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, 7th Ed.*, authors: Dr. R. Lyman Ott & Dr. 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 crime, 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.

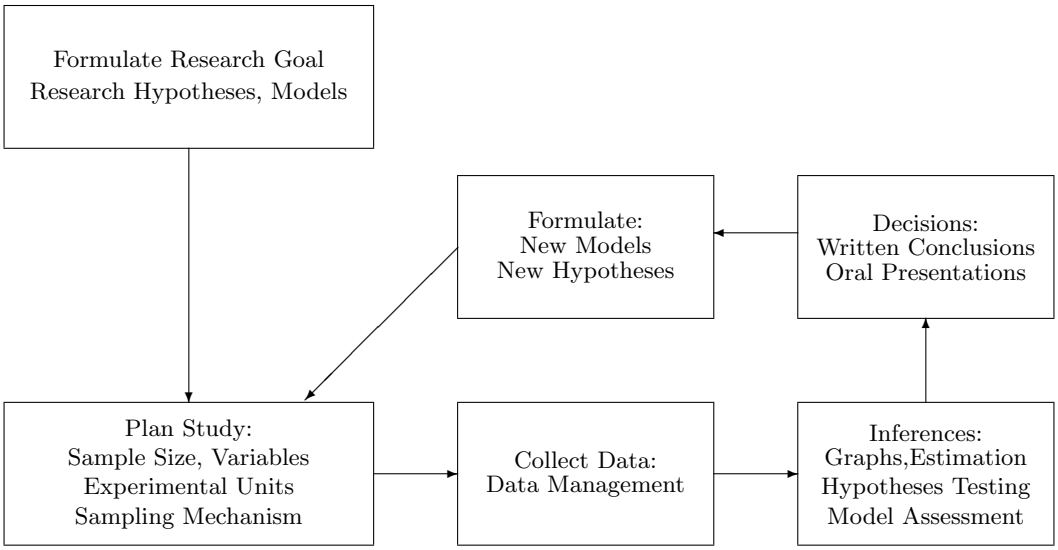
In this course, we will approach the study of statistics by considering a process by which we Learn from Data:

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 goals, 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.

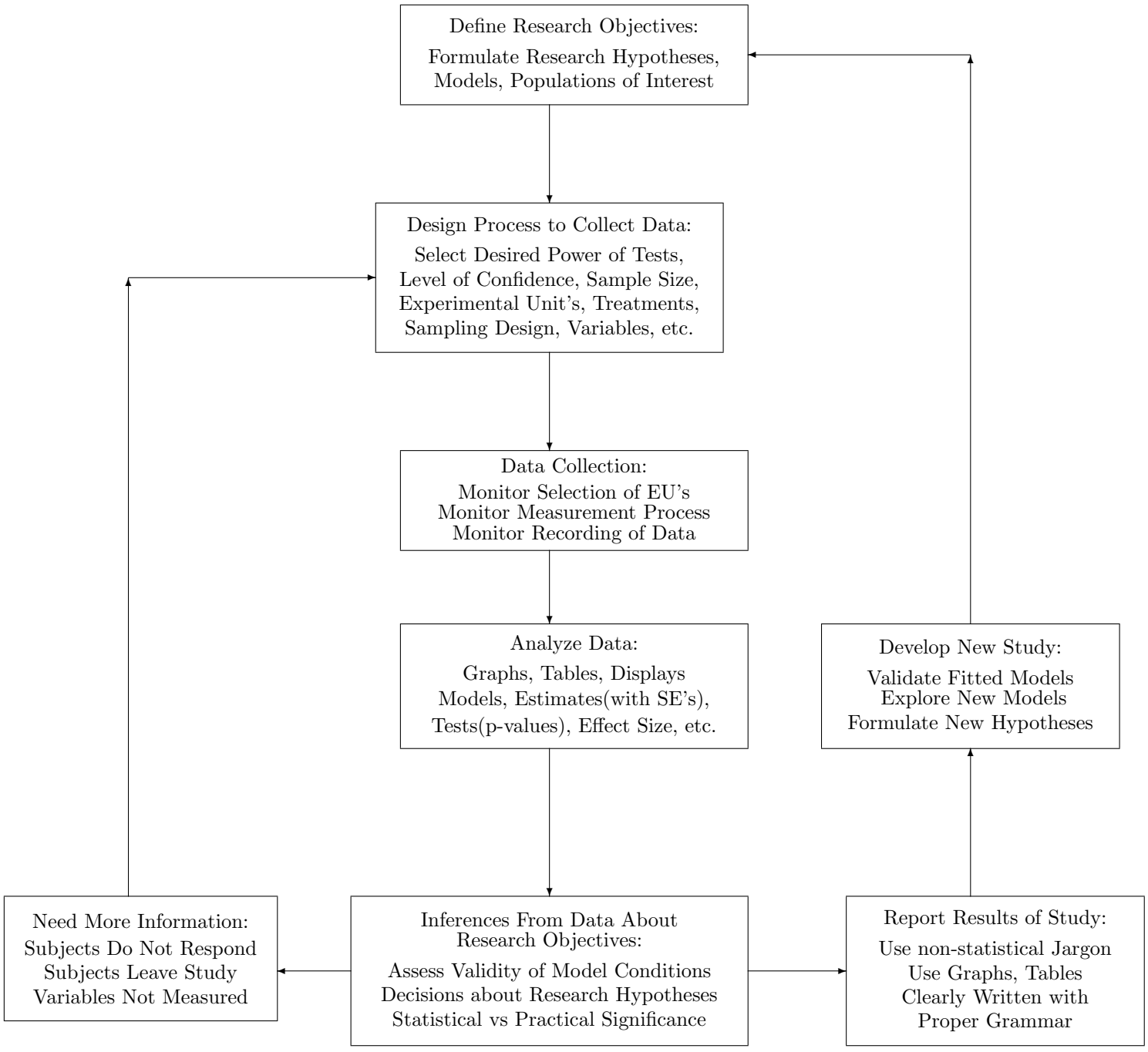
The design of experiments, collection of data, and analysis of data are integral components of the **Scientific Method**. Researchers use the results of studies and experiments to examine the validity of existing theories, to revise these theories, and eventually to formulate new theories. When the observed 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 many 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 or a 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 participant's 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.

There are "good samples", "bad samples", "biased samples", and "properly selected samples".

The problem is identifying when samples are bad and/or biased.

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 $25 per year, you as a student can be a member of ASA. Information on joining is given at the website: **www.amstat.org** under MEMBERSHIP then JOIN ASA!

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 cannot 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 is 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 in silicone breast implants, and a disease condition, such as an autoimmune disease. An epidemiologist who finds an association should try 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.

The Eagle • theeagle.com
Nation - B/CS Texas

Monday, August 21, 2017

Fallen forensics

Lawyers, scientists: Judges allowing disavowed methods

By DENISE LAVOIE
Associated Press

BOSTON — Two hairs that looked like the victim's; some dirt on a truck like that taken from the crime scene; a pattern on the bumper that resembled a design on the victim's popular brand of jeans. The case against Steven Barnes in the rape and murder of a 16-year-old girl seemed circumstantial, at best.

So the guilty verdict shocked him.

"I was saying, 'This can't be happening. You can't convict somebody on similarities, perhaps or maybes,'" Barnes said.

He spent the next 20 years in prison before DNA testing exonerated him, becoming one of hundreds of people convicted in whole or in part on forensic science that has come under fire during the past decade.

Some of that science — analysis of bite marks, latent fingerprints, firearms identification, burn patterns in arson investigations, footwear patterns and tire treads — was once considered sound, but is now being denounced by some lawyers and scientists who say it has not been studied enough to prove its reliability and in some cases has led to wrongful convictions.

Even so, judges nationwide continue to admit such evidence regularly.

"Courts — unlike scientists — rely too heavily on precedent and not enough on the progress of science," said Christopher Fabricant, director of strategic litigation for the Innocence Project. "At some point, we have to acknowledge that precedent has to be overruled by scientific reality."

Defense lawyers and civil rights advocates say prosecutors and judges are slow to acknowledge that some forensic science methods are flawed because they are the very tools that have for decades helped win convictions. And such evidence can be persuasive for jurors, many of whom who have seen it used dramatically on *Law & Order* and *CSI*.

Rulings in the past year show judges are reluctant to rule against long-accepted evidence even when serious questions have been raised about its reliability.

Two reports by scientific boards have sharply criticized the use of such forensic evidence, and universities that teach it are moving away from visual analysis — essentially, eyeballing it — and toward more precise biometric tools.

But some defense lawyers fear any progress on strengthening forensic science may be lost under President Donald Trump.

In April, Attorney General Jeff Sessions announced the Justice Department would disband the National Commission on Forensic Science, an independent panel of scientists, researchers, judges and attorneys that had been studying how to improve forensic practices.

The National Registry of Exonerations at the University of California Irvine has documented more than 2,000 exonerations since 1989. Nearly one-fourth list "false or misleading forensic evidence" as a contributing factor.

And a report last fall from the President's Council criticized several "feature-comparison" methods. The council said those methods — including analysis of shoeprints, tire tracks, latent fingerprints, firearms and spent ammunition — need more study to determine their reliability and error rates.

Forensic methods in doubt

BITE MARKS: Involves examining marks left on a victim or object, then comparing those with dental impressions taken from a suspect. Only a few empirical studies have been done to study the ability of examiners to accurately identify the source of a bite mark; those found false positive rates were so high that the method is "clearly scientifically unreliable at present."

FINGERPRINTS: Two recent black-box studies found that latent fingerprint analysis is foundationally valid, but also found that false positive rates could be as high as one error in 306 cases in one study and one error in 18 cases in the other. "Additional black-box studies are needed to clarify the reliability of the method."

FIREARMS: In which examiners attempt to determine whether ammunition came from a specific gun based on marks produced by guns on the ammunition. A 2016 report from the President's Council of Advisors on Science and Technology said that there is now only one appropriately designed study to measure its scientific validity and to estimate its reliability, and that it needs additional studies.

FOOTWEAR: In which examiners compare shoeprints found at a crime scene with specific shoes based on identifying marks. "Such conclusions are unsupported by any meaningful evidence or estimates of their accuracy and thus are not scientifically valid," the President's Council report said, adding that it needs more studies.

HAIR: Involves examiners comparing microscopic features of hair to determine whether a particular person may be the source. The FBI now acknowledges that microscopic hair analysis is inconclusive and uses it only in conjunction with DNA testing. The President's Council found that studies cited in a Department of Justice report "do not provide a scientific basis for concluding that microscopic hair examination is a valid and reliable process."

COMPLEX DNA MIXTURES: DNA analysis of complex mixtures of biological samples from multiple unknown people in unknown proportions; for example, from mixed blood stains. The President's Council found that subjective analysis of complex DNA mixtures by examiners has not been established as scientifically valid, but said computer programs that use algorithms to interpret complex mixtures in an objective way are a major improvement.

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 for 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 ensure 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 give 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 a 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 ask 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.

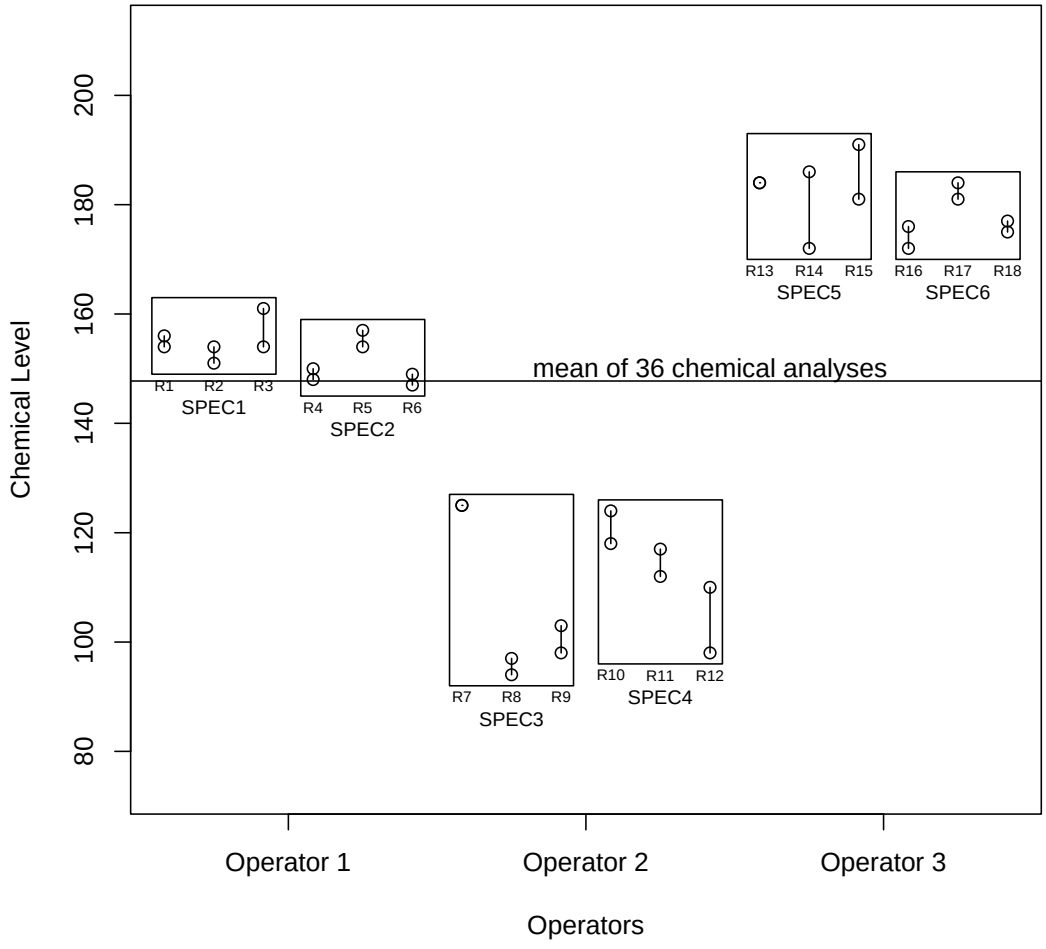
In order to investigate sources of variability for a chemical analysis, an experiment was designed and analyzed to ensure that relevant sources of variation could be identified and measured. The process engineer and operators discussed the situation and agreed that the four possible major sources of 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 following R code produced Figure 3

```
run=rep(1:18, each=2)

Res = c(156,154,151,154,154,161,148,150,154,157,147,149,125,125,94,97,
        98,103,118,124,112,117,98,110,184,184,172,186,181,191,172,176,
        181,184,175,177)
#Put in experimental value

data=matrix(data = Res,nrow = 18,ncol = 2,byrow = TRUE)
#Create Matrix to store the values

plot(run,Res,type="p",xlab="Operators",ylab="Chemical Level",
main="Figure 3: Chemical Variation Plot ",cex=.99,
ylim=c(min(Res)-20,max(Res)+20),xaxt="n")
#Creat Original plot

rec_d1=seq(0.75,15.75,by=3)
rec_d2=seq(3.25,18.25,by=3)#Set the rectangle range

tex=c('R1','R2','R3','R4','R5','R6','R7','R8','R9',
      'R10','R11','R12','R13','R14','R15','R16','R17','R18')
tex2=c('SPEC1','SPEC2','SPEC3','SPEC4','SPEC5','SPEC6')
#Set the text in plot

for (i in (1:6)) {
  low=3*i-2; high=3*i
  sub_data=data[low:high,]
  rect(rec_d1[i],min(sub_data)-2,rec_d2[i],max(sub_data)+2)
  segments(low,sub_data[1,1],low,sub_data[1,2])
  segments(low+1,sub_data[2,1],low+1,sub_data[2,2])
  segments(low+2,sub_data[3,1],low+2,sub_data[3,2])
  text(low,min(sub_data)-4,tex[low],cex=.55)
  text(low+1,min(sub_data)-4,tex[low+1],cex=.55)
  text(low+2,min(sub_data)-4,tex[low+2],cex=.55)
  text(low+1,min(sub_data)-8,tex2[i],cex=.75)
}#creat rectangle and segments within


axis(side=1,at=c(3.5,9.5,15.5),labels=c("Operator 1","Operator 2","Operator 3"))

segments(0,mean(Res),19,mean(Res))
text(12,mean(Res)+2,"mean of 36 chemical analyses")
#Create overall mean line
```

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:

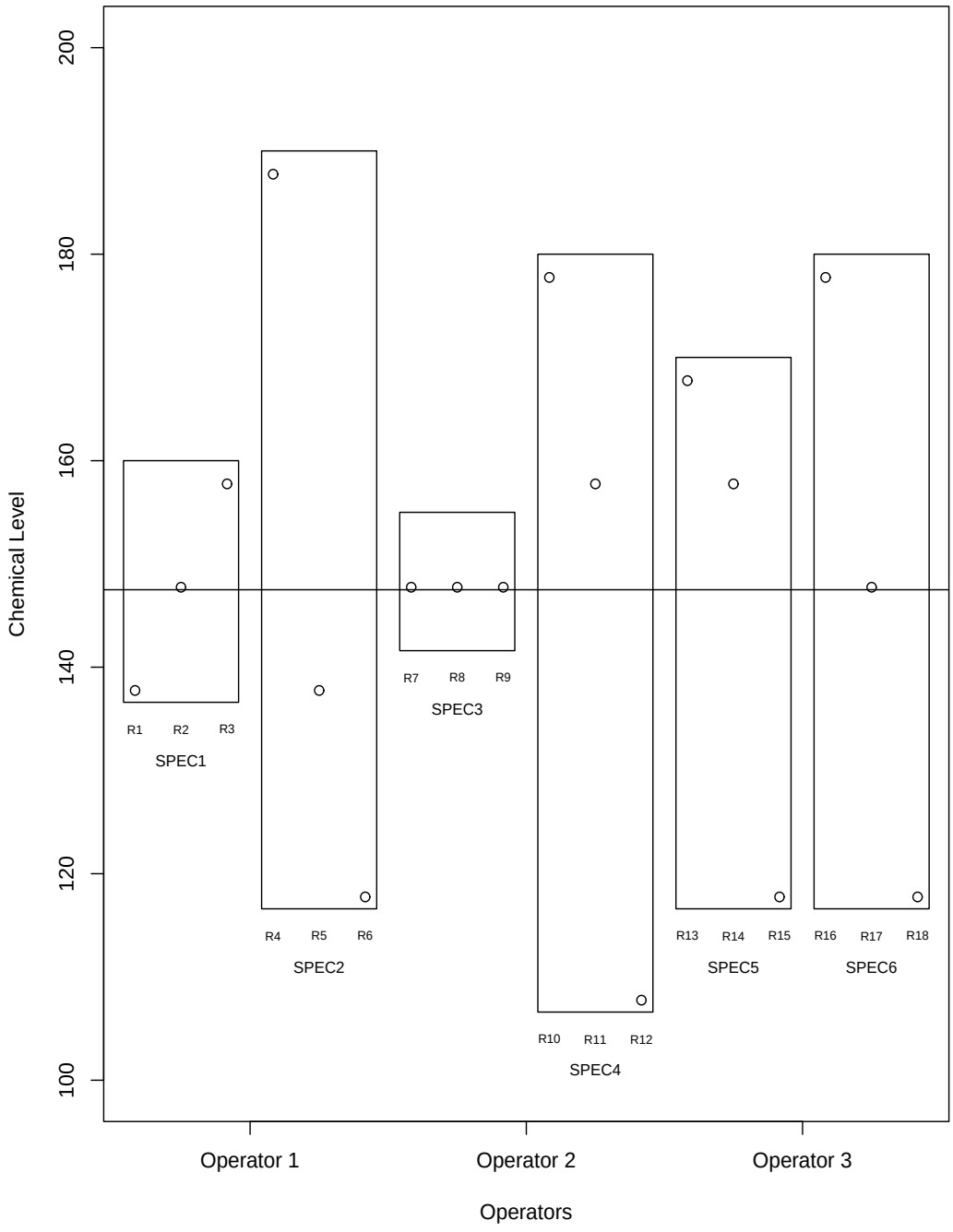
1. 94.80% Operator
2. 0.00% Specimen
3. 3.83% Combustion Run
4. 1.29% Chemical Analysis
5. 0.08% Other Sources

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, chemplot.r, used to produce the Figure 4:

```
run=rep(1:18, each=2)

Res = c(156,154,151,154,154,161,148,150,154,157,147,149,125,125,94,97,
        98,103,118,124,112,117,98,110,184,184,172,186,181,191,172,176,
        181,184,175,177) #Put in experimental value

data=matrix(data = Res,nrow = 18,ncol = 2,byrow = TRUE)#Create Matrix to store the values

plot(run,Res,type="p",xlab="Operators",ylab="Chemical Level",main="Chemical
     Variation Plot ",cex=.99,ylim=c(min(Res)-20,max(Res)+20),xaxt="n") #Creat Original plot

rec_d1=seq(0.75,15.75,by=3)
rec_d2=seq(3.25,18.25,by=3)#Set the rectangle range

tex=c('R1','R2','R3','R4','R5','R6','R7','R8','R9',
      'R10','R11','R12','R13','R14','R15','R16','R17','R18')
tex2=c('SPEC1','SPEC2','SPEC3','SPEC4','SPEC5','SPEC6')#Set the text in plot

for (i in (1:6)) {
  low=3*i-2; high=3*i
  sub_data=data[low:high,]
  rect(rec_d1[i],min(sub_data)-2,rec_d2[i],max(sub_data)+2)
  segments(low,sub_data[1,1],low,sub_data[1,2])
  segments(low+1,sub_data[2,1],low+1,sub_data[2,2])
  segments(low+2,sub_data[3,1],low+2,sub_data[3,2])
  text(low,min(sub_data)-4,tex[low],cex=.55)
  text(low+1,min(sub_data)-4,tex[low+1],cex=.55)
  text(low+2,min(sub_data)-4,tex[low+2],cex=.55)
  text(low+1,min(sub_data)-8,tex2[i],cex=.75)
}#creat rectangle and segments within


axis(side=1,at=c(3.5,9.5,15.5),labels=c("Operator 1","Operator 2","Operator 3"))

segments(0,mean(Res),19,mean(Res))
text(12,mean(Res)+2,"mean of 36 chemical analyses") #Create overall mean line
```

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:

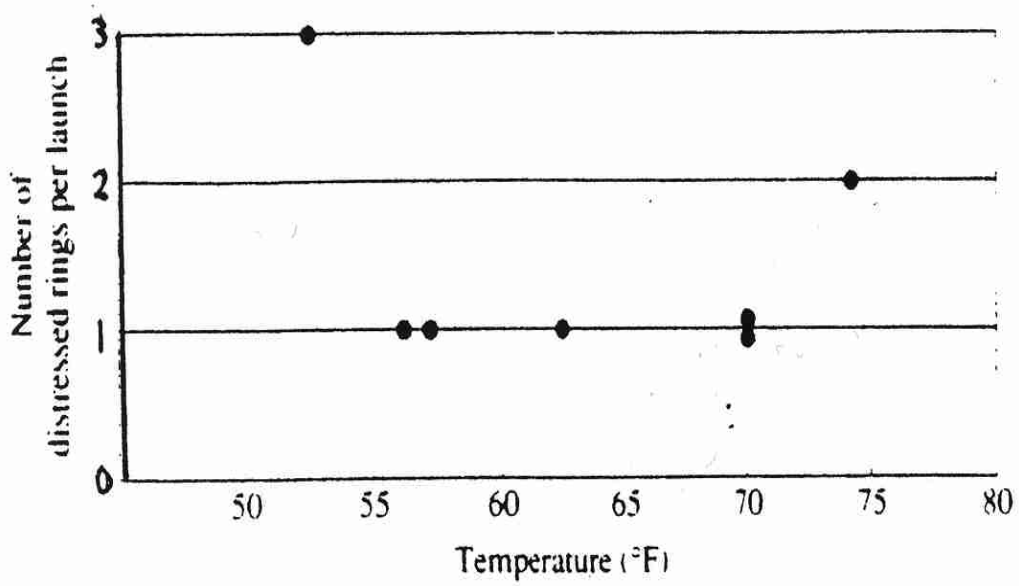


FIGURE 1.5-1 Scatter plot of number of distressed rings per launch against temperature.

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:

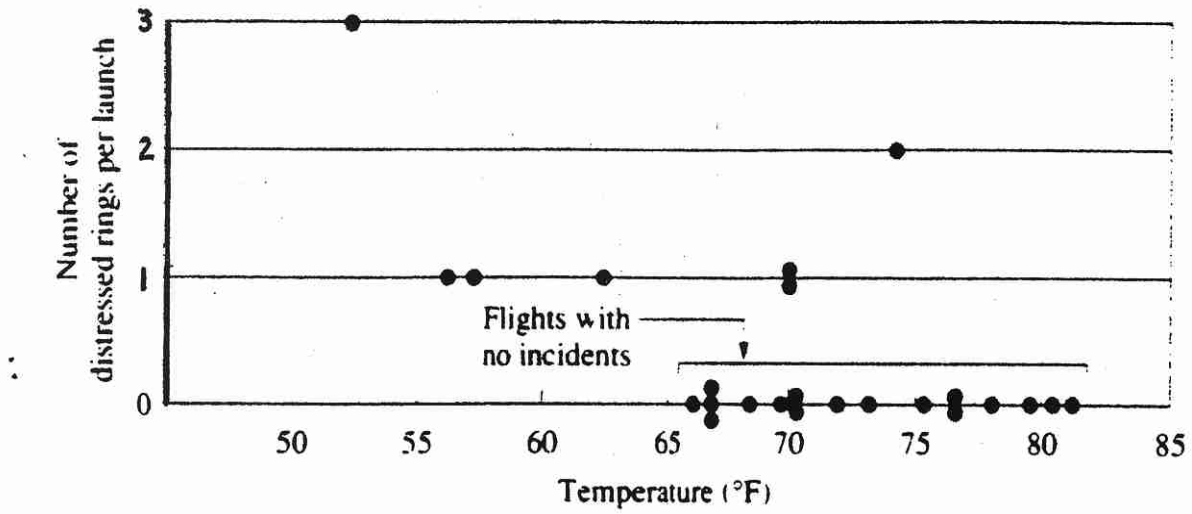


FIGURE 1.5-2 Scatter plot of number of distressed rings per launch against temperature (all data).

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

'Value at risk' involves a huge conceptual error

JAMES G. RICKARDS, Washington Post

Published 5:30 am, Friday, October 3, 2008

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.

HANDOUT #2 - TYPES OF STATISTICAL STUDIES

TOPICS

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

Supplemental Reading:

- Chapter 3 in Tamhane/Dunlop book

Sampling From a Population

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

Two basic types of studies: Observational and Experimental

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

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

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

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

  Examples:

  - Clinical Trials (Some control),

  - Laboratory Studies (More Control),

  - Agricultural Field Trials (Some Control),

  - Greenhouse Experiments (More Control),

  - Pilot Plants in Industry (More Control)

Observational studies are of four basic types:

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

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

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

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

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

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

  Examples:

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

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

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

The Study Begins

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

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

What Went Wrong?

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

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

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

The Study Ends and Reparation Begins

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

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

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

Preamble

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

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

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

General Principles

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

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

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

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

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

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

Sample Survey Example

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

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

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

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

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

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

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

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

Comparison of Retrospective and Prospective Studies

- Retrospective studies are generally cheaper and can be completed more rapidly than prospective studies.
- Retrospective studies have problems due to inaccuracies in data due to recall errors.

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

  Customer Survey: Was your shopping experience at our store enjoyable?
- Retrospective studies have no control over variables which may affect disease occurrence. Dietary Study: There are many other factors other than diet that may impact onset of Colon Cancer - Genetics, Occupation, Environment
- In prospective studies subjects can keep careful records of their daily activities

  Diet Diary, Check Ups, Record weight at 8am every day
- In prospective studies subjects can be instructed to avoid certain activities which may bias the study

  Exposure to risk factors - environmental toxins,work and personal stress factors, etc.
- Although prospective studies reduce some of the problems of retrospective studies they are still observational studies and hence the potential influences of confounding variables may not be completely controlled. It is possible to somewhat reduce the influence of the confounding variables by restricting the study to matched subgroups of subjects.

  Group subjects according to similar occupations, ethnicity, location of residency
- Both prospective and retrospective studies are often comparative in nature. Two specific types of such studies are **cohort studies** and **case-control studies**.
  - Cohort Studies: Follow a group of subjects forward in time to observe the differences in characteristics of subjects who develop a disease with those who do not. Prospective or Retrospective?
  - Case-Control Studies: Identify two groups of subjects, one with the disease and one without the disease. Next, gather information about the subjects from their past concerning risk factors which are associated with the disease. Prospective or Retrospective?
  - Case-Control studies are an improvement over just taking a random sample. Why?

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

Sampling versus Non-sampling Errors:

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

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

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

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

   Younger people participate at a lower rate than older persons

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

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

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

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

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

   Did you inhale?

   Do you use illegal drugs?

   Have you ever cheated on your income tax form?


ZIGGY
LAST QUESTION FOR OUR IMPARTIAL POLL: HOW ROTTEN A JOB IS OUR AWFUL PRESIDENT DOING SO FAR?
PRESS
POLL
©2017 Ziggy and Friends, Inc./Dist. by Andrews McMeel

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

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

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

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

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

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

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

7. Possible ways to deal with non-response:

   a. Design survey so that non-response is low

      Follow up phone calls to non-respondents

      Offer payment/donation to charity if survey is returned

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

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

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

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

What Went Wrong With the 2016 Polls?

The Atlantic November 9, 2016

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pollsters tackle what went wrong – Politico November 11, 2016

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

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

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

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

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

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

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

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

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

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

Went Wrong With the 2016 PollsP5.pdf

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

Here are the questions pollsters are asking this week:

Did Trump surge at the end?

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

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

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

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

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

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

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

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

Clinton aides blame loss on everything but themselves

By Annie Karni

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

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

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

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

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

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

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

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

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

Were there "shy Trump" supporters?

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

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

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

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

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

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

Who is a likely voter?

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

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

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

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

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

Meanwhile, turnout appears higher outside cities and suburbs.

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

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

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

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

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

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

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

SAMPLING PRINCIPLES

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

Why select a sample in place of a census?

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

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

What is the population of interest to the researcher?

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

What is the method by which the sample was selected?

- Examine all VW cars in used car lots in Houston

What is the population of interest to the researcher?

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

What is the method by which the sample was selected?

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

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

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

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

Probability Sampling

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

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

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

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

Example of non-probability sampling:

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

   Examples of convenience samples:

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

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

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

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

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

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

RANDOM SAMPLING

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

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

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

    EX.       sample(10)

              3  8 10  6  9  5  1  4  7  2


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

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

              sample(x)

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


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

    EX.       sample(x,5)

              67.00  21.00  45.00   0.25 -45.00


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

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

              -45.0   4.0  -3.0 -45.0   0.9
```

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

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

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


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

           0.25  0.25 45.00  0.25  0.25
```

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

"sample(N,n)"

```
EX.        sample(1000,10)

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

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

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

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

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

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

SYSTEMATIC RANDOM SAMPLING

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

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

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

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

$$\vdots$$

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

Randomly select 1 of the k clusters

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

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

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

The Sample then consists of the following units:

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

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

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

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

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

STRATIFIED RANDOM SAMPLING

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

Reasons for Using a Stratified Random Sample:

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

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

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

CLUSTER RANDOM SAMPLING

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

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

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

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

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

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

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

Stratified Sampling vs Cluster Sampling

Stratified Sampling:

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

Cluster Sampling:

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

Example 1:

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

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

This is a Multistage Cluster Random Sample

Example 2:

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

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

This is a Stratified Multistage Cluster Random Sample

ESTIMATION OF POPULATION MEAN: μ

Consider the estimation of μ under three different sampling Methods

Simple Random Sampling

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

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

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

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

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

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

Stratified Random Sampling

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

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

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

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

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

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

SINGLE STAGE CLUSTER Random Sampling

Let N be the number of clusters in the population;

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

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

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

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

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

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

The estimator of the population mean μ is

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

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

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

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

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

EXPERIMENTAL DESIGN PRINCIPLES

The rest of the handout will be covered in STAT 642

Two very important comments from noted pioneers of applied statistics:

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

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

Selected Comments from *Experimental Design* by W. Federer

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

   The variability of experimental responses."

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

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

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

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

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

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

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

   1. Review the relevant literature

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

   Too many 0's,

   Categorical data,

   Too few replications for projected variability,

   Correlated (nonindependent observations)

P3. Selection of Procedures for Conducting Research

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

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

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

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

P6. Quality of the Final Report:

1. Include well designed graphics

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

IV. Statistically Designed Experiments are

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

EXPERIMENTAL DESIGN TERMINOLOGY

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:

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

C2. Treatment Structure

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

C3. Measurement Structure

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

II. Specific Terms Used to Describe Designed Experiment:

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

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

EXAMPLE

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

- Designed Experiment Consists of Three Components:

  C1. Method of Randomization:

  C2. Treatment Structure:

  C3. Measurement Structure:

OTHER POSSIBLE WAYS OF CONDUCTING THE WAFER EXPERIMENT

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

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

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

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

COMMON PROBLEMS IN EXPERIMENTAL DESIGNS

I. Masking of Factor Effects

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

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

II. Uncontrolled Factors

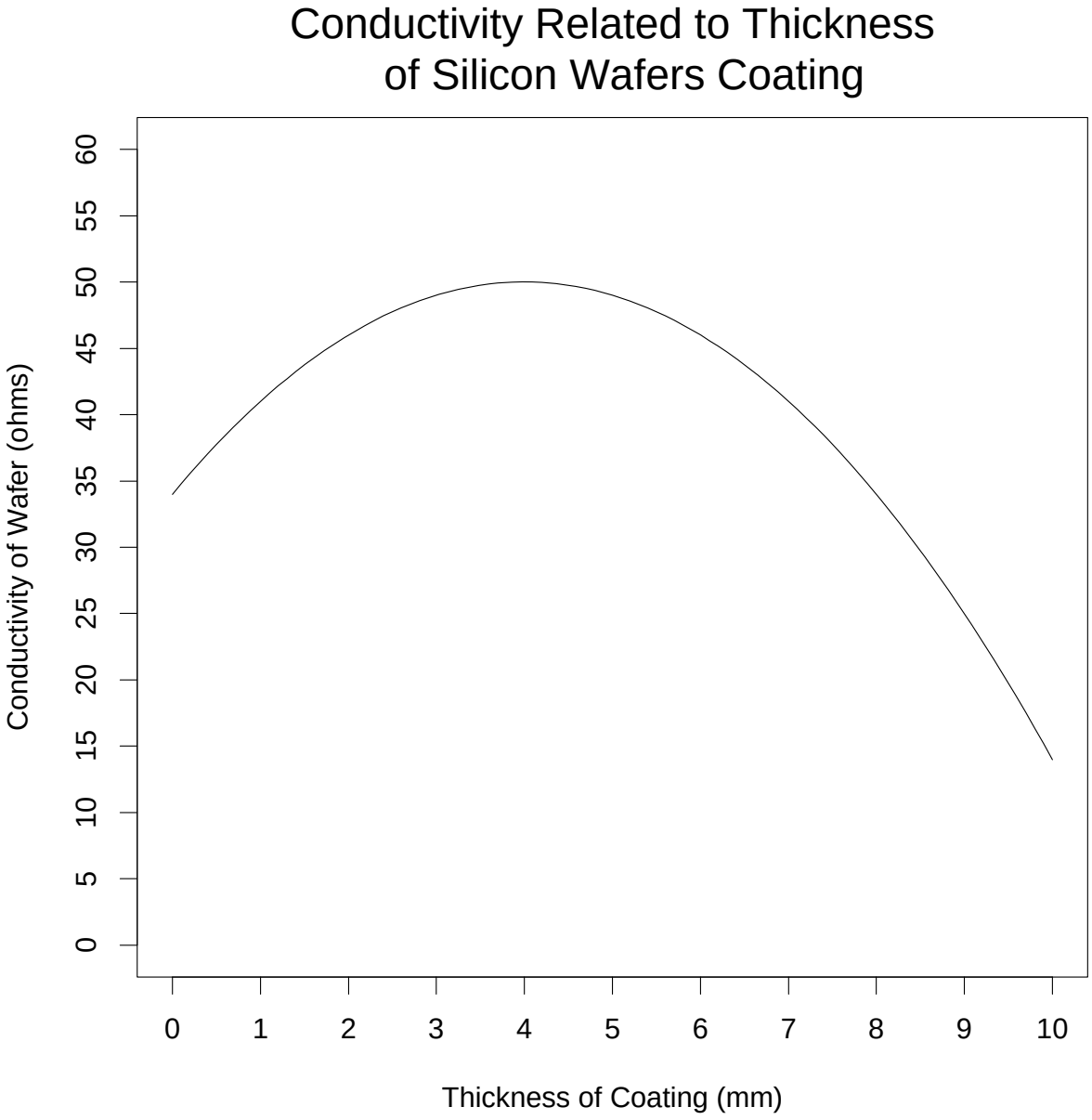
If factors are known to have an effect on the response variable, then these factors should be included in the experiment as either treatment or blocking variables. Failure to carefully consider all factors of importance can greatly compromise the extent to which conclusions can be drawn from the experimental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

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

III. Erroneous Principles of Efficiency

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

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


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

SELECTING AN APPROPRIATE EXPERIMENTAL DESIGN

I. Consideration of Objectives

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

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

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

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

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

3. Determination of observable variables

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

II. Factor Effects

1. Inclusion of all relevant factors avoids uncontrolled systematic variation.
2. Need to measure all important covariates to control heterogeneity of experimental units or conditions.
3. Anticipated interrelationships between factor levels helps to determine type of design:
   a. No interactions between factor levels: Use simple screening design
   b. Interactions exist: Need full factorial design
   c. Higher order relationships between factor levels may require a greater number of levels of the factors in order to be able to fit high order polynomials to the responses.

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

III. Precision - Efficiency of Experiment

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

IV. Randomization

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

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

DESIGNING FOR QUALITY: INDUSTRIAL PROCESSES

Two Basic Types of Experiments

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

   EVOP - Evolutionary Operation

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

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

Two Basic Goals in Experiments Involving Quality Improvement

1. Bring product On Target

   Average measurement of product characteristic are equal to the target value

2. Uniformity - Consistency
   Measured product characteristics have a small variability about the target value

Combining both of these criterions, we obtain
Minimize MSE = $(Bias)^2 + (StDev)^2$ = (Distance to Target)^2+Variance

Taguchi Approach:

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

HANDOUT #3 - Summaries of Population Distributions

TOPICS

1. Definition of Population/Process
2. Definition of Random Variable
3. Types of Random Variables
4. Functions which Characterize Random Variables/Populations
5. Families of Distributions Indexed by Parameters
6. Examples of Distributions: Discrete, Continuous, Mixtures
7. Interrelationships between Various Distributions
8. Simulation of Data from Specified Distributions
9. Functions Associated with Reliability/Survival Analysis

Supplemental Reading:

- Chapter 2 in Tamhane-Dunlop book
- *Statistical Distributions* by Forbes, Evans, Hastings, and Peacock

Definition of Population/Process/Random Variable

1. **Statistical Population -** Collection of all possible items or units possessing one or more common characteristics under specified experimental or observational conditions

2. **Process -** Repeatable series of actions that results in an observable characteristic or measurement

   Industrial and Laboratory experiments often are characterized as hypothetical populations. Why?

   Because the population of values does not exist at the beginning of the experiment or study and in fact, may never exist.

   Example 1: Study of the effect of dehydration on ticks by placing 100 ticks to a vessel having a very dry climate. What is the hypothetical population?

   • All ticks in a very dry climate

   Example 2: Study the gain in productivity of a random sample of 25 assembly workers who complete a new training program. What is the hypothetical population?

   • All workers that will sometime in the future attend the training program - a hypothetical population

3. **Random Experiment -** Procedure or operation whose outcome is uncertain and cannot be predicted in advance

   • Expose 3 rats to a potentially toxic chemical and observe the number of survivors 24 hours later.

4. **Sample Space -** Collection of all possible outcomes of a random experiment

   Let E_i be the event rat i was alive at the end of 24 hours

   $\overline{E_i}$ is the event rat i is dead at the end of 24 hours

   $2^3 = 8$ elements in the sample sample, $\mathcal{S}$

   $\mathcal{S} = \{E_1\, E_2\, E_3;\ \overline{E_1}\, E_2\, E_3;\ E_1\, \overline{E_2}\, E_3;\ E_1\, E_2\, \overline{E_3};\ \overline{E_1}\, \overline{E_2}\, E_3;\ \overline{E_1}\, E_2\, \overline{E_3};\ E_1\, \overline{E_2}\, \overline{E_3};\ \overline{E_1}\, \overline{E_2}\, \overline{E_3}\}$

5. **Random Variable (RV) -** A function, Y, which maps sample space to the real line

   $Y : \mathcal{S} \to (-\infty, \infty)$,

   Y assigns a unique numerical value to each element of the sample space

   For each s in $\mathcal{S}$, $Y(s)$ is a real number.

   Let N = number of rats surviving then $N : \mathcal{S} \to \{0, 1, 2, 3\}$

Example 1: Randomly select a sample of water (1 liter) from a river and record the amount, A, of PCB in the container in ppb

The sample space, S, is all possible 1 liter bottles of water from the river

A: $S \Rightarrow [0, \infty)$

Example 2: Randomly select light bulb from distribution center and measure one of the following characteristics of the bulb:

The sample S is all light bulbs in the distribution center

(a) Time, T, to failure of light bulb

$T : S \Rightarrow [0, 10000]$

where 10,000 is the maximum possible life length of the bulb

(b) Amount of protective coating, C, on bulb

$C : S \Rightarrow [0, .28]$

where .28 cm is the maximum possible coating thickness

(c) Determine if bulb is defective or not, with $D = 1$ if defective and $D = 0$ if not defective

$D : S \Rightarrow \{0, 1\}$

(d) Determine Quality, Q, of bulb, with $Q = 0$ if not defective, $Q = 1$ if defective but repairable, and $Q = 2$ if defective and non-repairable

$Q : S \Rightarrow \{0, 1, 2\}$

6. **Types of Random Variables -** Three major classifications:

   (a) **Discrete RV -** Collection of possible values of RV is at most a finite or countably infinite set

   - Random variables D and Q on previous page

   (b) **Continuous RV -** Collection of possible values of RV is one or more intervals on the real line (probability that it assumes any specific value is 0)

   - Random variables T and S on previous page

   (c) **Discrete-Continuous Mixture -** Collection of possible values of RV is one or more intervals on the real line and a set of distinct values

   Example 1: Let Y be the number of fish captured in a randomly placed net in the Gulf of Mexico divided by the length of time the net is in the water, Catch Per Unit Effort (CPUE)

   - 40% of the nets have no fish, Y=0, and the remaining 60% have a value of Y in the interval (0, 500).

   Example 2: Let X be the amount spent on health insurance per member of the household by a randomly selected employee at Texas A&M University.

   - 20% of the employees have no insurance, X=0, and the remaining employees have a value of X in the interval ($1200, $5000).

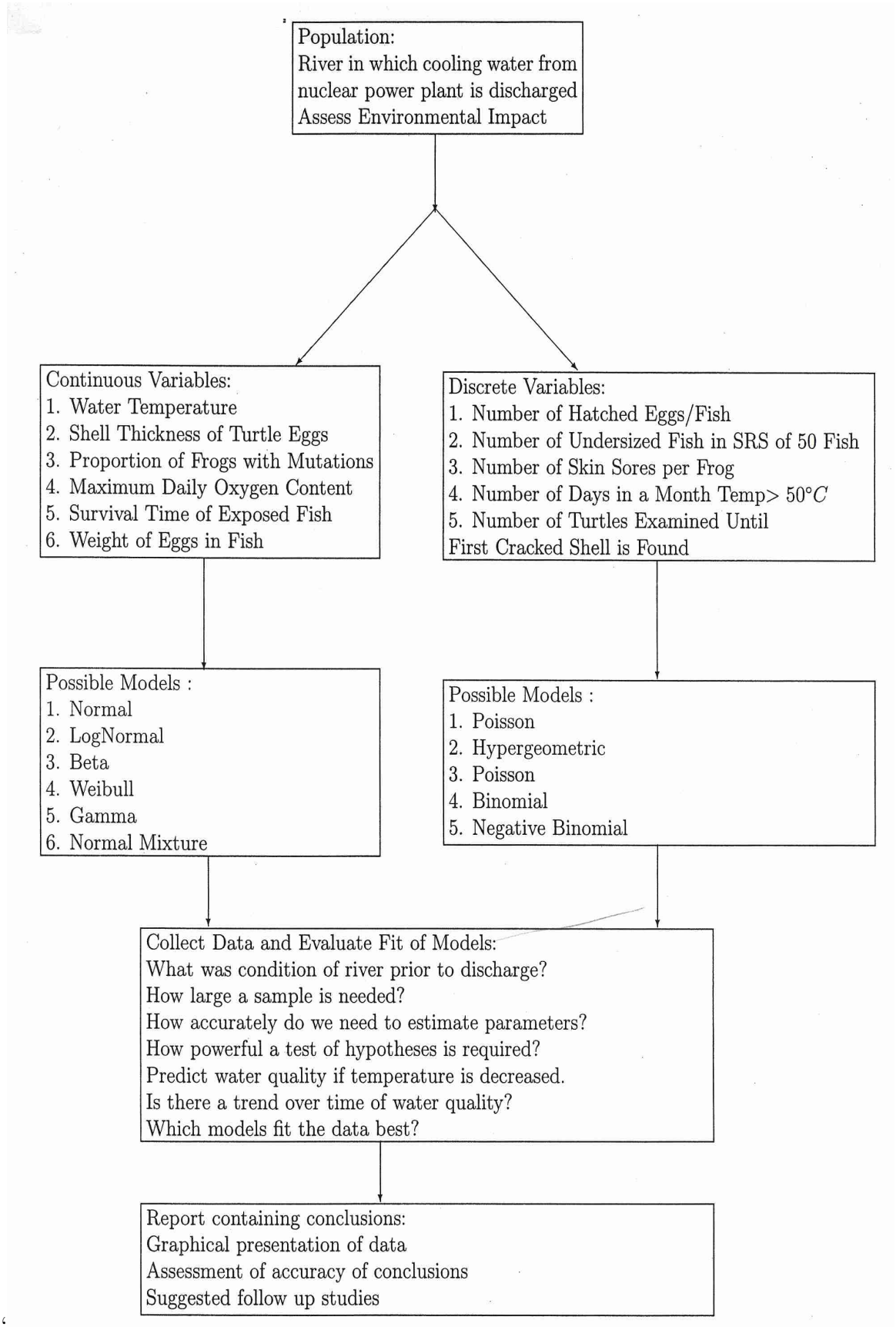
   Example 3: Let G be the growth rate of a randomly selected plant after receiving a prescribed amount of growth stimulant

   - 43% of the plants have no growth, G=0, and the remaining 57% have a value of G in (0, 28) cm.

   Example 4: Yearly payout on a health insurance policy

   - 46% of the policies have no payout, P=0, and the remaining 64% have a value of P in (0, $1,000,000)

The following diagram depicts a variety of RV's that may be defined on a single random experiment


Population:
River in which cooling water from
nuclear power plant is discharged
Assess Environmental Impact
Continuous Variables:
1. Water Temperature
2. Shell Thickness of Turtle Eggs
3. Proportion of Frogs with Mutations
4. Maximum Daily Oxygen Content
5. Survival Time of Exposed Fish
6. Weight of Eggs in Fish
Discrete Variables:
1. Number of Hatched Eggs/Fish
2. Number of Undersized Fish in SRS of 50 Fish
3. Number of Skin Sores per Frog
4. Number of Days in a Month Temp> 50°C
5. Number of Turtles Examined Until
First Cracked Shell is Found
Possible Models :
1. Normal
2. LogNormal
3. Beta
4. Weibull
5. Gamma
6. Normal Mixture
Possible Models :
1. Poisson
2. Hypergeometric
3. Poisson
4. Binomial
5. Negative Binomial
Collect Data and Evaluate Fit of Models:
What was condition of river prior to discharge?
How large a sample is needed?
How accurately do we need to estimate parameters?
How powerful a test of hypotheses is required?
Predict water quality if temperature is decreased.
Is there a trend over time of water quality?
Which models fit the data best?
Report containing conclusions:
Graphical presentation of data
Assessment of accuracy of conclusions
Suggested follow up studies

Characterization/Descriptions of Populations/Processes

Let **Y** be a R.V. associated with a Population/Process

Let **R(Y)** be the possible values of Y

Three Functions Which Completely Describe Y:

1. The Cumulative Distribution Function **(cdf)** of Y, $F(y)$:

$$F(y) = P[Y \leq y] \quad \text{for} \quad -\infty < y < \infty$$

That is, $F : (-\infty, \infty) \Rightarrow [0, 1]$ F maps $(-\infty, \infty)$ into [0,1]

$F(y)$ is the probability that the next observed value of the r.v. Y is less than or equal to y

$F(y)$ is the proportion of the population having values less than or equal to y

$F(y)$ is the proportion of the output of a process having values less than or equal to y

2. The Probability Mass Function **(pmf)** for discrete r.v.'s or Probability Density Function **(pdf)** for continuous r.v.'s

   (a) For Discrete R.V.'s:

$$f(y) = P[Y = y] = \quad \text{proportion of population values equal to y}$$

$$\text{cdf, F is related to pmf, f, by} \quad F(y) = P[Y \leq y] = \sum_{t \leq y} f(t)$$

   (b) For Continuous R.V.'s, the pdf is defined as that function, f, such that

$$f(y) \geq 0; \quad \text{with } F(y) = \int_{-\infty}^{y} f(t)dt \qquad \Rightarrow f(y) = \frac{dF(y)}{dy}$$

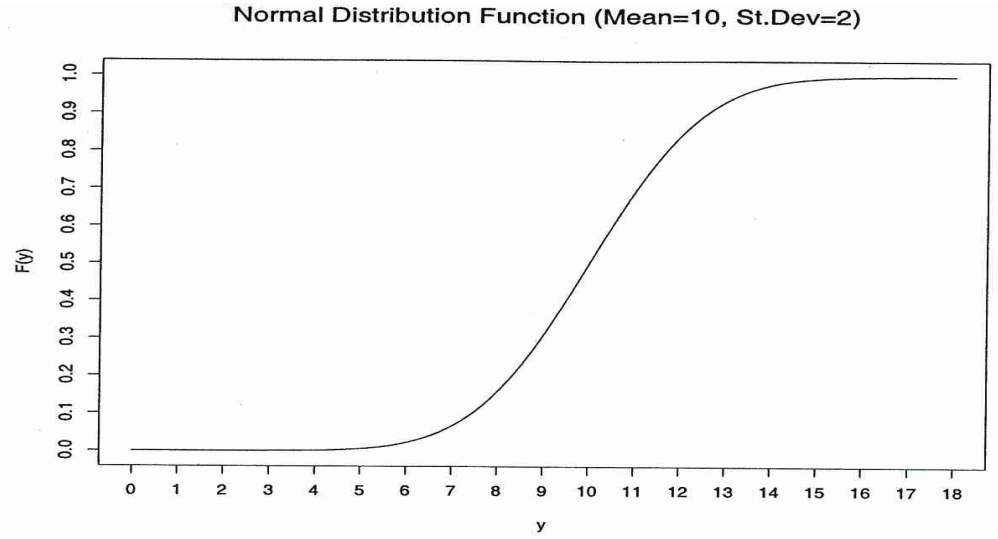
$$P[a \leq Y \leq b] = \int_a^b f(t)dt = \quad \text{area under} f() \quad \text{between a and b}$$

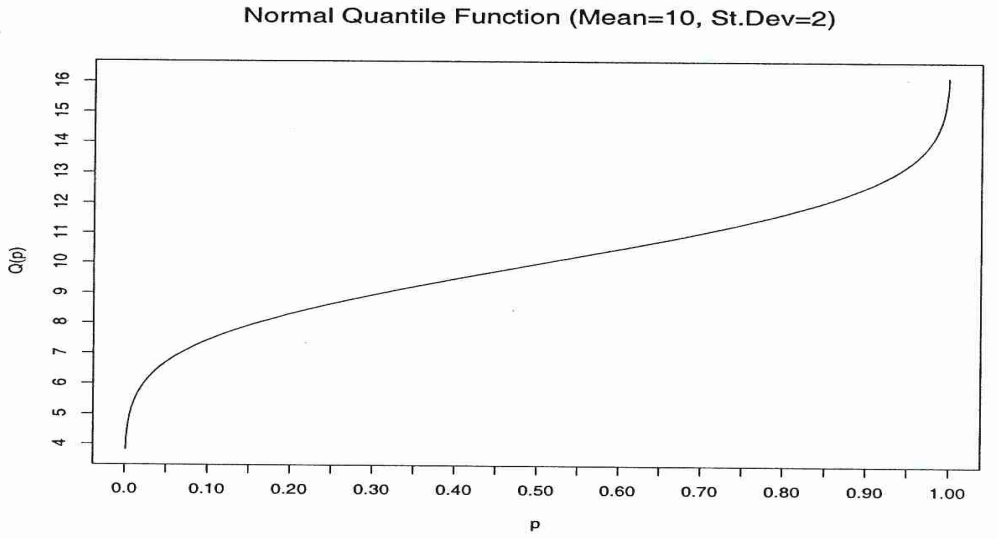
Note: $f(y)$ is the rate of increase in F at y. $f(y)$ **is not a probability**. In fact, it can have values greater than 1.0. For example, the exponential pdf with parameter $\lambda = 5$:

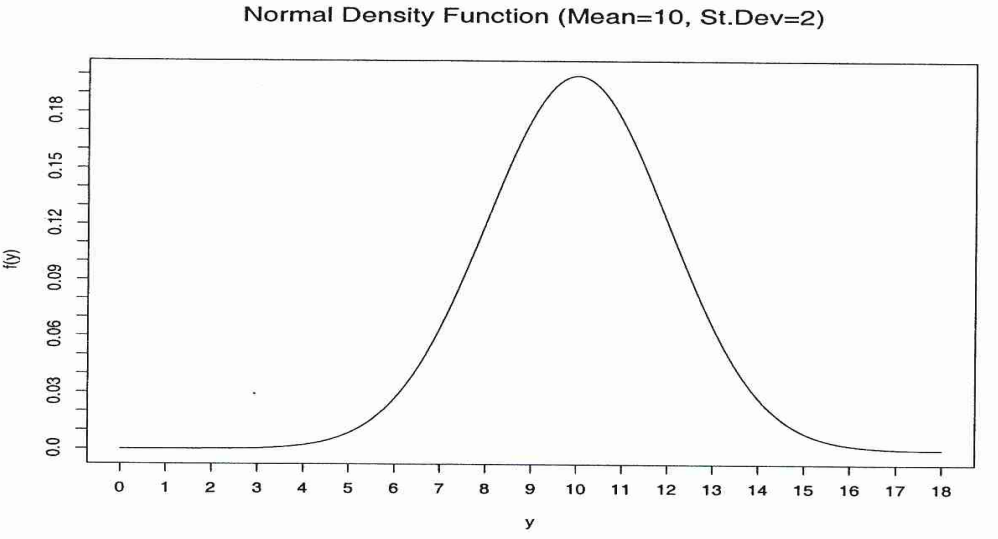
$$f(y) = 5e^{-5y} \text{ for } y \geq 0 \;\Rightarrow\; f(.04) = 5e^{-5(.04)} = 4.09 > 1$$

Using the definition, we have $f(y) = \frac{dF(y)}{dy} = \lim_{\Delta \to 0} \frac{F(y+\Delta) - F(y)}{\Delta} = \lim_{\Delta \to 0} \frac{P(Y \epsilon (y, y+\Delta))}{\Delta}$

Therefore, for very small Δ, $\Delta \cdot f(y) \approx P[Y \epsilon (y, y + \Delta)]$


Normal Distribution Function (Mean=10, St.Dev=2)
F(y)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
y


Normal Quantile Function (Mean=10, St.Dev=2)
Q(p)
4 5 6 7 8 9 10 11 12 13 14 15 16
0.0 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
p


Normal Density Function (Mean=10, St.Dev=2)
f(y)
0.0 0.03 0.06 0.09 0.12 0.15 0.18
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
y

Suppose D is a discrete random variable with possible values 0, 1, 2, 3, 4, 5 and probabilities:

$$f(d) = P(D = d) = \begin{cases} .03 & if \quad d = 0 \\ .16 & if \quad d = 1 \\ .35 & if \quad d = 2 \\ .25 & if \quad d = 3 \\ .15 & if \quad d = 4 \\ .06 & if \quad d = 5 \end{cases}$$

The pmf for D is $f(d) = P(D = d)$ with values given above.

The cdf for D is obtained from the expression: $F(d) = P(D \leq d) = \sum_{i=0}^{d} f(d)$, that is,

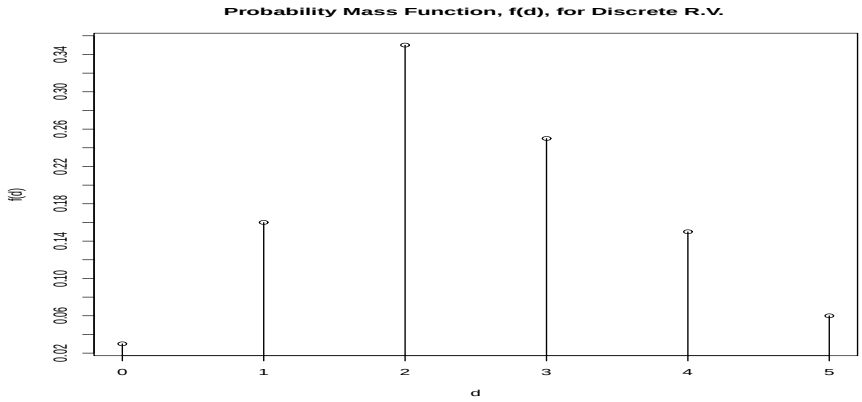
$$F(d) = P(D \leq d) = \begin{cases} 0 & if \quad d < 0 \\ .03 & if \quad 0 \leq d < 1 \\ .19 & if \quad 1 \leq d < 2 \\ .54 & if \quad 2 \leq d < 3 \\ .79 & if \quad 3 \leq d < 4 \\ .94 & if \quad 4 \leq d < 5 \\ 1 & if \quad 5 \leq d \end{cases}$$

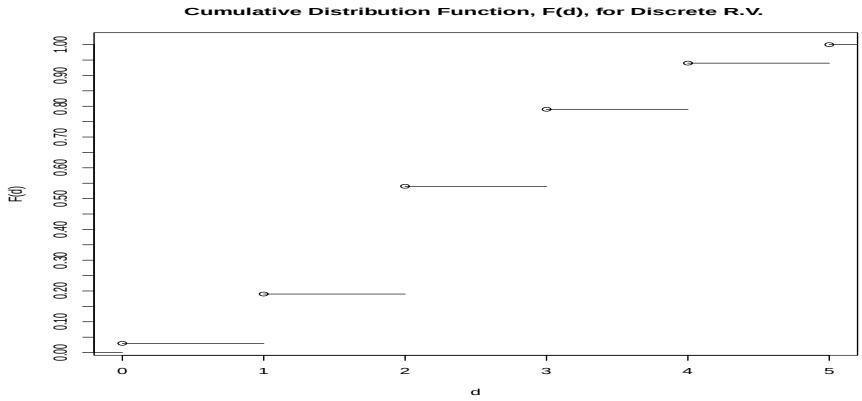
A graph of the cdf and pmf for the discrete r.v. D are given on the next page along with the quantile function, $Q(u)$ for $0 \leq u \leq 1$, which is the inverse of the cdf:

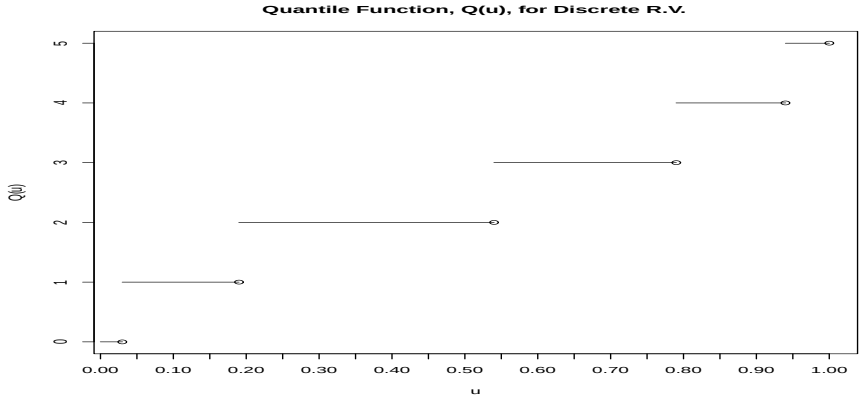
$$Q(u) = \begin{cases} 0 & if \quad 0 \leq u \leq .03 \\ 1 & if \quad .03 < u \leq .19 \\ 2 & if \quad .19 < u \leq .54 \\ 3 & if \quad .54 < u \leq .79 \\ 4 & if \quad .79 < u \leq .94 \\ 5 & if \quad .94 < u \leq 1 \end{cases}$$

Note: The pmf of D is obtained from the cdf by $f(d_j) = F(d_j) - F(d_{j-1})$.

That is, $f(3) = F(3) - F(2) = .79 - .54 = .25$


Probability Mass Function, f(d), for Discrete R.V.
f(d)
0.02
0.06
0.10
0.14
0.18
0.22
0.26
0.30
0.34
0
1
2
3
4
5
d

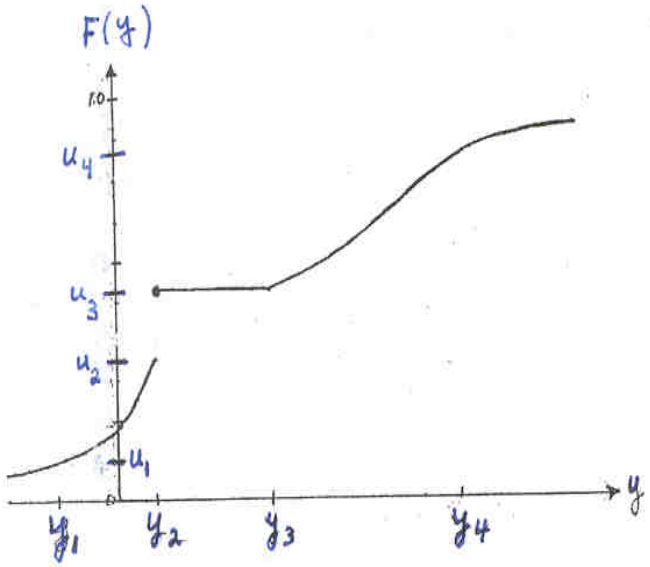

Cumulative Distribution Function, F(d), for Discrete R.V.
F(d)
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
0
1
2
3
4
5
d


Quantile Function, Q(u), for Discrete R.V.
Q(u)
0
1
2
3
4
5
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
u

Discrete - Continuous Mixture (point mass at $Y = y_2$)

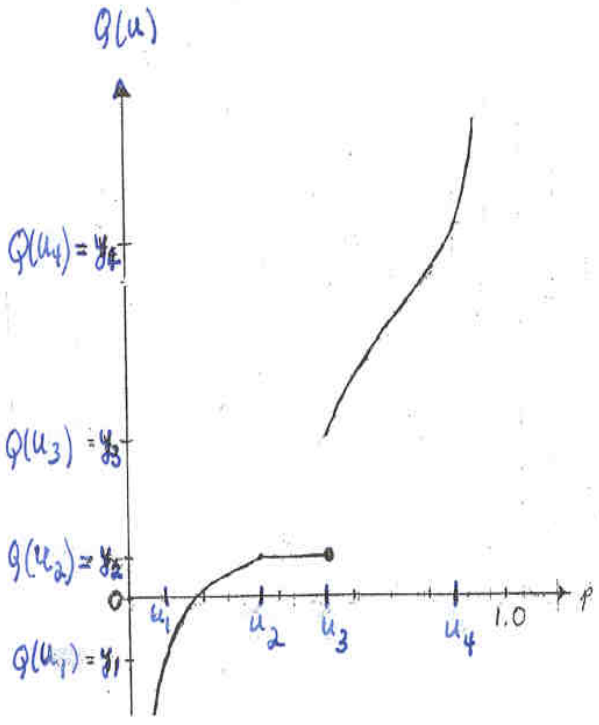
cdf

$F(y)$



Quantile

$Q(u)$



START Class Notes 9/8/21

3. The **Quantile Function** of Y, $Q(u)$:

Definition 1: Inverse of the cdf: $Q(u) = F^{-1}(u)$

$Q : [0, 1] \to (-\infty, \infty)$

- Case 1: For a continuous, strictly increasing cdf, $F(\cdot)$

$$Q(u) = y_u \quad \text{if and only if} \quad F(y_u) = u$$

- Case 2: For a discrete or discrete-continuous mixture r.v. the inverse of the cdf may not be defined for one of the following reasons:
  a. For specified $u\epsilon[0, 1]$ there may exists many real numbers y_u for which $F(y_u) = u$
  b. For specified $u\epsilon[0, 1]$ there is no real number y_u for which $F(y_u) = u$

In either case, the inverse of F would not be a valid function because it violates the definition of a function which states every value in the domain is assigned to one and only one value in the range of the function.

For example, see discrete-continuous mixture graphs on previous page:

1. For any value y satisfying $y_2 \leq y \leq y_3$ $F(y) = u_3$. Therefore, by the definition in Case 1,

   $$F(y_2) = u_3 \quad \text{and} \quad F(y_3) = u_3, \quad \text{in fact, for all } y \in [y_2,\ y_3], \quad F(y) = u_3$$

   Thus, by the definition of an inverse function,

   $$Q(u_3) = y \quad \text{for all } y \in [y_2,\ y_3]$$

   But this violates the definition of a function (a given value in the domain is mapped to multiple distinct values.)

2. For all u satisfying $u_2 < u < u_3$, there is no real number y_u for which $F(y_u) = u$.

   That is, there exists values in the domain which are not mapped by the function. Once again we have violated the definition of a function.

Thus, we have the following **Alternative Definitions:**

Definition 2: For $u\epsilon(0,1)$, the $100u$-quantile of the r.v. Y (or cdf F) is the real number $y_u = Q(u)$ such that

$$Q(u) = y_u = inf\{y : F(y) \geq u\}$$

That is, $Q(u)$ is the smallest value of y for which $F(y) \geq u$.

Special Case: If the r.v. Y is bounded below, that is, $Y \geq a$, then we define $Q(0) = a$.
Note: If we did not make the above specification, then Q(0) would be undefined because $F(y) = 0$ if $y < a$ thus $inf\{y : F(y) \geq 0\} = -\infty$ but $Y \geq a > -\infty$
Thus, we have the following: if $-\infty < a \leq Y \leq b < \infty$ then $Q(0) = a$ and $Q(1) = b$

Note that in our example of a discrete-continuous mixture distribution,

1. $Q(u_3) = y_2$ (y_2 is the smallest value of y for which $F(y) \geq u_3$)
2. For all u satisfying $u_2 < u < u_3$, $\quad Q(u) = y_2$

   (For u satisfying $u_2 < u < u_3$, $F(y_2) = u_3 > u$ and $F(y) < u$ for all $y < y_2$).

Note, the graph of $(u, Q(u))$ is a rotation of the mirror image of the graph of $(y, F(y))$:

- Jumps in the cdf F become flat regions in Q
- Flat regions in F become jumps in Q.

Remark: For $u\epsilon(0,1)$, if y_u is the $100u$ quantile of the r.v. Y or cdf F then

1. $P[Y \leq y_u] \geq u$ AND $P[Y \geq y_u] \geq 1 - u$

   or in terms of the cdf F

2. $F(y_u) \geq u$ AND $F(y_u^-) \leq u$

That is, $y_u = Q(u)$ is that value of Y such that at least $100u\%$ of the population values are less than or equal to y_u and that value of Y such that at least $100(1-u)\%$ of the population values are greater than or equal to y_u.

For distributions having cdf F strictly increasing on the support of the corresponding pdf and continuous, we can determine $Q(u)$ from $F(y)$ using the relationship:

$$y_u = Q(u) \text{ if and only if } F(y_u) = u.$$

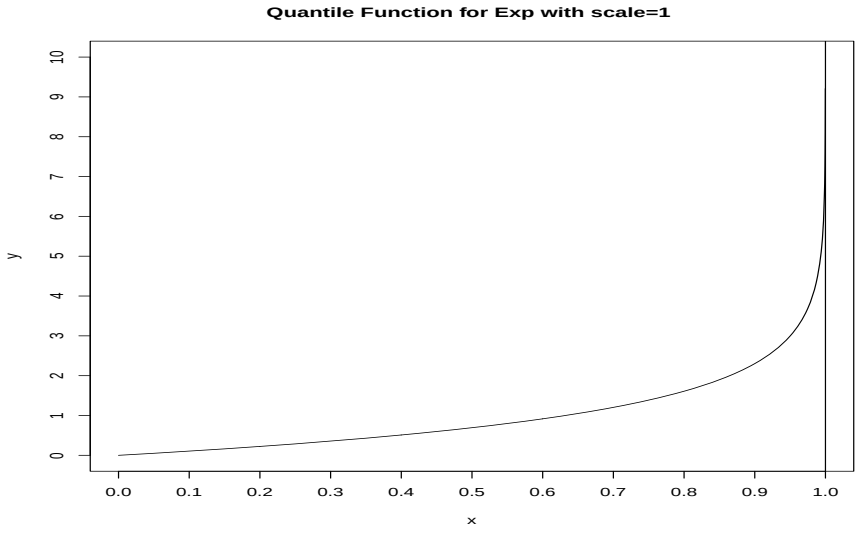
For example, suppose the random variable Y has cdf given by the following:

$$F(y) = \begin{cases} 1 - e^{-y} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases}$$

For $u > 0$, $u = F(y_u) = 1 - e^{-y_u} \Rightarrow log(1-u) = -y_u \Rightarrow$

$Q(u) = y_u = -log(1-u)$

Because $Y \geq 0$ we set $Q(u) = 0$ for $u = 0$



For further reading on the topic of quantile functions see Casella-Berger, *Statistical Inference* or John Rice, *Mathematical Statistics and Data Analysis*.

Location/Scale Families of CDF's

In many cases, the cdf or pdf of a r.v. Y is specified as a member of a family of distributions which are indexed by parameters:

$$Y \quad \text{has a pdf in the family} \quad \{f(y;\theta) : \theta \epsilon \Theta\}$$

The following examples will illustrate this notation:

Example 1 Y has pdf given by, for $-\infty < y < \infty$

$$f(y,\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi}\sqrt{\theta_2}} e^{-\frac{1}{2\theta_2}(y-\theta_1)^2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1,\theta_2) : \theta_1 \epsilon (-\infty,\infty), \theta_2 \epsilon (0,\infty)\}$$

Example 2 Y has pmf given by, for $y = 0, 1, \ldots, n$

$$f(y;\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for} \quad \theta \epsilon \Theta = [0,1]$$

Example 3 Y has pdf given by, for $0 < y < \infty$

$$f(y,\theta_1,\theta_2) = \frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}} y^{\theta_1 - 1} e^{-y/\theta_2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1,\theta_2) : \theta_1, \theta_2 > 0\}$$

Note: $\Gamma(c) = \int_0^\infty y^{c-1} e^{-y} dy$ is referred to as the gamma function

The following are some special cases of family of distributions:

1. The parameter θ in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta) : \theta \epsilon \Theta\}$ is said to be a **Location Parameter** if the distribution of $W = Y - \theta$ does not depend on θ, that is, if the pdf of W, $f_W(w) = f_Y(w + \theta)$ does not depend on θ.
   - W is referred to as the **Standard** member of a location family if $\theta = 0$.

2. The parameter θ in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta) : \theta \epsilon \Theta\}$ is said to be a **Scale Parameter** if the distribution of $W = Y/\theta$ does not depend on θ, that is, if the pdf of W $f_W(w) = \theta f_Y(\theta w)$ does not depend on θ.
   - W is referred to as the **Standard** member of a scale family if $\theta = 1$.

3. The parameters θ_1 and θ_2 in a family of pdf's for the r.v. Y, $\{f_Y(y;\theta_1,\theta_2) : \theta \epsilon \Theta\}$ are said to be a **Location-Scale Parameters** if the distribution of $W = (Y - \theta_1)/\theta_2$ does not depend on θ_1 nor θ_2, that is, if the pdf of W $f_W(w) = \theta_2 f_Y(\theta_2 w + \theta_1)$ does not depend on θ_1 nor θ_2.
   - W is referred to as the **Standard** member of a location-scale family if $\theta_1 = 0$ and $\theta_2 = 1$.

The following examples will illustrate these types of families:

Example 1. Let Y have a $N(\theta, 1)$ distribution.

Then θ is a location parameter as demonstrated by

$$f(y, \theta, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} \quad \text{for} \quad \theta \epsilon (-\infty, \infty)$$

Let $W = Y - \theta \quad \Rightarrow \quad f_W(w) = f_Y(w + \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(w+\theta)-\theta]^2} = \frac{1}{\sqrt{2\pi}} e^{-.5[w]^2}$

The pdf of W does not depend on θ.

W is the standard member of the normal family of distributions, that is, W has a normal distribution with parameter values 0 and 1, that is, $N(0,\ 1)$.

Example 2. Let Y have an Exponential Distribution with parameter λ, that is,

$$f(y; \lambda) = \begin{cases} \frac{1}{\lambda} e^{-y/\lambda} & if \quad y \geq 0 \\ 0 & if \quad y < 0 \end{cases}$$

Let $W = Y/\lambda$ then the pdf of W is given by

$$f_W(w) = \lambda f(\lambda w) = \lambda \left(\frac{1}{\lambda} e^{-(\lambda w)/\lambda} \right) = e^{-w}$$

The pdf of W does not depend on λ.

Thus, λ is a scale parameter and W has an exponential distribution with $\lambda = 1$.

W is the standard member of the exponential family.

Example 3. Let Y have a $N(6, \theta^2)$ distribution. Is θ a scale parameter in this distribution?

$$f(y, 6, \theta) = \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2\theta^2}(y-6)^2} \quad \text{for} \quad \theta \epsilon (0, \infty)\}$$

Let W $= Y/\theta$ then the pdf of W is given by

$$f_W(w) = \theta f(\theta w) = \theta \left(\frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2\theta^2}[(\theta w)-6]^2} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta w - 6]^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta^2 w^2 - 12\theta w + 36]}$$

Thus, the pdf of W depends on θ and therefore θ is not a scale parameter. It would be a shape parameter.

If Y had a $N(0,\ \theta^2)$ distribution, would θ be a scale parameter in this distribution?

yes.

Example 4. Let Y have a $N(\theta_1, \theta_2^2)$ distribution. Are (θ_1, θ_2) location- scale parameters in this distribution?

$$f(y, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}(y-\theta_1)^2} \quad \text{for} \quad \theta \epsilon \Theta = \{(\theta_1, \theta_2) : \theta_1 \epsilon (-\infty, \infty), \theta_2 \epsilon (0, \infty)\}$$

Let W $= \frac{Y-\theta_1}{\theta_2}$ then the pdf of W is given by

$$f_W(w) = \theta_2 f(\theta_2 w + \theta_1) = \theta_2 \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}[(\theta_2 w + \theta_1) - \theta_1]^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\theta^2}[\theta_2^2 w^2]} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[w^2]}$$

The pdf of W does not depend on θ_1 nor θ_2. Therefore, θ_1 and θ_2 are location-scale parameters for the family of distributions for Y.

Is the location parameter of a distribution equal to the mean of the distribution?

Is the scale parameter of a distribution equal to the standard deviation of the distribution?

The answer is both yes and no:

1. For the normal distribution the location parameter is the mean and scale parameter is the standard deviation.
2. For the Cauchy distribution, the mean and standard deviation do not exist but the Cauchy has both a location and scale paramenter.
3. For the double exponential distribution, the location parameter is the mean but the standard deviation is $\sqrt{2}$ times the scale parameter
4. For the 3-parameter Weibull distribution, both the mean and standard deviation are functions of all three paramenters.

Why is it important to be able to designate a family of distributions as being a location/scale family?

A few important reasons are given next:

1. If $F(y, \theta_1, \theta_2)$ is a cdf with location/scale parameters (θ_1, θ_2), then to tabulate the probability distribution we only need a table for the standard member of the family, F^*, where F^* has $\theta_1 = 0, \theta_2 = 1$. For any other member of the family we can find its probabilities by using the following expression:

   Let Y be a random variable with cdf F a member of the family and Y^* be a random variable with cdf F^*, then

   $$Y^* = \frac{Y - \theta_1}{\theta_2} \quad \text{"equals in distribution"}$$

   $$F(y) = P[Y \leq y] = P\left[Y^* \leq \frac{y - \theta_1}{\theta_2}\right] = F^*\left(\frac{y - \theta_1}{\theta_2}\right)$$

   That is to find a probability associated with Y just look up the standardized value in the Y^* table.

   This is what is done with the normal distribution.

   Example: Suppose we have a normal distribution with $\mu = 5, \sigma = 2.3$ and we want to know what proportion of the population has values less than 1.7:

   $P[Y \leq 1.7] = P\left[Y^* \leq \frac{1.7-5}{2.3}\right] = F^*(-1.435) = pnorm(-1.435) = 0.0756$ using the R function, **pnorm**.

2. Similarly, we can determine a percentile for the general member of the family using the percentiles from the standard member of the family:

   Let $Q_Y(u)$ be the quantile function for Y which has location-scale parameters (θ_1, θ_2) and let $Q_{Y^*}(u)$ be the quantile function for the standard member of the family. Then,

   $$Q_Y(u) = \theta_1 + \theta_2 Q_{Y^*}(u)$$

   Example: To find the 95th percentile of a normal distribution with $\mu = 5, \sigma = 2.3$. Look up the 95th percentile in the standard normal table, $Q_{Y^*}(.95) = 1.645 = qnorm(.95)$ and then compute

   $Q_Y(.95) = \theta_1 + \theta_2 Q_{Y^*}(.95) = 5 + (2.3)(1.645) = 8.7835$

Alternatively, use the R function, **qnorm**, to obtain qnorm(.95,5,2.3) = 8.783163

There are many examples of distributions having parameters which are neither location nor scale parameters. The following tables (Cassela & Berger) and figures will illustrate such distributions:

Table of Common Distributions

Discrete Distributions

Bernoulli(p)

pmf	$P(X=x\|p) = p^x(1-p)^{1-x}; \quad x=0,1; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = p, \quad \mathrm{Var}\,X = p(1-p)$
mgf	$M_X(t) = (1-p) + pe^t$

Binomial(n,p)

pmf	$P(X=x\|n,p) = \binom{n}{x}p^x(1-p)^{n-x}; \quad x=0,1,2,\ldots,n; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = np, \quad \mathrm{Var}\,X = np(1-p)$
mgf	$M_X(t) = [pe^t + (1-p)]^n$
notes	Related to Binomial Theorem (Theorem 3.2.2). The *multinomial* distribution (Definition 4.6.2) is a multivariate version of the binomial distribution.

Discrete uniform

pmf	$P(X=x\|N) = \frac{1}{N}; \quad x=1,2,\ldots,N; \quad N=1,2,\ldots$
mean and variance	$\mathrm{E}X = \frac{N+1}{2}, \quad \mathrm{Var}\,X = \frac{(N+1)(N-1)}{12}$
mgf	$M_X(t) = \frac{1}{N}\sum_{i=1}^{N} e^{it}$

Geometric(p)

pmf	$P(X=x\|p) = p(1-p)^{x-1}; \quad x=1,2,\ldots; \quad 0 \le p \le 1$
mean and variance	$\mathrm{E}X = \frac{1}{p}, \quad \mathrm{Var}\,X = \frac{1-p}{p^2}$

mgf $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad t < -\log(1-p)$

notes $Y = X - 1$ is negative binomial$(1, p)$. The distribution is *memoryless*: $P(X > s|X > t) = P(X > s - t)$.

Hypergeometric

pmf $P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0, 1, 2, \ldots, K;$
$M - (N - K) \leq x \leq M; \quad N, M, K \geq 0$

mean and variance $\mathrm{E}X = \frac{KM}{N}, \quad \mathrm{Var}\, X = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$

notes If $K \ll M$ and N, the range $x = 0, 1, 2, \ldots, K$ will be appropriate.

Negative binomial(r, p)

pmf $P(X = x|r, p) = \binom{r+x-1}{x}p^r(1-p)^x; \quad x = 0, 1, \ldots; \quad 0 \leq p \leq 1$

mean and variance $\mathrm{E}X = \frac{r(1-p)}{p}, \quad \mathrm{Var}\, X = \frac{r(1-p)}{p^2}$

mgf $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t < -\log(1-p)$

notes An alternate form of the pmf is given by $P(Y = y|r, p) = \binom{y-1}{r-1}p^r(1-p)^{y-r}$, $y = r, r+1, \ldots$. The random variable $Y = X + r$. The negative binomial can be derived as a gamma mixture of Poissons. (See Exercise 4.32.)

Poisson(λ)

pmf $P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, \ldots; \quad 0 \leq \lambda < \infty$

mean and variance $\mathrm{E}X = \lambda, \quad \mathrm{Var}\, X = \lambda$

mgf $M_X(t) = e^{\lambda(e^t-1)}$

Discrete Distributions - Examples

1. **Discrete Uniform R.V.** - Possible values: $1, 2, \ldots, N$ with equal probability, $\frac{1}{N}$

- Used in computing the odds in sporting events and card games
- During WWII, the Allied forces wanted to estimate the number of tanks placed in combat by the Germans. Because the tanks were consecutively numbered, the tank numbers in any given area formed a discrete uniform distribution. The Allied forces had a sample of such numbers from which they could estimate the total number of German tanks in a given region.

2. **Bernoulli R.V.** - Possible values: 0 (failure) and 1 (success), with probabilities $1 - p$ and p

- Inspector determines if part is defective or good
- Patient has the disease or not
- Gene is mutated or not

3. **Binomial R.V.** - Possible values: $0, 1, 2, \ldots, n$ with probabilities:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n$$

- The number of successes, S, in a fixed number, n, of independent Bernoulli Trials with identical probabilities, p for each trial. S is distributed $B(n, p)$
- A circuit board has 25 individual components with a .01% probability of an individual component being defective and the events that components are defective being independent events. Let C be the number of defective components on a randomly selected circuit board.
- A veterinarian inspects 50 deer for ticks. Let T be the number of deer having ticks.

4. **Binomial R.V.** - The number of Type A items, X, in a random sample of n units selected **with replacement** from a population containing N units consisting of two types of Units - Type A and Type B (not Type A).

- Let M be the number of Type A units in the population. Then $p = M/N$ in the Binomial pmf
- Sampling with replacement does not occur very often in practice but statisticians use sampling with replacement in approximating sampling distributions.

5. **Negative Binomial R.V.** - The number of trials, B, until the rth success in series of independent identically distributed (i.i.d.) Bernoulli trials

   Possible values: $r, r+1, r+2, \cdots$ with probabilities: $P(B=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$, for $k = r, r+1, \ldots$

- Let D be the number of patients examined until the 5th patient with a rare eye disease is found

- In a series of iid Bernoulli trials, the Binomial distribution has a fixed number of trials and a random number of successes

- In a series of iid Bernoulli trials, the Negative Binomial distribution has a fixed number of successes and a random number of trials

6. **Geometric R.V.** - The number of trials until the first success in series of i.i.d. Bernoulli trials

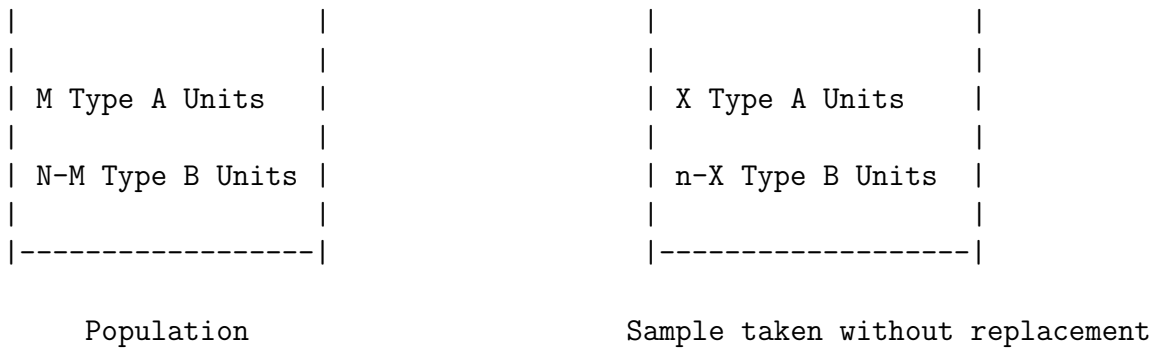
- Special case of Negative Binomial with $r = 1$

7. **Hypergeometric R.V.** - The number of Type A items, X in a random sample of n units selected **without replacement** from a population containing N units consisting of M Type A Units and $N - M$ Type B Units.

- Possible values: $k = 0, 1, 2, \ldots, n$ provided $M - (N - n) \leq k \leq M$ with probabilities:

  $P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ for $k = 0, 1, 2, \ldots, n$ and $M - (N - n) \leq k \leq M$

- Let X be the number of Type A units in a random sample of n units from the population of N units, sampling without replacement. The possible values of X are the integers between $max(0, M - (N - n))$ and $min(n, M)$



8. **Poisson R.V.** - The number of event occurrences, Y during a specified period of time or space of a Poisson process

   Poisson Postulates: The events occur according the following set of conditions:

   P1. The probability that exactly one event occurs in a very short period of time of length Δt is $\lambda \cdot \Delta t + o(\Delta t)$, where $\lambda > 0$ is a fixed constant

   P2. The probability of more than one event occurring during Δt is $o(\Delta t)$

   P3. The number of events occurring during the time interval Δt is independent of the number occurring in any other time interval.

- A more mathematical formulation of the Poisson postulates is given in the Casela-Berger book.
- λ is the average number of event occurrences in a unit period of time
- The Poisson postulates can be formulated in space as well as time where Δt is a unit of space in place of a unit of time.
- - Possible values: $0, 1, 2, \ldots$ with probabilities:

  $$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \text{ for } y = 0, 1, 2, \ldots$$

- Suppose the number of requests for assistance arrive at a computer service desk according to the Poisson postulates with $\lambda = 5$ per second. Let S be the number of requests in a given hour.
- Let F be the number of persons who are killed in plane accidents during a given year
- Let P be the number of lead particles in a square inch of painted wall in an old building
- The Poisson distribution is the limit of a Binomial distribution as n approaches infinity and p_n approaches 0 in such a manner that np_n approaches λ in the limit.
- Both the negative binomial and Poisson distributions are used to model R.V.'s in studies which do not exactly follow the about definitions: For example, we may randomly select plants in a field treated with an insecticide and count the number of insects on each of the plants. In a future handout, we will discuss how to measure how well a given distribution fits a given data set. *No model is correct but many are useful.* Quote from Dr. George Box.

The interrelationships between the various discrete distributions are given in the following display.

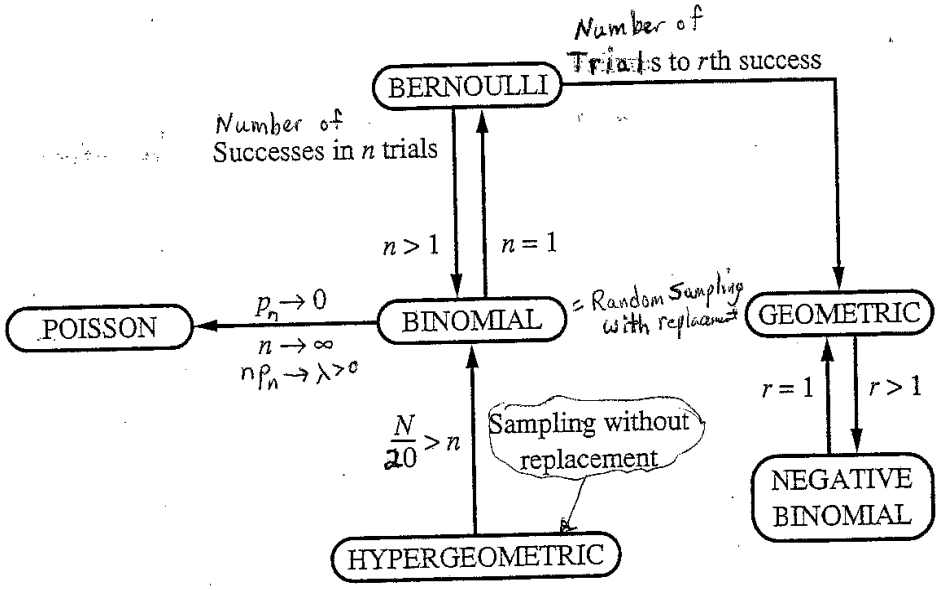


FIGURE
Relationships between the distributions
Statistical Analysis for Engineers/Scientists
J. Wesley Barnes

START: Class Notes 9/10/21:

Examples of Continuous Distributions

Beta(α, β)

pdf	$f(x\|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \quad \alpha > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
mgf	$M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$
notes	The constant in the beta pdf can be defined in terms of gamma functions, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Equation (3.2.18) gives a general expression for the moments.

Cauchy(θ, σ)

pdf	$f(x\|\theta,\sigma) = \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty; \quad -\infty < \theta < \infty, \quad \sigma > 0$
mean and variance	do not exist
mgf	does not exist
notes	Special case of Student's t, when degrees of freedom $= 1$. Also, if X and Y are independent n$(0,1)$, X/Y is Cauchy.

Chi squared(p)

pdf	$f(x\|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}; \quad 0 \le x < \infty; \quad p = 1, 2, \ldots$
mean and variance	$\mathrm{E}X = p, \quad \mathrm{Var}\, X = 2p$
mgf	$M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \quad t < \frac{1}{2}$
notes	Special case of the gamma distribution.

Double exponential(μ, σ)

pdf	$f(x\|\mu,\sigma) = \frac{1}{2\sigma} e^{-\|x-\mu\|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$
mean and variance	$\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = 2\sigma^2$
mgf	$M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2}, \quad \|t\| < \frac{1}{\sigma}$
notes	Also known as the *Laplace* distribution.

Exponential(β)

pdf $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \beta, \quad \mathrm{Var}\, X = \beta^2$

mgf $M_X(t) = \frac{1}{1-\beta t}, \quad t < \frac{1}{\beta}$

notes Special case of the gamma distribution. Has the *memoryless* property. Has many special cases: $Y = X^{1/\gamma}$ is *Weibull*, $Y = \sqrt{2X/\beta}$ is *Rayleigh*, $Y = \alpha - \gamma \log(X/\beta)$ is *Gumbel*.

F

pdf $f(x|\nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}};$
$0 \le x < \infty; \quad \nu_1, \nu_2 = 1, \ldots$

mean and variance $\mathrm{E}X = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2,$

$\mathrm{Var}\, X = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4$

moments (mgf does not exist) $\mathrm{E}X^n = \frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_2}{\nu_1}\right)^n, \quad n < \frac{\nu_2}{2}$

notes Related to chi squared ($F_{\nu_1,\nu_2} = \left(\frac{\chi^2_{\nu_1}}{\nu_1}\right) / \left(\frac{\chi^2_{\nu_2}}{\nu_2}\right)$, where the χ^2s are independent) and t ($F_{1,\nu} = t^2_\nu$).

Gamma(α, β)

pdf $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha, \beta > 0$

mean and variance $\mathrm{E}X = \alpha\beta, \quad \mathrm{Var}\, X = \alpha\beta^2$

mgf $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \quad t < \frac{1}{\beta}$

notes Some special cases are exponential ($\alpha = 1$) and chi squared ($\alpha = p/2$, $\beta = 2$). If $\alpha = \frac{3}{2}$, $Y = \sqrt{X/\beta}$ is *Maxwell*. $Y = 1/X$ has the *inverted gamma distribution*. Can also be related to the Poisson (Example 3.2.1).

Logistic(μ, β)

pdf $f(x|\mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \beta > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = \frac{\pi^2\beta^2}{3}$

mgf $M_X(t) = e^{\mu t}\Gamma(1-\beta t)\Gamma(1+\beta t), \quad |t| < \frac{1}{\beta}$

notes The cdf is given by $F(x|\mu,\beta) = \frac{1}{1+e^{-(x-\mu)/\beta}}$.

Lognormal(μ, σ^2)

pdf $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}, \quad 0 \le x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean and variance $\mathrm{E}X = e^{\mu+(\sigma^2/2)}, \quad \mathrm{Var}\, X = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$

moments (mgf does not exist) $\mathrm{E}X^n = e^{n\mu+n^2\sigma^2/2}$

notes Example 2.3.5 gives another distribution with the same moments.

Normal(μ, σ^2)

pdf $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean and variance $\mathrm{E}X = \mu, \quad \mathrm{Var}\, X = \sigma^2$

mgf $M_X(t) = e^{\mu t+\sigma^2 t^2/2}$

notes Sometimes called the *Gaussian* distribution.

Pareto(α, β)

pdf $f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad a < x < \infty, \quad \alpha > 0, \quad \beta > 0$

mean and variance $\mathrm{E}X = \frac{\beta\alpha}{\beta-1}, \quad \beta > 1, \quad \mathrm{Var}\, X = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2$

mgf does not exist

---

t

pdf $f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{\left(1+\left(\frac{x^2}{\nu}\right)\right)^{(\nu+1)/2}}, \quad -\infty < x < \infty, \quad \nu = 1, \ldots$

mean and variance $\mathrm{E}X = 0, \quad \nu > 1, \quad \mathrm{Var}\, X = \frac{\nu}{\nu-2}, \quad \nu > 2$

moments (mgf does not exist) $\mathrm{E}X^n = \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{\nu-n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\nu^{n/2}$ if $n < \nu$ and even,
$\mathrm{E}X^n = 0$ if $n < \nu$ and odd.

notes Related to F ($F_{1,\nu} = t_\nu^2$).

Uniform(a, b)

pdf	$f(x\|a,b) = \frac{1}{b-a}, \quad a \le x \le b$
mean and variance	$\mathrm{E}X = \frac{b+a}{2}, \quad \mathrm{Var}\,X = \frac{(b-a)^2}{12}$
mgf	$M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$
notes	If $a = 0$ and $b = 1$, this is a special case of the beta ($\alpha = \beta = 1$).

Weibull(γ, β)

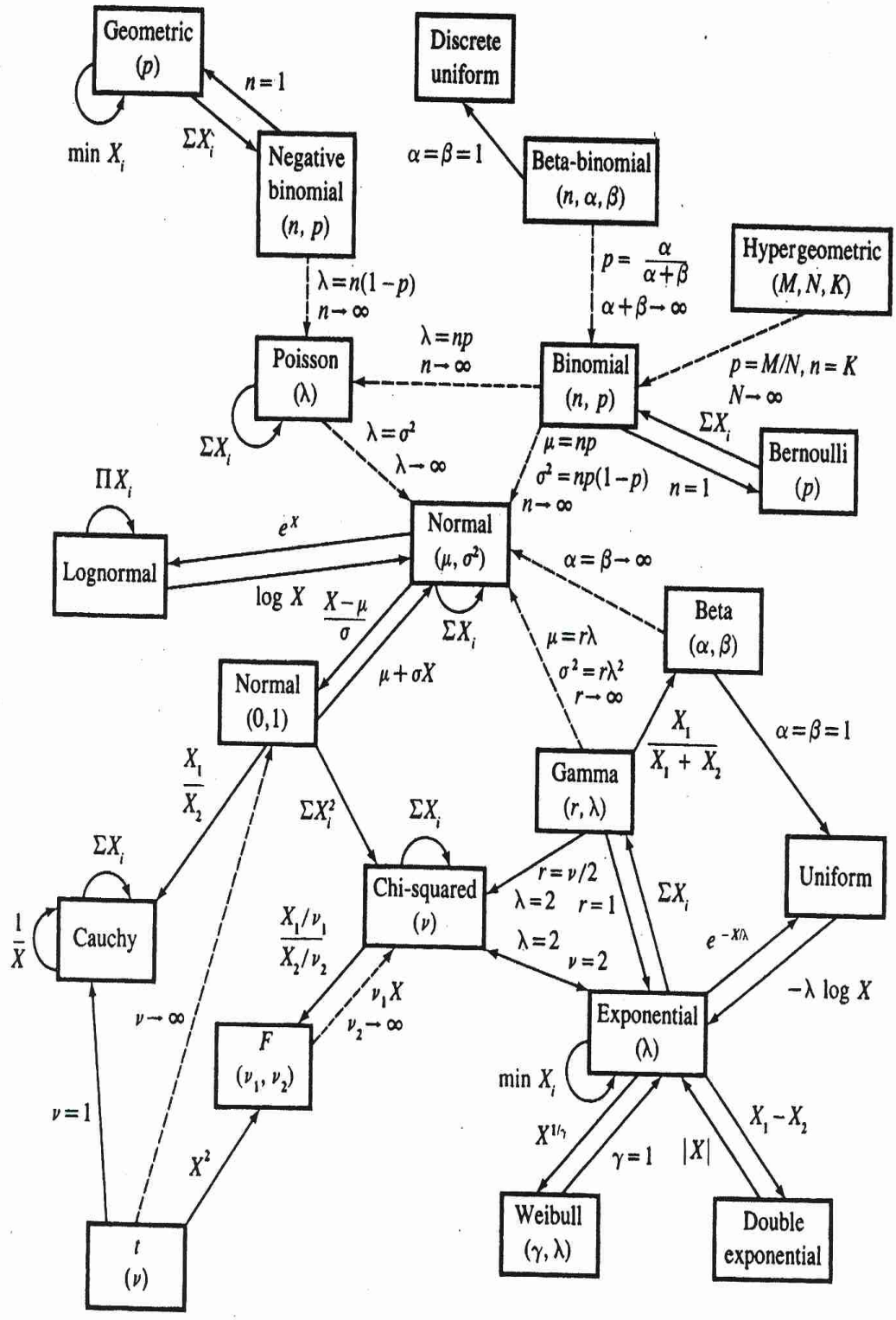
pdf	$f(x\|\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta}, \quad 0 \le x < \infty, \quad \gamma > 0, \quad \beta > 0$
mean and variance	$\mathrm{E}X = \beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right), \quad \mathrm{Var}\,X = \beta^{2/\gamma}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$
moments	$\mathrm{E}X^n = \beta^{n/\gamma}\Gamma\left(1+\frac{n}{\gamma}\right)$
notes	The mgf exists only for $\gamma \ge 1$. Its form is not very useful. A special case is exponential ($\gamma = 1$).

Note that the Weibull distribution is often expressed with alternative parameters as

$$f(y|\gamma,\alpha) = \frac{\gamma}{\alpha}\left(\frac{y}{\alpha}\right)^{\gamma-1} e^{-(y/\alpha)^\gamma} \quad \text{for} \quad y \ge 0$$

That is, $\beta = \alpha^\gamma$

This is the form used in R and SAS.



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

From Casella-Berger, "Statistical Inference"

Continuous Distributions

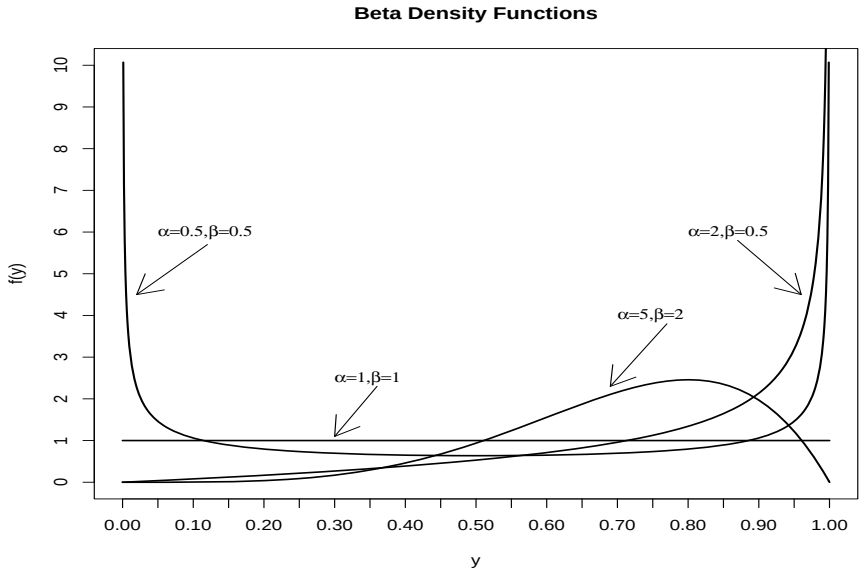
1. **Uniform**(a, b) **R.V.** - probability outcome of r.v. Y is in interval (c, d) is proportional to the length of the interval.

- Symmetric distribution on the interval (a, b)
- Used in simulations and modelling events which are equally likely

← good when modeling probabilities.

2. **Beta**(α, β) **R.V.** - Generalization of the Uniform on $[0, 1]$ R.V.

- Can be symmetric, right-skewed, or left-skewed depending on the values of α and β
- Distribution is Uniform on $[0, 1]$ if $\alpha = \beta = 1$
- Let p be the proportion of defectives in the plant producing a given product, where a company has many plants. A model for p could be the Beta
- Model for R.V.s with bounded realizations. Let Z have a Beta distribution on $[0, 1]$ and let $Y = (b - a)Z + a$. Then Y has a Beta distribution on the interval $[a, b]$



3. **Normal**(μ, σ^2) **R.V.** - Also referred to as a Gaussian R.V.

- Used to model r.v.s with symmetric distributions having tails of moderate weight, not too many extreme values
- Used in many statistical applications due to the Central Limit Theorem

4. **Chi-squared**(ν) **R.V.**

- Right skewed distribution which becomes more symmetric (normal like) as ν becomes large
- Special case of the Gamma and Weibull R.V.s
- parameter ν is referred as degrees of freedom
- Related to the standard normal$(\mu = 0, \sigma = 1)$ distribution through the relationship:

  If $Z_1, Z_2, \ldots, Z_k$ are i.i.d. $N(0, 1)$ r.v.s, then $Y = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ has a Chi-squared distribution with $\nu = k$.

5. **Student t**(ν) **R.V.** - Generally just referred to as the t-distribution

- Symmetric distribution with tails heavier than standard normal
- Converges to a standard normal as ν converges to ∞
- parameter ν is referred as degrees of freedom
- Used to model populations in which extreme values occur more frequently than would be expected in a normal distribution, for example, financial data
- Related to the Chi-squared and standard normal distribution:

  Let Z have a $N(0, 1)$ distribution, W have a Chi-squared distribution with parameter ν with Z and W independent. Then the r.v. $Y = \frac{Z}{\sqrt{W/\nu}}$ has a t-distribution with parameter ν

6. **Fisher**(ν_1, ν_2) **R.V.** - Generally referred to as an F with degrees of freedom, df $= (\nu_1, \nu_2)$

- Right skewed distribution
- Most often used as a test statistic in ANOVA and Regression
- Related to the Chi-squared distribution:

  Let W_1 have a Chi-squared distribution with parameter ν_1, W_2 have a Chi-squared distribution with parameter ν_2, with W_1 and W_2 independent. Then the r.v. $Y = \frac{W_1/\nu_1}{W_2/\nu_2}$ has an F-distribution with parameters ν_1, ν_2

- If T has a t-distribution with parameter ν then T^2 has an F-distribution with parameters $\nu_1 = 1, \nu_2 = \nu$

β_1 = skewness
β_2 = kurtosis

df	μ	σ^2	β_1	β_2
1	—	—	—	—
2	0	—	—	—
5	0	1.7	0	9
9	0	1.3	0	4.2
12	0	1.2	0	3.75
20	0	1.1	0	3.38
25	0	1.09	0	3.29
30	0	1.07	0	3.23
∞	0	1.00	0	3

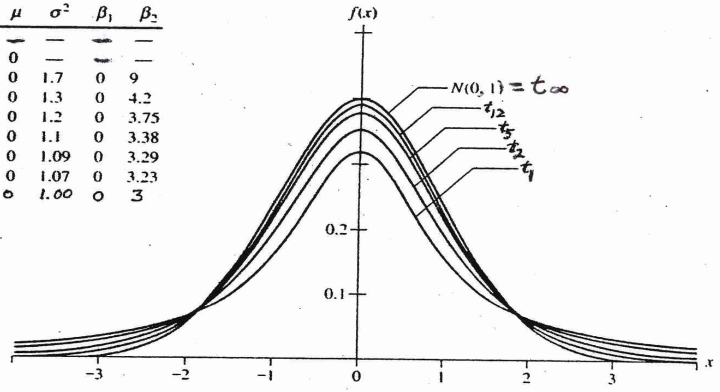
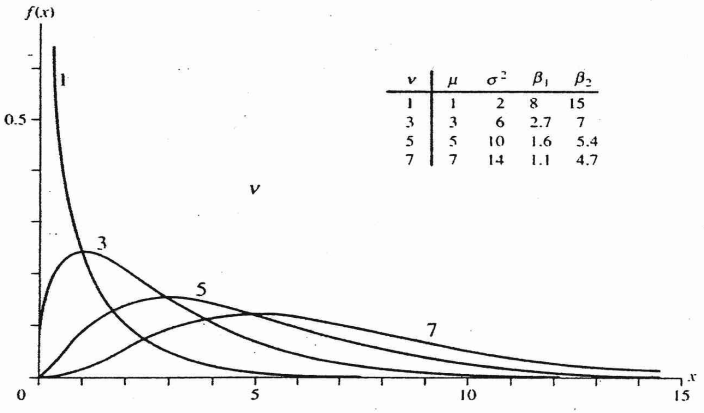


FIGURE 4-6
The relationship between the standard normal distribution and the t distribution.



ν	μ	σ^2	β_1	β_2
1	1	2	8	15
3	3	6	2.7	7
5	5	10	1.6	5.4
7	7	14	1.1	4.7

FIGURE 4-5
Representative χ^2 distribution plots with low degrees of freedom.

n	m	μ	σ^2	β_1	β_2
1	10	1.25	4.69	33.3	109
3	6	1.50	5.25	—	—
5	10	1.25	1.35	14.9	53.9
10	20	1.11	0.43	3.37	9.89

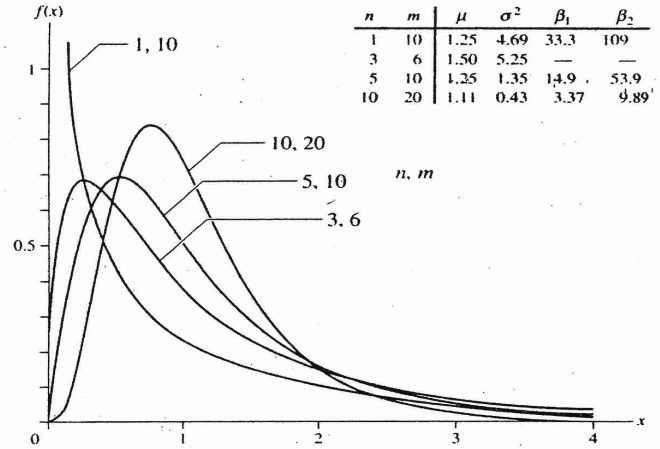
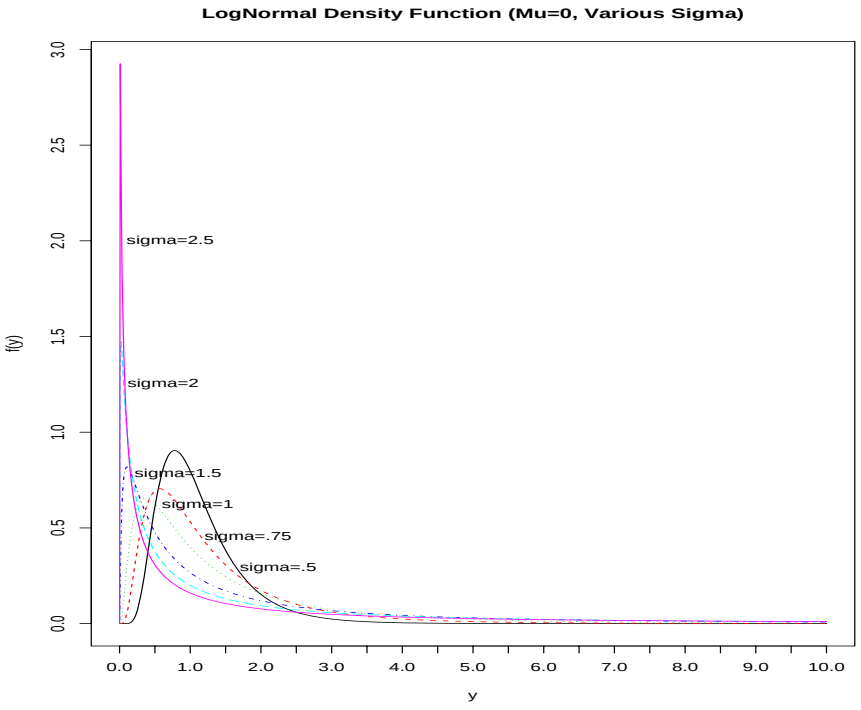


FIGURE 4-7
Plots of selected F distributions.

7. **Lognormal**(μ, σ^2) **R.V.**

- Right skewed distribution
- Used to model the growth of plants, tumors, and in reliability, the time to failure of a device or the time until a tumor is no longer detectable
- Related to the normal distribution:

  Let Y have a $N(\mu, \sigma^2)$ distribution then $X = e^Y$ has a lognormal distribution, that is, $log(X)$ has a normal distribution



8. **Cauchy**(θ_1, θ_2) **R.V.**

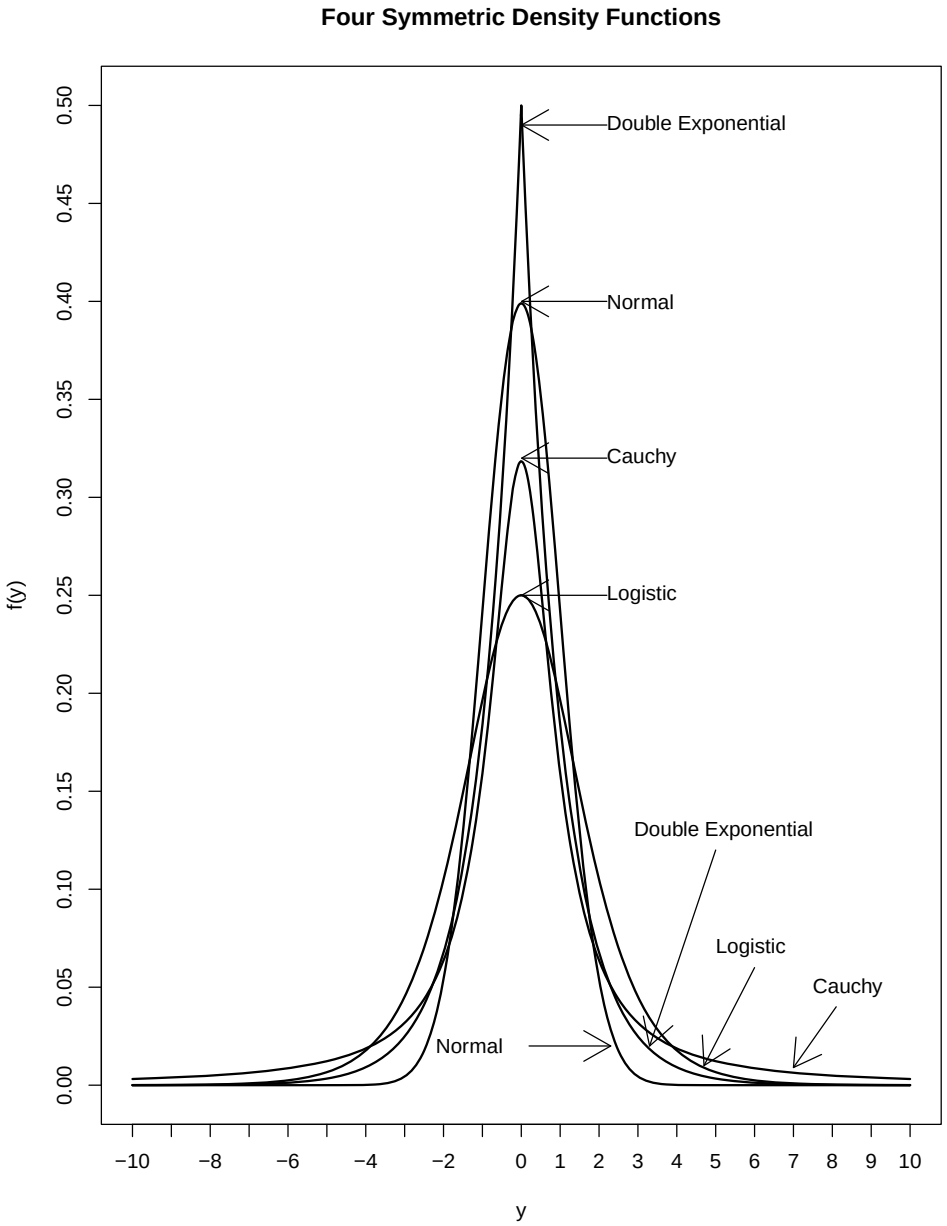
- Symmetric distribution with very heavy tails, so heavy that its mean and variance do not exist
- Models populations in which very large, relative to the point of symmetry, θ_1, occur much more frequently than would be expected in a normal distribution - financial models
- If T has a Student t-distribution with $\nu = 1$ then T has a Standard Cauchy distribution, $(\theta_1 = 0, \theta_2 = 1)$
- If Z_1 and Z_2 are independent $N(0, 1)$ r.v.s then $Y = Z_1/Z_2$ has a standard Cauchy distribution

9. **Double Exponential**(θ_1, θ_2) **R.V.** - Also, referred to as Laplace R.V.

- Symmetric distribution with a sharp peak and tails heavier than the normal distribution but lighter than Cauchy

10. **Logistic**(θ_1, θ_2) **R.V.**

- Symmetric distribution with tail weights in between the double exponential and Cauchy



11. **Exponential**(β) **R.V.**

- In a Poisson process with λ being the average number of occurrences in a unit of time, let T be the time between occurrence of 2 events. T has an exponential distribution with $\beta = 1/\lambda$
- Used in reliability or survival analysis to model the time until failure or death when the time to failure/death has a **memoryless property**, that is, given that the device has survived until at least time y, the probability that the device will function an additional t time units is equal to the probability that the device will survive until time t,

  $P[T \geq t + y | T \geq y] = P[T \geq t]$

  Also, referred as a constant failure rate

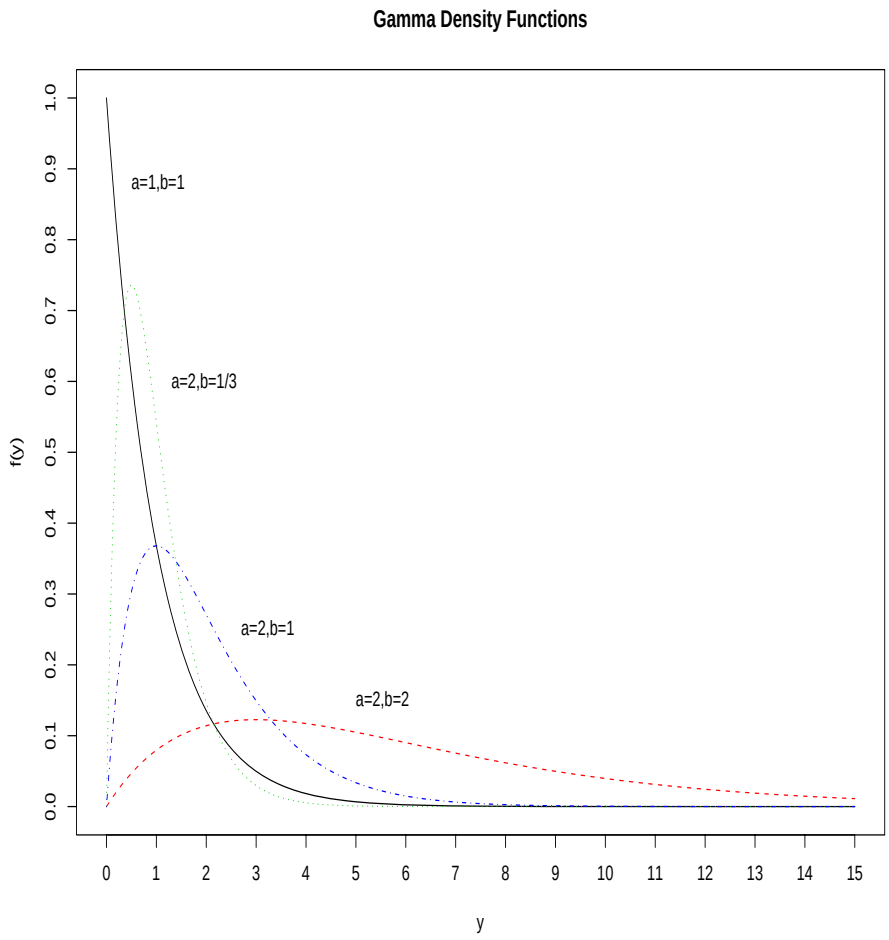
  The failure times of most devices/human subjects do not have the memoryless property over their total lifetime but may have over a limited range

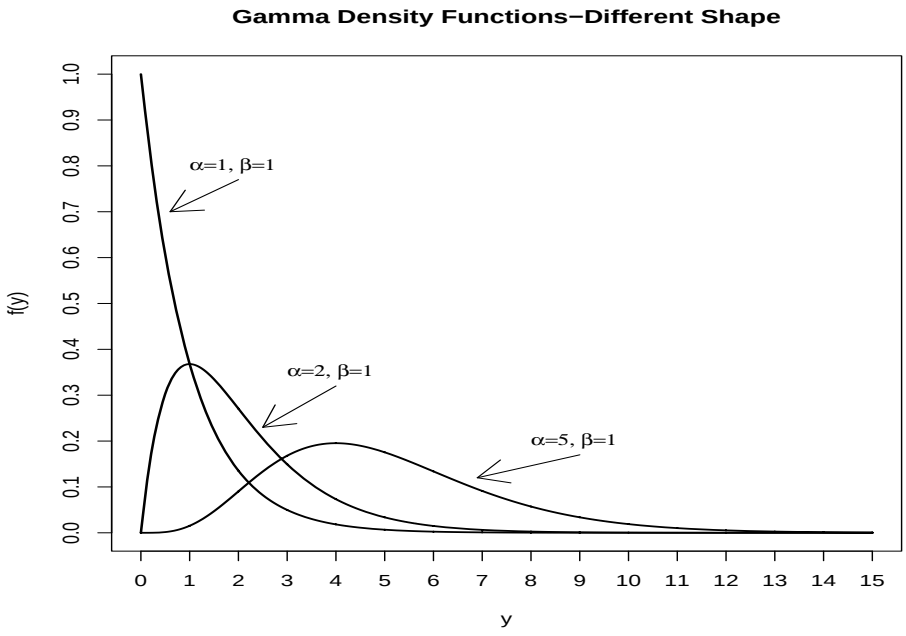
12. **Gamma**(α, β) **R.V.**

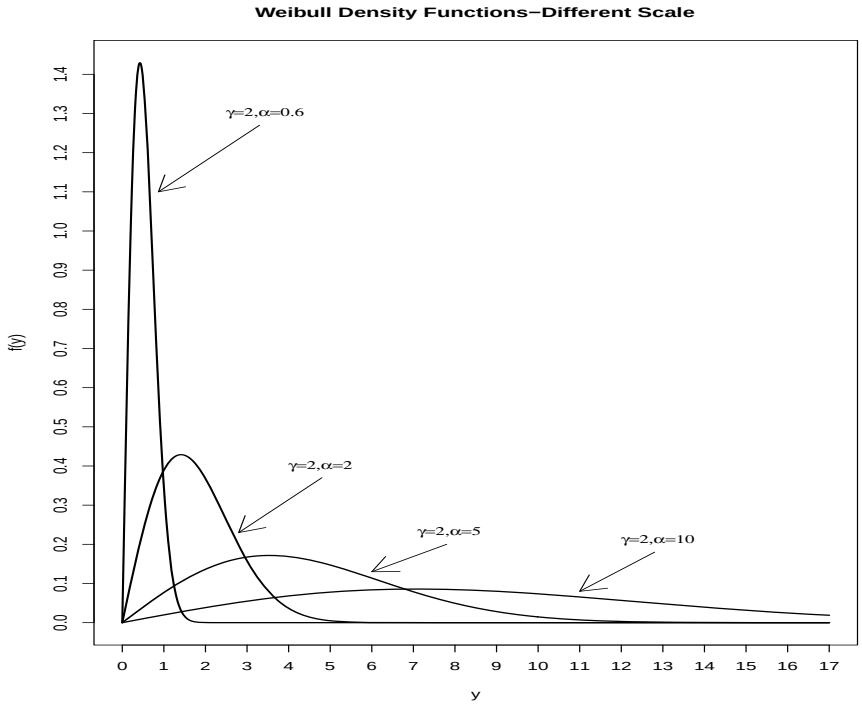
- Let T be the time between k events in a Poisson process with parameter λ then T has a Gamma distribution with $\alpha = k, \beta = 1/\lambda$
- It follows that if $E_1, E_2, \ldots, E_k$ are i.i.d. Exponential r.v.s with parameter β then $T = E_1 + E_2 + \ldots + E_k$ has a Gamma distribution with parameters $\alpha = k, \beta$. This form of the Gamma distribution is referred to as the Erlang distribution
- Exponential(β) distribution is a Gamma($\alpha = 1, \beta$) distribution
- Chi-squared(ν) distribution is a Gamma($\alpha = \nu/2, \beta = 2$) distribution
- If Y has a Gamma(α, β) distribution then $W = \sqrt{Y/\beta}$ has a Maxwell distribution which is used to model particle speeds in gases under special conditions.

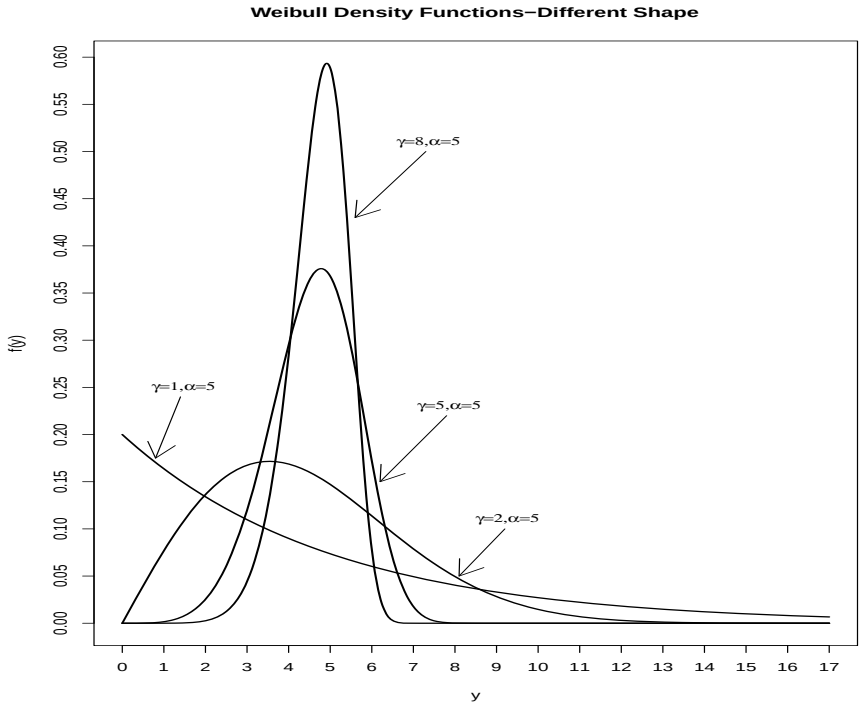
13. **Weibull**(γ, β) **R.V.**

- Generalization of the exponential distribution so that the failure rate is time raised to a power
- Let Y have an exponential(β) distribution and let $X = Y^{1/\gamma}$ then X has a Weibull(γ, β) distribution
- Exponential(β) is a special case of Weibull($\gamma = 1, \beta$)
- The Weibull distribution is a special case of a class of distributions called the Extreme Value Distributions used to model the extreme observations in a populations, either minimums or maximums. For example, largest crack in a 40 feet long pipe, shortest time to death of 100 ticks exposed to an environment of high temperature/low humidity
- Alternative parametrization of the Weibull has parameters (α, γ) with $\alpha = (\beta)^{1/\gamma}$ Need to check software to determine which parametrization they are using.


Gamma Density Functions
a=1,b=1
a=2,b=1/3
a=2,b=1
a=2,b=2
f(y)
y


Gamma Density Functions−Different Shape
α=1, β=1
α=2, β=1
α=5, β=1
f(y)
y


Weibull Density Functions−Different Scale
γ=2,α=0.6
γ=2,α=2
γ=2,α=5
γ=2,α=10
f(y)
y


Weibull Density Functions−Different Shape
γ=8,α=5
γ=1,α=5
γ=5,α=5
γ=2,α=5
f(y)
y

Mixture Distribution

In a number of situations neither a discrete nor a continuous distribution will adequately model the population/process. The population/process is a combination of the realizations from several discrete and/or continuous distributions. We model such situations using Mixture Distributions. In other situations, we may have several populations/processes producing the observed data.

The following examples will illustrate these situations.

Example 1: A central warehouse receives the output from 5 production facilities, e.g., Firestone Tires. The tires are inspected and D_{ij} is the deviation from the specified adhesive strength of Tire j from Facility i. Suppose D_{ij} has a $N(\mu_i, \sigma_i^2)$ distribution, that is, the distribution of D may be differ from facility to facility. Let p_i be the proportion tires in the warehouse from Facility i with $\sum_{i=1}^{5} p_i = 1$. Let X be the measurement obtained from a randomly selected tire in the central warehouse. What is the distribution of D?

Example 2: The summer ozone level data in Houston where the data is the combined data from 10 detection devices placed within 1000 meters of 10 different chemical plants.

In general, if the population of interest is a combination of the values from k distinct populations, where Population i has pdf f_i and cdf F_i, then the Mixture Population has cdf and pdf given by

$$F(x) = \sum_{i=1}^{k} p_i F_i(x) \qquad f(x) = \sum_{i=1}^{k} p_i f_i(x) \quad \text{with} \quad \sum_{i=1}^{k} p_i = 1$$

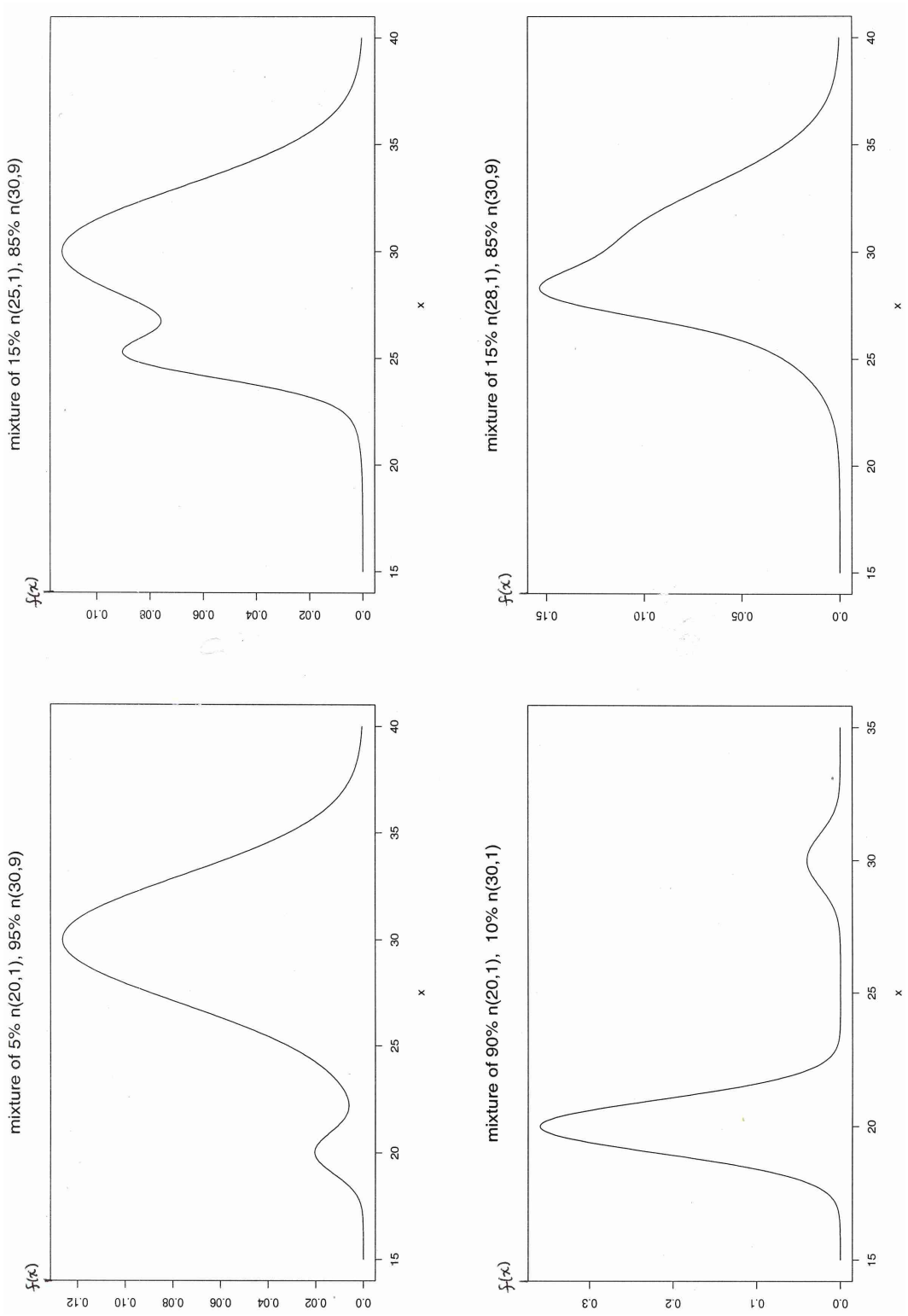
The graphs on the following pages illustrate mixtures of two normal populations.

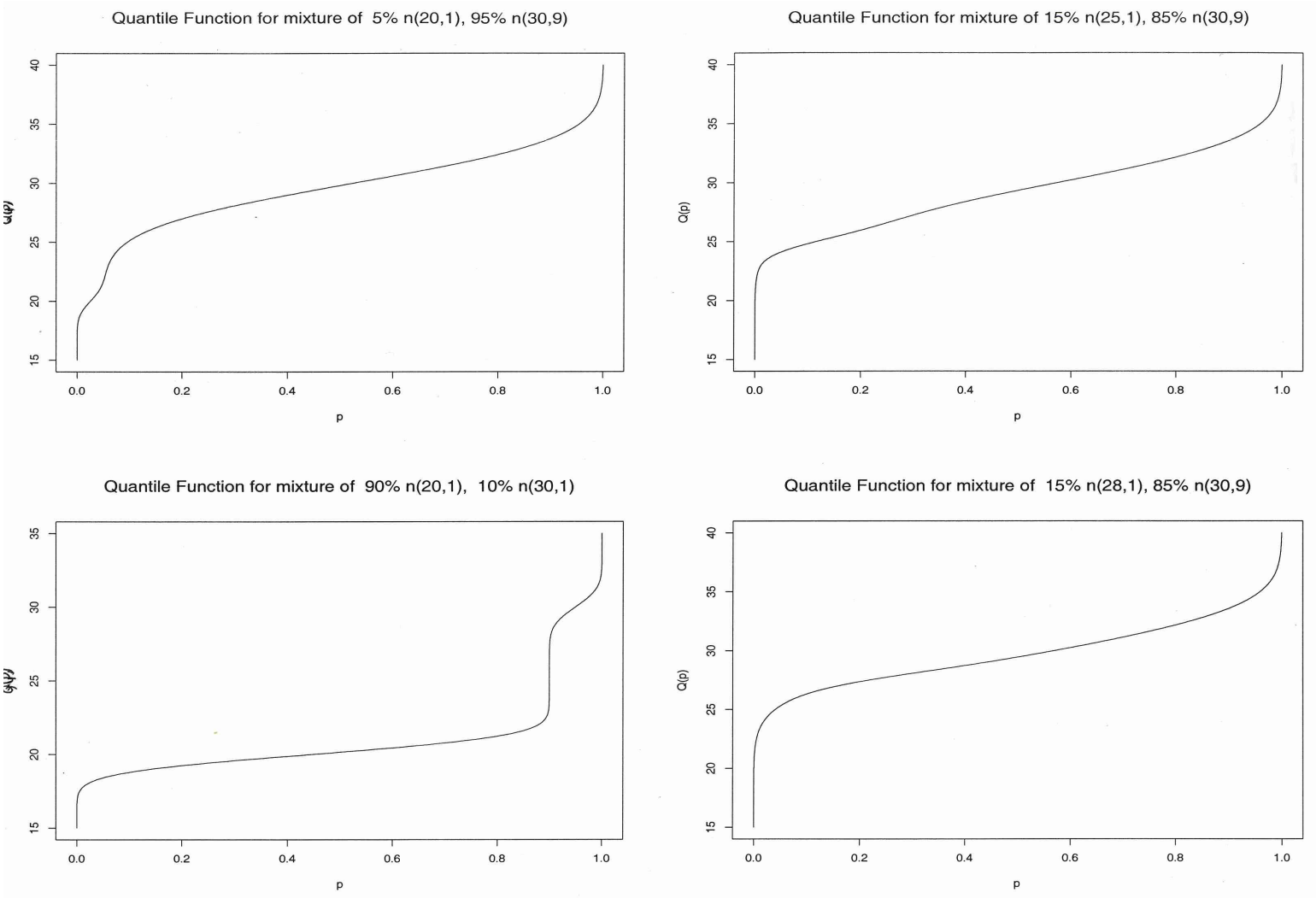
For the situation where we have a random variable, Y, which has probability p of equalling 0 and a continuous cdf, F^* on $(0, \infty)$, then the cdf of Y is

$$F(y) = pI(y \geq 0) + (1 - p)F^*(y)$$

Example: The National Oceanic and Atmospheric Administration (NOAA) wants to determine the amount of game fish caught in shrimp nets. Let Y be the Catch per Unit Effort (CPUE) for a shrimp boat whose nets have been in the water H hours. That is, $Y = N/H$, where N is the total number of game fish caught in a net which was H hours in the water. A proportion of nets, say 20%, will have $N = 0$ and hence $Y = 0$ whereas the values with $Y > 0$ will have a right skewed continuous distribution, such as the log-normal distribution. The cdf of Y is given by

$$F(y) = .2I(y \geq 0) + .8F^*(y), \quad \text{where } F^* \text{ is the cdf of log-normal distribution}$$


mixture of 5% n(20,1), 95% n(30,9)
mixture of 15% n(25,1), 85% n(30,9)
mixture of 90% n(20,1), 10% n(30,1)
mixture of 15% n(28,1), 85% n(30,9)
f(x)
x


Quantile Function for mixture of 5% n(20,1), 95% n(30,9)
Quantile Function for mixture of 15% n(25,1), 85% n(30,9)
Quantile Function for mixture of 90% n(20,1), 10% n(30,1)
Quantile Function for mixture of 15% n(28,1), 85% n(30,9)
Q(p)
p
15
20
25
30
35
40
0.0
0.2
0.4
0.6
0.8
1.0

Simulation of Observations from Specified Distributions

Simulate Observation from Strictly Increasing Continuous cdf F

Let Y have a strictly increasing continuous cdf F_Y. Then,

the quantile function of Y, $Q_Y(u) = F^{-1}(u)$ is a well defined function.

Let U have a Uniform on (0,1) distribution and define the r.v. $W = Q_Y(U)$.

The cdf for U is

$F_U(u) = u$ for $0 \leq u \leq 1$ and $F_U(u) = 0$ for $u < 0$; $F_U(u) = 1$ for $u > 1$

Claim: W has cdf $F_W(w) = F_Y(w)$ for all w.

$$F_W(w) = P[W \leq w] = P[Q_Y(U) \leq w] = P[F_Y(Q_Y(U)) \leq F_Y(w)] = P[U \leq F_Y(w)] = F_Y(w)$$

That is, W is a realization from the distribution of Y.

Thus, we only need a method for generating observations from the Uniform on (0,1) distribution in order to obtain realizations from any continuous distribution that has $Q(u) = F^{-1}(u)$ expressed in a closed form.

Example:

Generate 1000 observations from an Exponential distribution with $\beta = 4$.

$$F_Y(y) = 1 - e^{-y/4} \text{ for } y \geq 0$$

1. Find $Q_Y \quad \Rightarrow \quad u = F(y_u) = 1 - e^{-y_u/4}$
2. Solve for $y_u = -4log(1-u) = Q_Y(u)$
3. Generate observation from Uniform on (0, 1) distributions
   Using R: U = runif(1) = .27
4. The observation from an Exp(4) distribution would be
   $Y = Q_Y(.27) = -4log(1-.27) = 1.259$
5. Repeat steps 3. and 4. 1000 times

The above method of generating random observations only works when the cdf and hence quantile function can be expressed in a closed form. For example, the normal distribution and gamma distribution do not have cdf with a close form. They can only be expressed as indefinite integrals:

$$F(t) = \int_0^t \frac{1}{\Gamma(\alpha)\beta^\alpha} y^t e^{-y/\beta} \mathrm{d}y$$

For a particular value of t, the integral must be solved numerically. A similar problem occurs with the normal cdf.

Special algorithms exist for generating observations from these distributions. For example, Cassella-Berger have an algorithm for generating 2 independent observations from a $N(\mu, \sigma^2)$ distribution.

1. Generate 2 observations from a $U_{(0,1)}$ distribution: U_1, U_2
2. Let $R = \sqrt{-2log(U_1)}$
3. Let $Z_1 = R\cos(2\pi U_2)$ and $Z_1 = R\sin(2\pi U_2)$
   Z_1 and Z_2 have independent $N(0,1)$ distributions
4. Let $Y_1 = \mu + \sigma Z_1$ and $Y_2 = \mu + \sigma Z_2$
   Y_1 and Y_2 have independent $N(\mu, \sigma^2)$ distributions

B: we use this algorithm to generate a few vars $Z_1, \ldots, Z_5 \sim N(0,1)$. Once we have a generator for $N(0,1)$ then we can easily generate values from lots of different distributions:

Ex. 1.) Generate obs from chi-squared w/ 4 DOF.

$$\chi^2_4 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$$

2.) Generate a value from t-dist w/ 4 DOF

$$t_4 = \frac{Z_5}{\sqrt{\chi^2_4/4}}$$

START: Class Notes 9/13/21

Simulation from Discrete cdf F p: want to sim from Poisson($\lambda = .5$) dist

Let D have a discrete distribution with cdf F and pmf

$$f(d_i) = p_i \quad \text{for} \quad d_1 < d_2 < \cdots < d_k$$

$$F(d) = \sum_{d_i \le d} p_i$$

Alt Def of Quantile function

Generate an observation, U, from a Uniform (0,1) distribution.

To obtain an observation on D, let $D = F^{-1}(U) = inf\{d : F(d) \ge U\}$

smallest value of d s.t. the CDF F(d) exceeds U; i.e. $F(d) \ge U$

$$D = \begin{cases} d_1 & \text{if} \quad 0 \le U \le F(d_1) \\ d_2 & \text{if} \quad F(d_1) < U \le F(d_2) \\ \vdots & \vdots \\ d_i & \text{if} \quad F(d_{i-1}) < U \le F(d_i) \\ \vdots & \vdots \\ d_{k-1} & \text{if} \quad F(d_{k-2}) < U \le F(d_{k-1}) \\ d_k & \text{if} \quad F(d_{k-1}) < U \le 1 = F(d_k) \end{cases}$$

Prove that the above method results in D having cdf F.

For U having a Uniform on (0,1) distribution, the cdf for U is given by

$$G(u) = P[U \le u] = \begin{cases} 0 & if \quad u < 0 \\ u & if \quad 0 \le u \le 1 \\ 1 & if \quad u \ge 1 \end{cases}$$

Let f be the pmf of D. Prove that $f(d_j) = F(d_j) - F(d_{j-1})$

$$\begin{aligned} f(d_j) = P[D = d_j] &= P[F(d_{j-1}) < U \le F(d_j)] \\ &= P[U \le F(d_j)] - P[U \le F(d_{j-1})] \\ &= G(F(d_j)) - G(F(d_{j-1})) \\ &= F(d_j) - F(d_{j-1}) \end{aligned}$$

Example: Generate observations from a geometric distribution with $p = .2$:

$$f(i) = p(1-p)^{i-1} = (.2)(.8)^{i-1} \quad \text{for} \quad i = 1, 2, 3, \ldots$$

$$F(i) = \sum_{k=1}^{i} f(i)$$

Suppose we observe $U = .63$ then the corresponding realization from a Geometric distribution with $p = .2$ would be that integer C such that $F(C-1) < .63 \leq F(C)$. Determine C:

$$F(1) = .2, \quad F(2) = .2 + (.2)(.8) = .36, \quad F(3) = .36 + (.2)(.8)^2 = .488$$

$$F(4) = .488 + (.2)(.8)^3 = .590, \quad F(5) = .590 + (.2)(.8)^4 = .672$$

Therefore, we have that with $U = .63$, $C = inf\{y : F(y) \geq .63\}$ which implies

$$F(4) = .590 < .63 < .672 = F(5) \quad \Rightarrow C = 5$$

The following several pages provide SAS and r code for generating random samples from specified distributions and drawing graphs of pdfs and cdfs.

The following table from *Modern Applied Statistics with S* by W.N. Venables and B.D. Ripley describes the r function names and parameters for a number of standard probability distributions. The first letter of the function name indicates which of the probability functions it describes. For example:

- **dnorm** is the density function(pdf) of the normal distribution
- **pnorm** is the cumulative disribution function(cdf) of the normal distribution
- **qnorm** is the quantile function of the normal distribution

Example Let Y have a $N(\mu = 2,\ \sigma = 5)$ distribution.

The value of the pdf at Y=4 is $f(4) = dnorm(4, 2, 5) = .0735$

The value of the cdf at Y=4 is $F(4) = pnorm(4, 2, 5) = .65542$

The value of the quantile at u=.95 is $Q(.95) = qnorm(.95, 2, 5) = 10.224$

Table 5.1: S function names and parameters for standard probability distributions.

Distribution	S name	Parameters
beta	`beta`	`shape1, shape2`
binomial	`binom`	`size, prob`
Cauchy	`cauchy`	`location, scale`
chi-squared	`chisq`	`df`
exponential	`exp`	`rate`
F	`f`	`df1, df2`
gamma	`gamma`	`shape, rate`
geometric	`geom`	`prob`
hypergeometric	`hyper`	`m, n, k`
log-normal	`lnorm`	`meanlog, sdlog`
logistic	`logis`	`location, scale`
negative binomial	`nbinom`	`size, prob`
normal	`norm`	`mean, sd`
Poisson	`pois`	`lambda`
T	`t`	`df`
uniform	`unif`	`min, max`
Weibull	`weibull`	`shape, scale`
Wilcoxon	`wilcox`	`m, n`

These functions can be used to replace statistical tables. For example, the 5% critical value for a (two-sided) t test on 11 degrees of freedom is given by `qt(0.975, 11)`, and the P value associated with a Poisson(25)-distributed count of 32 is given by (by convention) `1 - ppois(31, 25)`. The functions can be given vector arguments to calculate several P values or quantiles.

SAS Program to Generate Random Values

The following SAS program will generate 10 observations from a N(0,1) distribution and 10 observations from a Uniform on (0, 1) distribution:

```
DATA;
DO I=1 TO 10;
X = RANNOR(0);
U = RANUNI(0);
OUTPUT;
END;
RUN;
PROC PRINT X U;
RUN;
```

The following are the functions for a number of other distributions:

1. RANPOI(0,L) - POISSON DISTRIBUTION WITH PARAMETER λ = L
2. RANTRI(O,H) - TRIANGULAR DISTRIBUTION WITH PARAMETER H
3. RANBIN(0,N,P) - BINOMIAL DISTRIBUTION WITH PARAMETERS N and P
4. RANCAU(0) - CAUCHY DISTRIBUTION WITH LOC = 0, SCALE = 1
5. RANEXP(0) - EXPONENTIAL DISTRIBUTION WITH SCALE = 1
6. RANGAMMA(0,A) - GAMMA DISTRIBUTION WITH SHAPE = A, SCALE = 1

The following R commands will generate 10 observations from a N(0,1) and one observation from a Uniform on (0,1) distribution.

```
y = rnorm(10,0,1)
u = runif(1,0,1)
```

Survival Analysis and Reliability Theory

In survival analysis and reliability theory, we are interested in the time to the occurrence of an event: death, failure of a machine, cancer-free examination. Let T be the time at which the event occurs, with cdf F and pdf f, then three functions related to the r.v. T are

1. **Survival Function** is the probability that the event occurs after time t:

$$S(t) = P[T > t] = 1 - F(t)$$

(probability device works greater than t units of time).

2. **Hazard Function** (Failure Rate or Intensity Function) is the risk of failure of a device at time t given device is working at time t:

$$h(t) = \frac{f(t)}{S(t)} \Rightarrow \quad (\Delta t)h(t) \approx P[T \leq t + \Delta t | T > t] \quad \text{for very small } \Delta t$$

$$\begin{aligned} h(t) = \frac{f(t)}{S(t)} = \frac{1}{S(t)} F'(t) &= \frac{1}{P[T > t]} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{P[T > t]} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P[t < T \leq t + \Delta t]}{P[T > t]} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < T \leq t + \Delta t | T > t] \end{aligned}$$

$h(t)$ is generally reported as the number of failures per unit of time. It specifies the instantaneous rate of failure at time t given that the device is working at time t.

3. **Cumulative Hazard Function**:

$$H(t) = \int_0^t h(\tau) d\tau$$

Accumulated instantaneous risk at time t.

For a r.v. having a continuous strictly increasing cdf, $F(t)$, the following table displays the interrelationships between the various functions: pdf - $f(t)$; Survival function - $S(t)$; Hazard function - $h(t)$; Cumulative hazard function - $H(t)$

	Lifetime Distribution Relationships			
	$f(t)$	$S(t)$	$h(t)$	$H(t)$
$f(t)$	*	$S(t) = \int_t^\infty f(\tau)d\tau$	$h(t) = \frac{f(t)}{\int_t^\infty f(\tau)d\tau}$	$H(t) = -ln\left(\int_t^\infty f(\tau)d\tau\right)$
$S(t)$	$f(t) = -S^{'}(t)$	*	$h(t) = \frac{-S^{'}(t)}{S(t)}$	$H(t) = -ln(S(t))$
$h(t)$	$f(t) = h(t)e^{-\int_0^t h(\tau)d\tau}$	$S(t) = e^{-\int_0^t h(\tau)d\tau}$	*	$H(t) = \int_0^t h(\tau)d\tau$
$H(t)$	$f(t) = H^{'}(t)e^{-H(t)}$	$S(t) = e^{-H(t)}$	$h(t) = H^{'}(t)$	*

Note: For a Discrete distributions we have the following relationships:

1. pmf: $f(t) = P[T = t]$ with $\sum_t f(t) = 1$
2. cdf: $F(t) = P[T \leq t] = \sum_{\tau \leq t} f(\tau)$
3. Survival: $S(t) = P[T > t] = \sum_{\tau > t} f(\tau) = 1 - \sum_{\tau \leq t} f(\tau) = 1 - F(t)$

Exponential Dist.

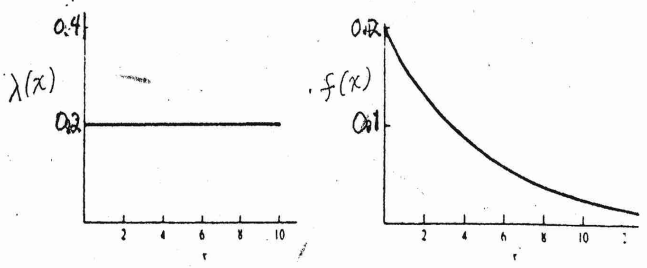
$\lambda(x) = \frac{1}{\beta}$ for $x > 0$

$f(x) = \frac{1}{\beta} e^{-\frac{1}{\beta}x}$ for $x > 0$

EX $\beta = 5$

$\lambda(x) = .2$

Failure Rate pdf



← failure rate flat

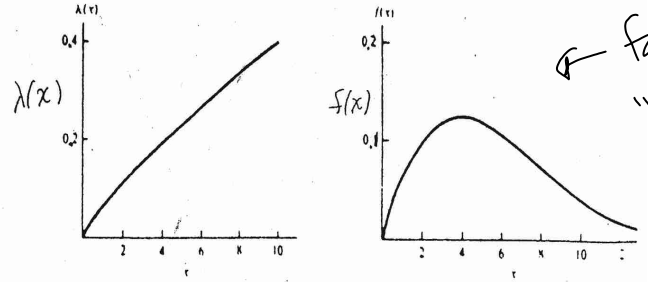
Weibull Dist.

$\lambda(x) = \gamma x^{\gamma-1}/\beta$ for $x > 0$

$f(x) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-\frac{1}{\beta}x^{\gamma}}$ for $x > 0$

EX $\gamma = 1.8$, $\beta = 28.26$

$\lambda(x) = 1.8 x^{1.8-1}/28.26 = .0637 x^{.8}$



← failure rate increasing at decreasing rate

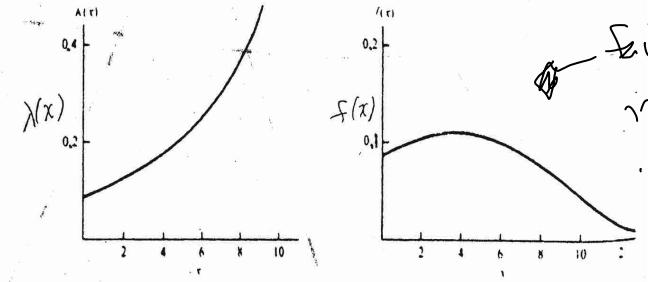
Gompertz Dist.

$\lambda(x) = c e^{bx}$ for $x > 0$

$f(x) = c \exp\{bx - \frac{c}{b} e^{bx} + \frac{c}{b}\}$ for $x > 0$

EX $b = \ln(1.2)$ $c = .087$

$\lambda(x) = .087 e^{.1823x}$ for $x > 0$



← failure rate increasing at increasing rate

FIGURE 3.3-2 Plot of the failure rate $\lambda(x)$ and the p.d.f. $f(x)$ of the following three distributions: (a) exponential with $\beta = 5$; (b) Weibull with $\beta = 6.4$ and $\gamma = 1.8$; (c) Gompertz with $b = \ln(1.2)$ and $c = 0.087$.

$Y \sim Exp(\beta)$, $X = Y^{1/\gamma}$

$F(x) = P(X \le x) = P(Y^{1/\gamma} \le x)$

$= P(y \le x^{\gamma}) = 1 - e^{-x^{\gamma}/\beta}$

$f(x) = F'(x) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^{\gamma}/\beta}$

Hogg - Ledolter

Applied Statistics for Engrg's & Physical Scientists

Leemis (1995) "Reliability, Probability Models an STATISTICAL Methods"

TABLE 4.2 TWO-PARAMETER UNIVARIATE LIFETIME DISTRIBUTIONS

Distribution	$f(t)$	$S(t)$	$h(t)$	$H(t)$	Parameters
Weibull	$\kappa\lambda^{\kappa}t^{\kappa-1}e^{-(\lambda t)^{\kappa}}$	$e^{-(\lambda t)^{\kappa}}$	$\kappa\lambda^{\kappa}t^{\kappa-1}$	$(\lambda t)^{\kappa}$	$\lambda>0;\kappa>0$
Gamma	$\frac{\lambda(\lambda t)^{\kappa-1}e^{-\lambda t}}{\Gamma(\kappa)}$	$1-I(\kappa,\lambda t)$	$\frac{\lambda(\lambda t)^{\kappa-1}e^{-\lambda t}}{\Gamma(\kappa)[1-I(\kappa,\lambda t)]}$	$-\log(1-I(\kappa,\lambda t))$	$\lambda>0;\kappa>0$
Uniform	$\frac{1}{b-a}$	$\frac{b-t}{b-a}$	$\frac{1}{b-t}$	$-\log\left(\frac{b-t}{b-a}\right)$	$0\le a<b$
Log normal	$\frac{1}{\sigma t\sqrt{2\pi}}e^{\frac{-(\log t-\mu)^2}{2\sigma^2}}$	$\int_t^{\infty} f(\tau)d\tau$	$\frac{f(t)}{S(t)}$	$-\log S(t)$	$-\infty<\mu<\infty;\sigma>0$
Log logistic	$\frac{\lambda\kappa(\lambda t)^{\kappa-1}}{[1+(\lambda t)^{\kappa}]^2}$	$\frac{1}{1+(\lambda t)^{\kappa}}$	$\frac{\lambda\kappa(\lambda t)^{\kappa-1}}{1+(\lambda t)^{\kappa}}$	$\log[1+(\lambda t)^{\kappa}]$	$\lambda>0;\kappa>0$
Inverse Gaussian	$\sqrt{\frac{\lambda}{2\pi t^3}}e^{-\frac{\lambda}{2\mu^2 t}(t-\mu)^2}$	$\int_t^{\infty} f(\tau)d\tau$	$\frac{f(t)}{S(t)}$	$-\log S(t)$	$\lambda>0;\mu>0$
Exponential Power	$(e^{1-e^{\lambda t^{\kappa}}})e^{\lambda t^{\kappa}}\lambda\kappa t^{\kappa-1}$	$e^{(1-e^{\lambda t^{\kappa}})}$	$e^{\lambda t^{\kappa}}\lambda\kappa t^{\kappa-1}$	$e^{\lambda t^{\kappa}}-1$	$\lambda>0;\kappa>0$
Pareto	$\frac{\kappa\lambda^{\kappa}}{t^{\kappa+1}}$	$\left(\frac{\lambda}{t}\right)^{\kappa}$	$\frac{\kappa}{t}$	$\kappa\log\left(\frac{t}{\lambda}\right)$	$\lambda>0;\kappa>0$
Gompertz	$\delta\kappa^{t}e^{[-\delta(\kappa^{t}-1)/\log\kappa]}$	$e^{[-\delta(\kappa^{t}-1)/\log\kappa]}$	$\delta\kappa^{t}$	$\frac{\delta(\kappa^{t}-1)}{\log\kappa}$	$\kappa>1;\delta>0$

START CLASS NOTES: Mon 9/20/21

HANDOUT #5: PARAMETRIC SUMMARIES OF POPULATIONS AND PROCESSES

1. Summaries of Center/Location in a Distribution
   (a) Population/Process Mean (μ)
   (b) Population/Process Median ($\tilde{\mu}$)
   (c) Population/Process Quartiles ($Q(.25), \tilde{\mu}, Q(.75)$)
   (d) Population/Process Trimmed Mean ($\mu_{(\alpha)}$)
2. Summaries of Level of Dispersion/Variability in a Distribution
   (a) Population/Process Range (R)
   (b) Population/Process Semi-interquartile Range (SIQR)
   (c) Population/Process Standard Deviation (σ)
   (d) Population/Process Median Absolute Deviation (MAD)
3. Summaries of Shape/Tail Weight in a Distribution
   (a) Population/Process Skewness (β_1)
   (b) Population/Process Kurtosis (β_2)
4. Do Mean and Standard Deviation Summarize Distribution?
5. How Are Mean and Median Related?
6. Correlation and AutoCorrelation ($\rho_{y,x}, \rho_k$)

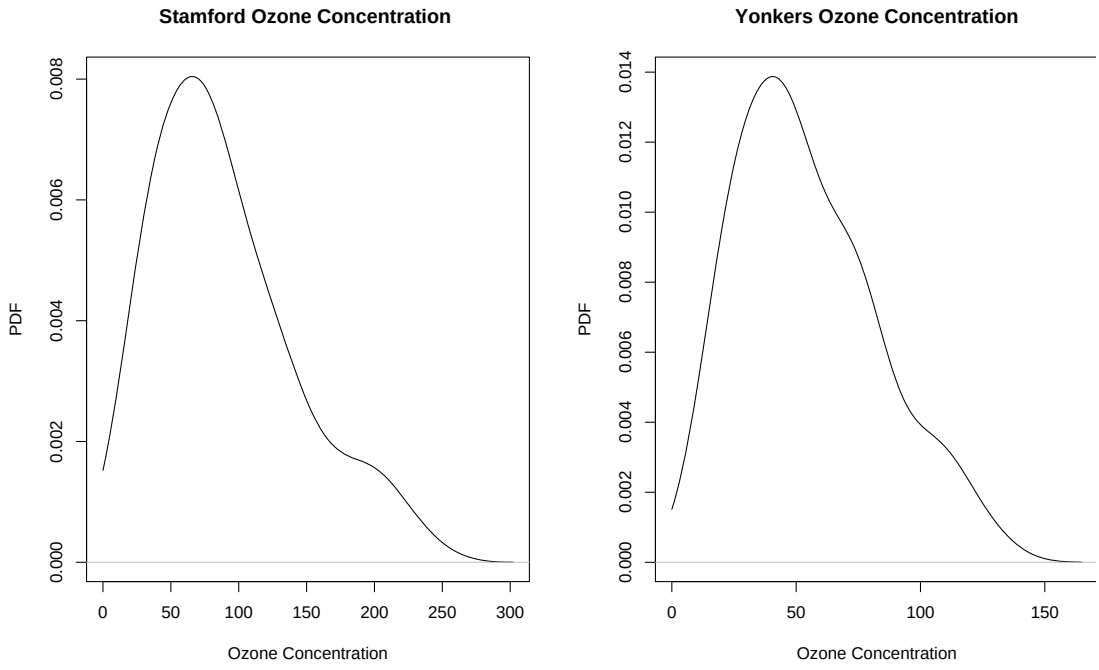
Supplemental Reading:

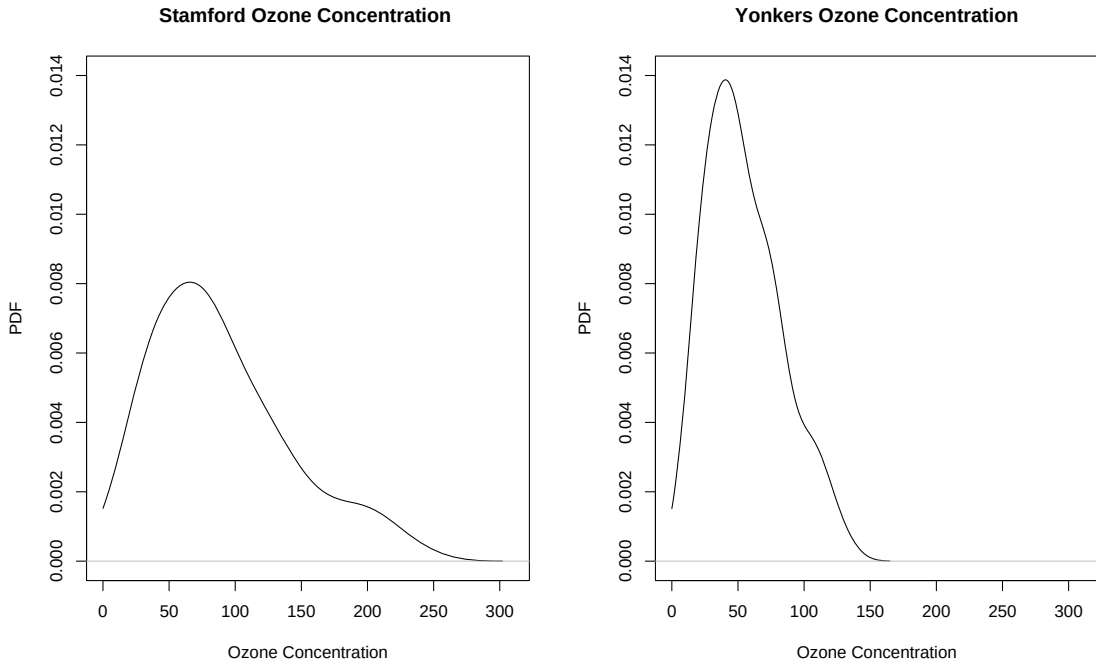
- Chapter 2 and Sections 4.1, 4.2, and 4.3.4 in the Tamhane/Dunlop book

Comparing Several Population Distributions

A process or population is completely described by its cdf or quantile function or pdf. However, it is often very difficult to compare several populations or processes using their pdfs or cdfs. Also, its difficult to precisely quantify the differences in several pdfs or cdfs. The distribution of maximum ozone concentration in two cities, Stamford, Connecticut and Yonkers, New York over the summer months are given on the next two pages. In the first set of graphs, we plotted the pdfs with the default settings for the scale on the Y-axis and X-axis. From these two plots the distribution of maximum ozone would appear to be nearly the same in the two cities. When we set the range of the scales to be the same for the two pdfs, a very large difference in the two distributions is apparent. However, if we were EPA regulators or state environmental monitors, how can we describe the distribution of maximum ozone concentration or state precisely the difference in the distributions for the two cities? Furthermore, if we wanted to quantify changes in ozone concentrations in the two cities after new environmental controls on the emissions from automobiles or industrial plants were implemented, would it be possible to make precise statements about any such changes that the public or congress would understand using just the plots of the distributions?

Thus, we will define several parameters which will summarize the information contained in the cdf or pdf concerning the distribution of values in a population or process. These numerical summaries will not completely describe the differences in several populations or processes but may provide an adequate description of changes or differences in many situations.





Parametric Summarization of a Population Distribution

Definition A *parameter* is a numerical characteristic of a population or process.
It may be a functional of the cdf or a constant contained within the formula for the cdf.

Example 1 The location, scale, or shape parameters in a cdf:

The three parameter Weibull distribution has a location parameter θ, a scale parameter β and shape parameter γ:

$$F(y;\theta,\beta,\gamma) = \begin{cases} 0 & \text{if } y < \theta \\ 1 - e^{-((y-\theta)/\beta)^{\gamma}} & \text{if } y \geq \theta \end{cases}$$

$$\text{with } -\infty < \theta < \infty, \quad \beta > 0, \quad \gamma > 0.$$

The three quantities θ, β, and γ are parameters.

Example 2 The median or IQR of the expenses per household in Louisiana resulting from the lack of a rapid local/state/federal response to problems caused by Katrina.
Both the median and IQR would be parameters associated with the distribution of household damage expenses.

Example 3 The mean and standard deviation of the mileage (MPG) of a new hybrid automobile.
The mean and standard deviation of MPG would be used to compare MPG of hybrids to conventional automobiles.

We will now define a variety of parameters which will attempt to describe different aspects of a distribution. Many of these parameters will be associated with a functional of the cdf called the *expected value* of a function of the r.v Y, **E[h(Y)]**:

Definition: Expected Value of h(Y): The expected value of a function, h(Y) of a r.v. Y having cdf F_Y and pdf f_Y is defined as

For a continuous strictly increasing cdf:

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y)f(y)dy$$

For discrete cdf:

$$\mu = E[h(Y)] = \sum_{i=1}^{\infty} h(y_i)f(y_i)$$

From the above formulas, the expected value of a random variable is a weighted average of the values of the random variables with weights being the pdf of the random variable.

EXAMPLE Let Y be the number of defective parts in a randomly selected container of 10 parts. The pmf for Y is given in the following table. The cost per container of repairing defective parts is given by $C = h(Y) = 5Y + 3$. The average cost of repairs is then obtained by weighting the values of $h(Y)$ by the corresponding values of the pmf and then summing these values.

y	0	1	2	3	4	5	6	7	8	9	10	
c	3	8	13	18	23	28	33	38	43	48	53	Total
$f(y)$	.68	.18	.05	.03	.015	.014	.012	.008	.007	.003	.001	1.000
$y * f(y)$	0	0.180	0.100	0.090	0.060	0.070	0.072	0.056	0.056	0.027	0.010	0.721
$c * f(y)$	2.040	1.440	0.650	0.540	0.345	0.392	0.396	0.304	0.301	0.144	0.053	6.605

From the above table we have that the average number of defects per container is $E(Y) = .721$ and the average cost of repairing the parts in a container is $E(C) = 6.605$.

As is often true for random variables taking on only integer values, $E(Y)$ is not one of the possible values of Y and $E(C)$ is not one of the possible values for C.

Note if we just took the average number of defects per container, $E(Y)$, .721 and used it in the cost formula we would obtain $E(C) = 5(.721)+3 = 6.605$.

Optional Material: Riemann vs Riemann-Stieltjes Integral

Many parameters are defined by specifying the function h(y):

Mean value of Y, $\mu = E[Y]$, has $h(y) = y$,

Variance of Y, $\sigma^2 = E[(Y-\mu)^2]$, has $h(y) = (y-\mu)^2$.

It is somewhat cumbersome having two separate definitions for $E[h(Y)]$ depending on whether the distribution of Y is discrete or continuous. The two definitions can be consolidated through the use of a generalization of the Riemann Integral, the Riemann-Stieltjes Integral:

Let g be a bounded function on $[a, b)$.

Let $P_n = a = x_{0,n} < x_{1,n} < \cdots < x_{n-1,n} < x_{n,n} = b$ be a partition of $[a, b)$ with mesh size,

$$||P_n|| = \max_{1 \le i \le n} |x_{i,n} - x_{i-1,n}|.$$

$||P_n||$ is the maximum gap in P_n.

Riemann Integral: Let P_n be a sequence of partitions of $[a, b)$ for which each term in the sequence is a refinement of its predecessor and for which $\lim_{n\to\infty} ||P_n|| = 0$. Let $x^*_{i,n} \epsilon [x_{i-1,n}, x_{i,n})$ and $R(P_n) = \sum_{i=1}^n g(x^*_{i,n})(x_{i,n} - x_{i-1,n})$ be an approximating sum for the area under the curve $g(\cdot)$ between $[a, b)$. If $\lim_{n\to\infty} R(P_n) = R$ for all such sequences of partitions then $R = \int_a^b g(x)dx$ is the Riemann Integral of g over $[a, b)$.

Generalization of Riemann Integral is Riemann-Stieltjes Integral of g with respect to F:

Riemann-Stieltjes Integral of g wrt F: Let F be a cdf and g be a continuous, real valued function. Let P_n be a sequence of partitions of $[a, b)$ for which each term in the sequence is a refinement of its predecessor and for which $\lim_{n\to\infty} ||P_n|| = 0$. Let $x^*_{i,n} \epsilon [x_{i-1,n}, x_{i,n})$ and

$$RS(P_n) = \sum_{i=1}^{n} g(x^*_{i,n}) \left[F(x_{i,n}) - F(x_{i-1,n})\right]$$

be an approximating sum. If $\lim_{n\to\infty} RS(P_n) = RS$ for all such sequences of partitions then $RS = \int_a^b g(x)dF(x)$ is the Riemann-Stieltjes Integral of g w.r.t. F over $[a, b)$.

We will not need the full power of the Riemann-Stieltjes integral but a few examples will illustrate its usefulness in defining population parameters and later in defining sample estimators of these parameters.

Example 1 Let F be a continuous strictly increasing cdf with pdf f. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Example 2 Let F be a cdf of a discrete r.v. Y with pmf f and jumps in F at $y_1, y_2, \ldots$ of size $p_i = f(y_i)$. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x) = \sum_{i=1}^{\infty} g(y_i)f(y_i) = \sum_{i=1}^{\infty} p_i g(y_i)$$

Example 3 Let F be a cdf with jumps in F at $Y_1, Y_2, \ldots, Y_n$ of size $1/n$. That is, F is the empirical cdf. Then

$$\int_{-\infty}^{\infty} g(y)dF(y) = \sum_{i=1}^{n} \frac{1}{n} g(Y_i) = \frac{1}{n} \sum_{i=1}^{n} g(Y_i)$$

In particular, if $g(y) = y$ then

$$\int_{-\infty}^{\infty} g(y)dF(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

Note that F is the empirical distribution function (edf) for the n data values $Y_1, Y_2, \ldots, Y_n$

Parametric Summaries of the Center and Dispersion of a Distribution

We will define several parameters associated with a distribution which describe the "average" value in the distribution or the amount of "dispersion" in the distribution. That is, how spread out the values are about the "center" of the distribution. Let Y be a r.v. which represents a random selected value from a population or process. Suppose Y has cdf F and pdf(pmf) f.

Definition: The kth Moment of Y The kth moment of Y is

$$\begin{aligned} m_k = E[Y^k] &= \int_{-\infty}^{\infty} y^k dF(y) \\ &= \int_{-\infty}^{\infty} y^k f(y) dy \quad \text{if F is continuous} \\ &= \sum_{i=1}^{\infty} y_i^k f(y_i) \quad \text{if F is discrete} \end{aligned}$$

We usually denote m_1 as μ.

Definition: The kth Central Moment of Y about the center μ

The kth central moment of Y is

$$\begin{aligned} \mu_k = E[(Y-\mu)^k] &= \int_{-\infty}^{\infty} (y-\mu)^k dF(y) \\ &= \int_{-\infty}^{\infty} (y-\mu)^k f(y) dy \quad \text{if F is continuous} \\ &= \sum_{i=1}^{\infty} (y_i-\mu)^k f(y_i) \quad \text{if F is discrete} \end{aligned}$$

where $\mu = m_1$ is called the mean or expected value of the r.v. Y.

Note: $\mu_1 = 0$

Mean of a Population or Distribution

Several key parameters used in describing distributions are either central moments or are functions of several central moments. We will now define two of these parameters, the mean and standard deviation. These are the two most widely used parameters (correctly used or not) in describing the center and dispersion in distributions.

Definition: The expected value or mean value of a r.v. Y or of its distribution F is defined as

$$\mu = E[Y] = m_1$$

For a r.v. Y having pdf $f(y)$,

$$\mu = E[Y] = \int_{-\infty}^{\infty} yf(y)dy$$

Example Suppose Y has pdf $f(y) = \frac{1}{\beta}e^{-y/\beta}$ for $y \geq 0$ and 0 otherwise.

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy = \int_{-\infty}^{0} y(0)dy + \int_{0}^{\infty} y\frac{1}{\beta}e^{-y/\beta}dy = 0 - \beta e^{-y/\beta}|_0^{\infty} = \beta$$

This is a "weighted" average value of the occurrence of the values of Y in the population or process where the weights are the values of the pdf. The idea of weighting is more clearly observed in the case of a discrete distributions.

For example, suppose Y has k possible values $y_1, y_2, \cdots, y_k$ and pmf f, where $f(y_i) = Pr[Y = y_i]$. Then, μ is simply the weighted average value of the $y_i s$, with weights $f(y_i)$:

$$\mu = E[Y] = \sum_{i=1}^{k} y_i f(y_i).$$

In particular, for a population containing N units with distinct values of Y: $y_1, y_2, \cdots, y_N$ occuring with frequencies: $f_i = 1/N$, we have

$$\mu = \sum_{i=1}^{N} y_i f_i = \sum_{i=1}^{N} y_i 1/N = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}$$

Thus μ is our common notation for the mean of a population.

For some continuous, heavy-tailed distributions μ does not exist because the integral is indeterminate. For the Cauchy distribution,

$$\mu = E[Y] = \int_{-\infty}^{\infty} yf(y)dy = \int_{-\infty}^{\infty} y\left(\pi\theta_2\left[1 + \left(\frac{y-\theta_1}{\theta_2}\right)^2\right]\right)^{-1} = \infty - \infty \neq 0$$

Standard Deviation of a Population or Distribution

The most widely used measure of how spread out the values of Y are about the measure of the center μ is the standard deviation of Y:

Definition: The Variance of a r.v. Y or of its distribution F is defined as

$$\sigma^2 = Var(Y) = E\left[(Y-\mu)^2\right] = \int_{-\infty}^{\infty}(y-\mu)^2 dF(y)$$

σ^2 is the weighted average squared distance of the values of Y about the mean μ. Its units are the square of the units of Y.

When Y has a pdf $f(y)$, $\sigma^2 = \int_{-\infty}^{\infty}(y-\mu)^2 f(y)dy$

A little algebra demonstrates that $\sigma^2 = E\left[(Y-\mu)^2\right] = E[Y^2] - (\mu)^2 = m_2 - m_1^2 = \mu_2$

Example Suppose Y has pdf $f(y) = \frac{1}{\beta}e^{-y/\beta}$ for $y \geq 0$ and 0 otherwise.

$\mu = E[Y] = \beta$ and

$$E[Y^2] = \int_0^{\infty} y^2 \frac{1}{\beta}e^{-y/\beta}dy = 2\beta^2 \;\Rightarrow\; \sigma^2 = E[Y^2] - \mu^2 = 2\beta^2 - (\beta)^2 = \beta^2$$

For any continuous distributions in which the mean does not exist, then the variance would also not exist.

For example, the Cauchy distribution.

For the t-distribution with df $= 2$, $E[Y] = 0$ but $E[Y^2] = \infty$.

For the t-distribution with df $= 1$, the mean and variance do not exist (t with df=1 is a Cauchy distribution).

In particular, for a population containing N units with distinct values of Y: $y_1, y_2, \cdots, y_N$ occurring with frequencies: $f_i = 1/N$, we have

$$\sigma^2 \;=\; \sum_{i=1}^{N}(y_i-\mu)^2 f_i \;=\; \sum_{i=1}^{N}(y_i-\mu)^2 1/N \;=\; \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$$

Note that when we are describing the population variance, the divisor is N and not $N-1$, as would be in the sample standard deviation when μ is unknown.

Definition: The Standard Deviation of a r.v. Y or of its distribution F is defined as

$$\sigma = \sqrt{Var(Y)} = \sqrt{E\left[(Y-\mu)^2\right]}$$

The standard deviation has the same units as Y and hence is more easily interpreted as a measure of the dispersion of the population values about the mean than is the variance which has units the square of the units of the r.v.

Example 1 For the example on page 4 concerning Y, the number of defective parts in a randomly selected container of 10 parts. We have the pmf for Y is given in the following table:

y	0	1	2	3	4	5	6	7	8	9	10	
$f(y)$	.68	.18	.05	.03	.015	.014	.012	.008	.007	.003	.001	1.000

The expected value of Y is given by

$$\mu = E[Y] = \sum_{y=0}^{10} yf(y) \Rightarrow$$

$$\mu = 0(.68) + 1(.18) + 2(.05) + 3(.03) + 4(.015) + 5(.014) + 6(012) + 7(.008) + 8(.007) + 9(.003) + 10(.001) = 0.721$$

That is, the average number of defectives per container is 0.721.

A simplified formula for computing variance is given by

$$\sigma^2 = E[(Y-\mu)^2] = E[Y^2 - 2Y\mu + \mu^2] = E[Y^2] - 2\mu E[Y] + \mu^2] = E[Y^2] - \mu^2$$

For our defectives example, we have

$$E[Y^2] = 0(.68)+1(.18)+4(.05)+9(.03)+16(.015)+25(.014)+36(012)+49(.008)+64(.007)+81(.003)+100(.001) = 2.855$$

Therefore, the standard deviation has value,

$$\sigma = \sqrt{Var(Y)} = \sqrt{E[Y^2] - \mu^2} = \sqrt{2.855 - (.721)^2} = 1.528$$

Example 2 Let X have an exponential pdf: $f(x) = \frac{1}{\beta}e^{-x/\beta}$ for $x > 0$

$$\begin{aligned}
\mu &= E[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_0^{\infty} x\frac{1}{\beta}e^{-x/\beta}\mathrm{d}x = -xe^{-x/\beta}|_0^{\infty} - \beta e^{-x/\beta}|_0^{\infty} = \beta \\
E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x = \int_0^{\infty} x^2\frac{1}{\beta}e^{-x/\beta}\mathrm{d}x = 2\beta^2 \\
\sigma^2 &= Var(X) = E[X^2] - (\mu)^2 = 2\beta^2 - (\beta)^2 = \beta^2 \Rightarrow \\
\sigma &= \beta
\end{aligned}$$

Note: For the exponential distribution, $\sigma = \beta = \mu$

How completely do the pair (μ, σ) describe a family of distributions?

One answer to the above question is to examine the proportion of the population values falling within $k\sigma$ units of μ, $p(k)$. That is, determine whether or not $p(k)$ varies depending on

1. the values of (μ, σ) or
2. the family of distributions or
3. is it a constant for all distributions?

We will answer the above questions by examining:

$$p(k) = P[Y \text{ is within } k\sigma \text{ units of } \mu] \Rightarrow$$

$$p(k) = P[|Y - \mu| \leq k\sigma] = P[\mu - k\sigma \leq Y \leq \mu + k\sigma] = F[\mu + k\sigma] - F[\mu - k\sigma]$$

General Bound: Chebyshev's Inequality If Y is a r.v. with mean μ, $0 < \sigma < \infty$ and $k > 1$ then

$$p(k) = P[|Y - \mu| \leq k\sigma] \geq 1 - \frac{1}{k^2}$$

How sharp is this bound? That is, how close is the true probability of this event to the value given by the bound, $1 - \frac{1}{k^2}$?

Case 1: Location-Scale Family Suppose the mean and standard deviation (μ, σ) of the r.v. Y are also location-scale parameters in a family of distributions. Then, with $Z = (Y - \mu)/\sigma$,

$$P[|Y - \mu| \leq k\sigma] = P[|Y - \mu|/\sigma \leq k] = P[|Z| \leq k] = F_Z[k] - F_Z[-k]$$

The proportion of the distribution within k standard deviations of the mean is the same value for every member of the family. It does not depend on the values of (μ, σ). However, the proportion does vary from family to family because the cdf of Z, F_Z would be different for each family of distributions, that is, the proportion for the normal family of distributions is different from the value for the Cauchy or Double Exponential family of distributions.

Case 2: General Family of Distributions In general, the exact proportion will vary greatly within even the same family and will depend on the values of (μ, σ) when (μ, σ) are not location-scale parameters for the population.

We will illustrate these ideas with a few examples.

The following table contains the proportion of the distribution within k standard deviations of the population mean for various values of k and distributions:

$$p(k) = P[|Y - \mu| \leq k\sigma] = F_Y(\mu + k\sigma) - F_Y(\mu - k\sigma)$$

By Chebyshev inequality, $p(k) \geq 1 - \frac{1}{k^2}$

	k						
Distribution	.5	1	1.25	1.5	2	2.5	3
Bound$(1 - \frac{1}{k^2})$	-3.0	0	.36	.556	.750	.840	.889
Normal	.383	.683	.789	.866	.955	.988	.9973
Cauchy	.295	.500	.570	.626	.705	.758	.7950
ChiSq(df=1)	.3970	.880	.904	.923	.950	.967	.978
ChiSq(df=5)	.3820	.724	.848	.916	.955	.978	.987
ChiSq(df=10)	.3823	.701	.815	.893	.959	.980	.991
ChiSq(df=50)	.3828	.686	.794	.871	.956	.986	.9954

From the above table we note the following:

1. If $k \leq 1$, the Chebyshev bound is not useful, $1 - \frac{1}{k^2} \leq 0$.
2. The Chebyshev bound is not very close to the true proportion for the normal distribution or the Chi-square distribution. Not surprising because the bound must be true for all distributions having a finite standard deviation.
3. The Cauchy distribution has proportions which are in fact smaller than the values specified by the Chebyshev bound. Does this fact demonstrate that the Chebyshev bound is not valid?
4. The proportion of the chi-square distribution that is within 1.0 standard deviations of its mean varies widely from 0.880 to 0.686 as the df increase from 1 to 50.

df	1	5	10	50	75	100	3000
$\mu = df$	1	5	10	50	75	100	3000
$\sigma = \sqrt{2df}$	1.414	3.162	4.472	10	12.2474	14.1421	77.4597
$P[\|Y - \mu\| \leq \sigma]$	.8798	.7236	.7007	.6860	.6849	.6843	.6827

5. The proportion of the chi-square distribution that is within 1 standard deviations of its mean is close to the proportion for the normal distribution, .6827, when $df \geq 3000$.

Thus, we observe that the pair (μ, σ) do not adequately describe a distribution. We need to know much more about the distribution in order to provide a complete picture of the distribution, for example, the peakedness of the distribution or the tail behavior of the distribution. In an attempt to provide this information, we define two more parameters based on the central moments.

STOP : 9/20/21

Skewness and Kurtosis

Definition: **The skewness** of a r.v. Y or its distribution F is defined as

$$\beta_1 = \frac{E[(Y-\mu)^3]}{\sigma^3} = \frac{\mu_3}{(\mu_2)^{3/2}}.$$

We note the following properties of β_1:

1. The distribution of Y is said to be symmetric about θ if

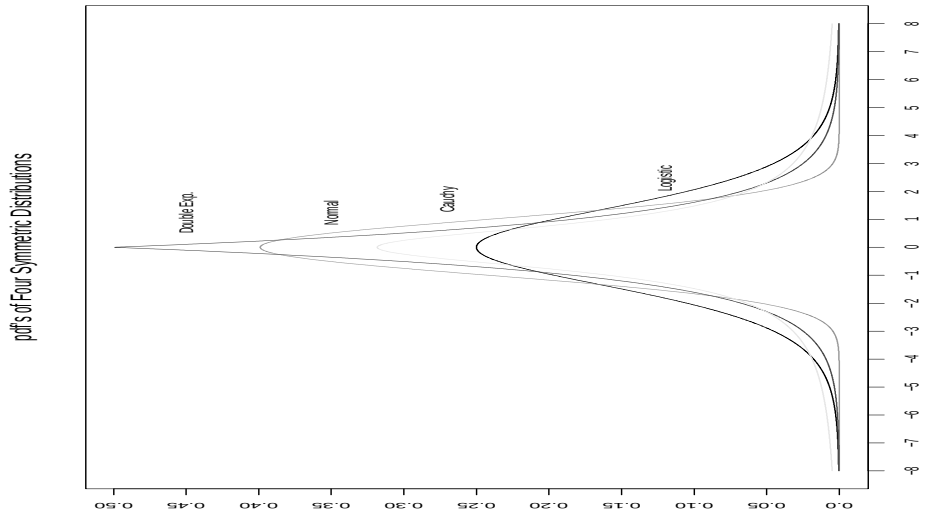
$$P[Y \leq \theta - y] = P[Y \geq \theta + y] \text{ for all } y,$$

that is,

$$F(\theta - y) = 1 - F((\theta + y)^-) \text{ for all } y,$$

If F is symmetric with pdf f then $F(\theta - y) = 1 - F(\theta + y)$ for all $y \Rightarrow$

$$f(\theta - y) = f(\theta + y) \text{ for all } y.$$



- If the distribution of Y is symmetric about θ then $\beta_1 = 0$ and $\theta = \mu$ provided μ exists.

2. The converse of the above is not true in general. That is, $\beta_1 = 0$ does not necessarily imply that the cdf is symmetric See Ord (1968) *Annals of Mathematical Statistics*, 39, pp. 1513-1516. This article shows for the discrete t-distribution that $\beta_1 = 0$ but the distribution is not symmetric.

3. Right skewed distributions have $\beta_1 \geq 0$ and left skewed distributions have $\beta_1 \leq 0$.

4. Although, $\beta_1 = 0$ does not imply symmetry in the distribution. The following is true

- If $\beta_1 \neq 0$, then the pdf f is not symmetric.

Definition: **The kurtosis** of a r.v. Y or its distribution F is defined as

$$\beta_2 = \frac{E[(Y-\mu)^4]}{\sigma^4} = \frac{\mu_4}{(\mu_2)^2} \quad \text{scaled 4th moment}$$

What does kurtosis actually measure. The following statement is taken from our textbook:

The kurtosis measures the tail-heaviness (the amount of probability in the tails) of the distribution. For the normal distribution, $\beta_2 = 3$. The normal distribution is considered a light-tailed distribution because the probability in its tails beyond, say, three standard deviations from the mean is negligible (.0027). Thus, depending on whether $\beta_2 > 3$ or $\beta_2 < 3$, a distribution is heavier tailed or lighter tailed than the normal distribution.

The above interpretation of kurtosis is often stated. However, this interpretation is very inaccurate. There are many articles addressing the interpretation of kurtosis. Two of these articles are

Kurtosis: A Critical Review, by K. Balanda and H. MacGillivray, *The American Statistician*, May 1988, Vol. 42, pp. 111-120.

The Meaning of Kurtosis: Darlington Reexamined, by J. Moors, *The American Statistician*, November 1986, Vol. 40, p. 283.

A few comments will be extracted from these articles:

- From Moors: *A valid interpretation (of kurtosis) may be formulated as follows: kurtosis measures dispersion around two values $\mu \pm \sigma$, it is an inverse measure of the concentration in these two points. High kurtosis, therefore, may arise in two situations:*

1. *concentration of probability mass near μ, corresponding to a peaked unimodal distribution and*

2. *concentration of probability mass in the tails of the distribution.*

- From Balanda and MacGillivray: Because of the "averaging" nature of moments, however, the relationship of β_2 to shape is far from clear;. . .

1. An error commonly associated with kurtosis is that the sign of $\beta_2 - 3$ compares the value of the density (function) at the center with that of the corresponding normal density.

2. The value of β_2 is affected by so many different aspects of a distribution that . . . a given value of β_2 can correspond to several different distributional shapes.

3. Figure 2 contains a number of standardized symmetric densities with $\mu = 0, \sigma = 1, \beta_2 = 3$ (the value for the normal distribution). Although Curve 3 has finite support (and thus short tails) it is a good approximation to the normal distribution. Curve 4 is bimodal whereas Curve 2, although it has infinite support and is unimodal, is considerably more peaked than the standard normal distribution. Thus, there is no logical connection between the value of the density (function) of the standardized distribution at the center and the sign of $\beta_2 - 3$.

Curve 1 - Normal Distribution
Curve 2 - TukeyLambda[L=5.2] Distribution
Curve 3 - TukeyLambda[L=.135] Distribution
Curve 4 - Double Gamma Distribution

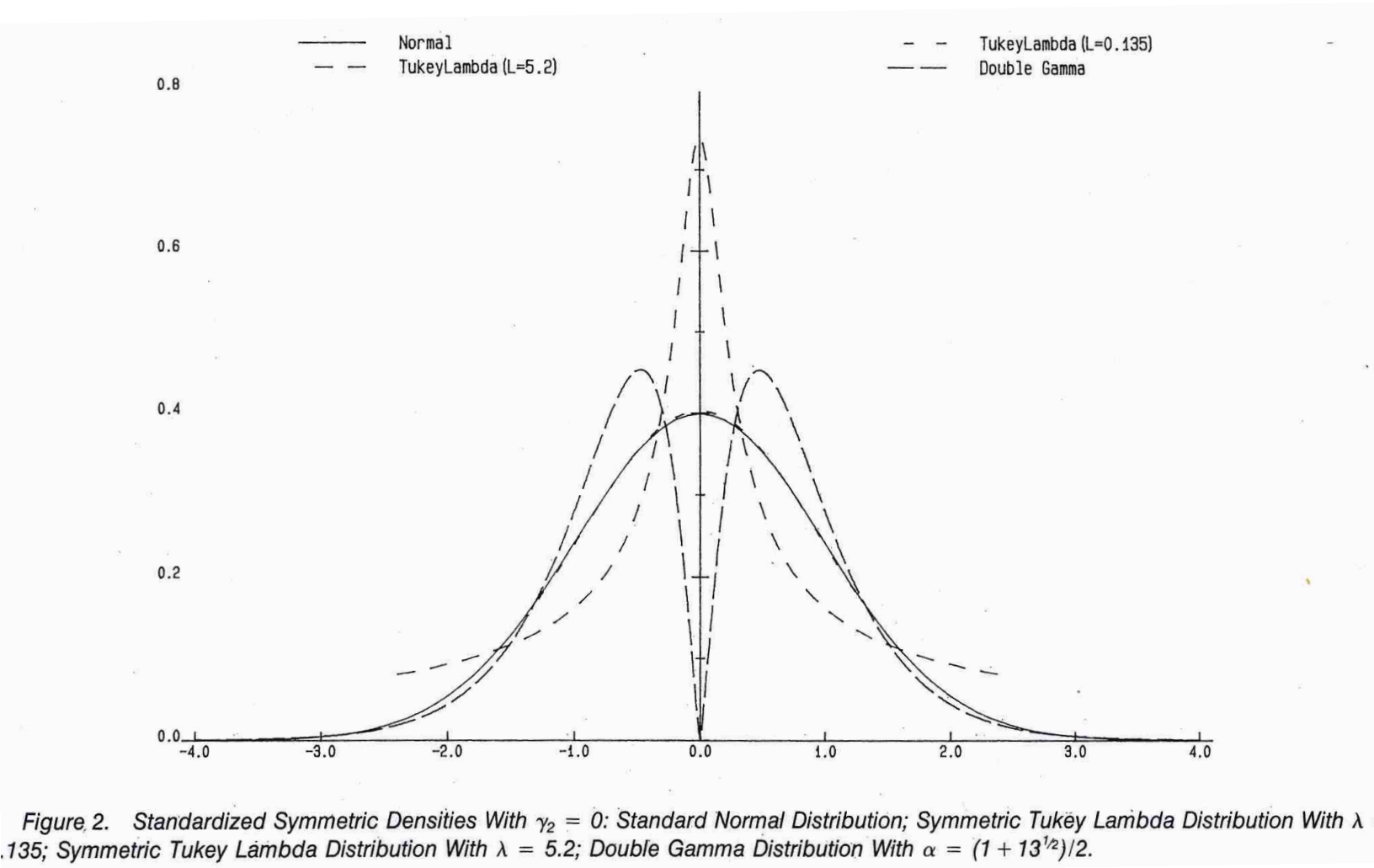


Figure 2. Standardized Symmetric Densities With $\gamma_2 = 0$: Standard Normal Distribution; Symmetric Tukey Lambda Distribution With λ .135; Symmetric Tukey Lambda Distribution With $\lambda = 5.2$; Double Gamma Distribution With $\alpha = (1 + 13^{1/2})/2$.

The following discussion was sent to me but I do not have the reference:

There are two fallacies that are commonly associated with the shape parameters, skewness and kurtosis. The first fallacy is that two distributions having the same mean, standard deviation, skewness, and kurtosis have the same shape. In fact, Karl Pearson stated in the early 1900's that knowing $\mu, \sigma, \beta_1, \beta_2$ would completely describe a distribution. The second fallacy is that a distribution with a skewness parameter of zero will be symmetric. That these are indeed fallacies will be illustrated by the following examples. Consider the following two pdfs:

$$f_1(y) = \begin{cases} 0.6391 + 1.0337y & if \quad -0.0091 < y < 0.5387 \\ 1.7527 - 1.0337y & if \quad 0.5387 < y < 1.0864 \\ 0 & if \quad y \leq -0.0091 \text{ or } y \geq 0.5387 \end{cases}$$

$$f_2(y) = \begin{cases} (18.1484)(.0629)y^{-19.1484}\left[1 + y^{-18.1484}\right]^{-1.0629} & if \quad y > 0 \\ 0 & if \quad y \leq 0 \end{cases}$$

The two pdfs have the same values for the parameters:

$$\mu = .5387 \qquad \sigma = .2907 \qquad \beta_1 = 0 \qquad \beta_2 = 2$$

The pdfs are graphed below. Thus, even though the two pdfs have the same values for mean, standard deviation, skewness and kurtosis, they definitely do not have the same shape. Also, pdf f_2 has skewness coefficient equal to 0 when it is obvious that f_2 is nonsymmetric.

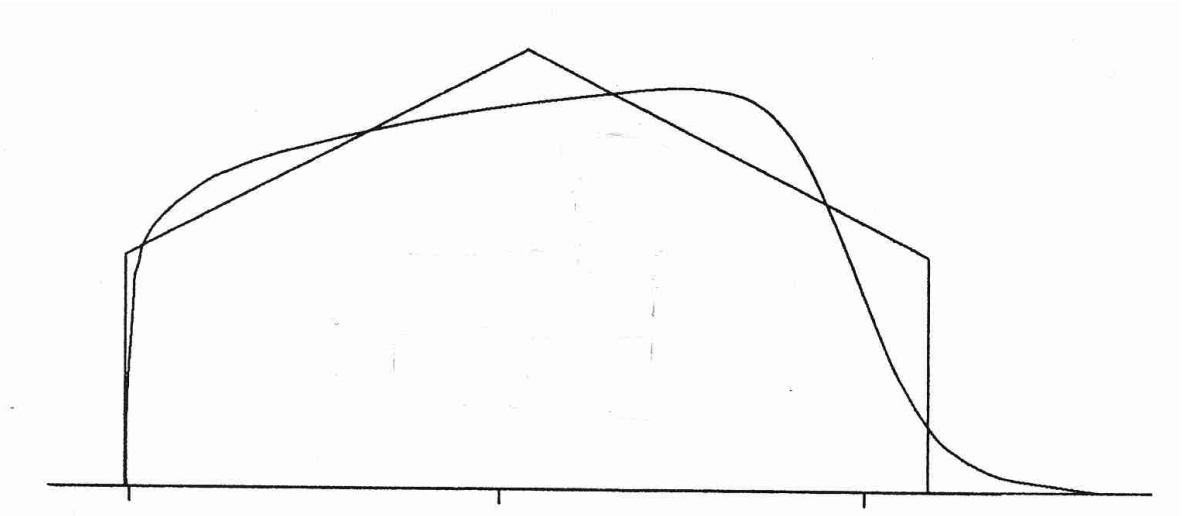


Figure 13.7: The Usefulness of Skewness and Kurtosis

Casella-Berger on page 64 display two pdf's which have all their moments, m_r for $r = 1, 2, 3, \cdots$, equal, but the two pdf's are very different. The moments of a distribution, m_r, uniquely determine the distribution, only if the Carleman's Condition holds: $\sum_{r=1}^{\infty} m_{2r}^{-1/2r} = \infty$

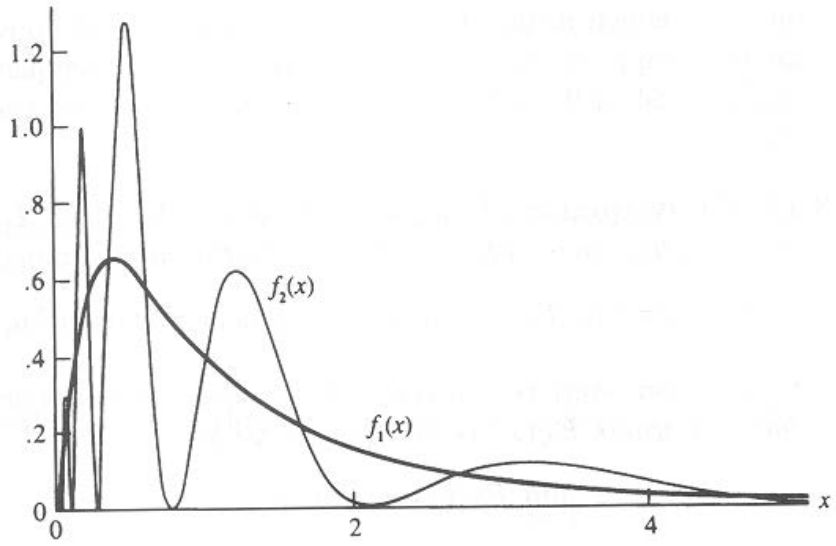


Figure 2.3.2. *Two pdfs with the same moments:* $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}$ *and* $f_2(x) = f_1(x)[1 + \sin(2\pi \log x)]$

The following table contains values for the four parameters μ, σ, β_1, and β_2 for a wide variety of distributions.

TABLE 4.1
Some frequently encountered continuous probability distribution functions

Distribution	Probability Density Function		Mean	Variance
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	$-\infty < \mu < \infty$ $\sigma > 0$	μ	σ^2
χ^2	$f(x) = \frac{x^{(n/2)-1}e^{-(x/2)}}{2^{n/2}\Gamma(n/2)}$	$x > 0$ $n > 0$	n	$2n$
t	$f(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{[(n+1)/2]}$	$n > 0$	0	$\frac{n}{n-2}$ $(n>2)$
F	$f(x) = \frac{n^{n/2}m^{m/2}}{\beta(n/2, m/2)} x^{(n-2)/2}(m + nx)^{-(n+m)/2}$	$x > 0$ $m,\ n > 0$	$\frac{m}{m-2}$ $(m > 2)$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ $(m > 4)$
Gamma	$f(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x}$	$x > 0$ $\lambda, r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Exponential	$f(x) = \lambda e^{-\lambda x}$	$\lambda > 0$ $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Weibull	$f(x) = \alpha\beta x^{\beta-1}\exp(-\alpha x^\beta)$	$x > 0$ $\alpha, \beta > 0$	$\alpha^{-1/\beta}\Gamma\left(1 + \frac{1}{\beta}\right)$	$\alpha^{-\frac{2}{\beta}}\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right\}$
Lognormal	$f(x) = \frac{1}{\beta\sqrt{2\pi}} x^{-1} \exp\left[-(\ln x - \alpha)^2/2\beta^2\right]$	$x > 0$ $\beta > 0$	$\exp\left(\alpha + \frac{\beta^2}{2}\right)$	$\exp(2\alpha + \beta^2)\left[\exp(\beta^2) - 1\right]$
Beta	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$0 < x < 1$ $\alpha, \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Uniform	$f(x) = \frac{1}{b-a}$	$a \le x \le b$ $-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Distribution	β_1	β_2
Normal	0	3
χ^2	$\frac{8}{n}$	$3\left(\frac{4}{n} + 1\right)$
t	0	$\frac{6}{n-4} + 3 \quad (n > 4)$
F	$\frac{8(2n+m-2)^2(m-4)}{n(m-6)^2(n+m-2)} \quad (m > 6)$	$\frac{(m-2)^3(m-4)(n+6)(n+4)(n+2)}{4(m-6)(m-8)(n+m-2)^2n^2} - \frac{8(m-4)(2n+m-2)}{(m-6)(n+m-2)} - \frac{3n(m-4)}{(n+m-2)} - \frac{n^2(m-4)}{4(n+m-2)^2}$ $(m > 8)$
Gamma	$\frac{4}{r}$	$\frac{6}{r} + 3$
Exponential	4	9
Weibull	$\left\{\Gamma\left(1 + \frac{3}{\beta}\right) - 3\Gamma\left(1 + \frac{1}{\beta}\right)\Gamma\left(1 + \frac{2}{\beta}\right) + 2\Gamma^3\left(1 + \frac{1}{\beta}\right)\right\}^2 *$ $* \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right]^3$	$\Gamma\left(1 + \frac{4}{\beta}\right) - 4\Gamma\left(1 + \frac{1}{\beta}\right)\Gamma\left(1 + \frac{3}{\beta}\right) + 6\Gamma^2\left(1 + \frac{1}{\beta}\right)\Gamma\left(1 + \frac{2}{\beta}\right) - 3\Gamma^4\left(1 + \frac{1}{\beta}\right) *$ $* \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right]^2$
LogNormal	$\left[\exp(\beta^2) - 1\right]\left[\exp(\beta^2) + 2\right]^2$	$\left[\exp(\beta^2) - 1\right]\left[\exp(3\beta^2) + 3\exp(2\beta^2) + 6\exp(\beta^2) + 6\right] + 3$
Beta	$\frac{4(\beta-\alpha)^2(\alpha+\beta+1)}{\alpha\beta(\alpha+\beta+2)^2}$	$\frac{3(2\alpha^2 + \alpha^2\beta - 2\alpha\beta + \alpha\beta^2 + 2\beta^2)(\alpha+\beta+1)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$
Uniform	0	1.8

From: Statistical Analysis for Engineers and Scientists by J.W. Barnes

Alternative Measures of Location/Center in a Distribution

When the population/process distribution has very heavy tails, the mean μ may not exist, for example, the Cauchy distribution. Also, when the distribution is heavily right or left skewed, the mean μ may be a very poor representation of the "center" or "typical value" of the population or process distribution. If both of these cases, we will seek an alternative to the mean μ.

The median $\tilde{\mu}$ is the most commonly used alternative to the μ as a representation of the center of a population distribution. To illustrate its performance relative to the mean for a highly skewed distribution we will examine the lognormal distribution:

Mean and Median of Lognormal Distribution

To illustrate the distribution of a population about its mean and median, we will next calculate the proportion of the values of a random variable Y less than its Mean μ_Y and the proportion less than its Median $\tilde{\mu}_Y$, when the distribution of Y is Lognormal distribution with parameters $\theta_1 = 0$, and θ_2, for various values of θ_2.

Let X have a $N(\theta_1, \theta_2^2)$ distribution then $\theta_1 = \mu_X = \tilde{\mu}_X$

- $Y = e^X$ has a lognormal distribution

- $\mu_Y = e^{\theta_1 + \frac{1}{2}\theta_2^2}$ (Using mgf of X)

- $\tilde{\mu}_Y = e^{\theta_1}$ This follows from

$$.5 = P[Y \leq \tilde{\mu}_Y] = P[X \leq log(\tilde{\mu}_Y)] \quad \Rightarrow \quad log(\tilde{\mu}_Y) = \tilde{\mu}_X = \theta_1$$

- Set $\theta_1 = 0$, then $\mu_Y = e^{\frac{1}{2}\theta_2^2}$ and $\tilde{\mu}_Y = e^0 = 1$. We then have that

$$P\left(Y \leq \mu_Y\right) = P\left(Z \leq \frac{\theta_2}{2}\right),$$

  where Z is N(0,1).

  This follows from

$$P(Y \leq \mu_Y) = P(log(Y) \leq log(\mu_Y)) = P\left(X \leq \frac{\theta_2^2}{2}\right) = P\left(\frac{X}{\theta_2} \leq \frac{\theta_2}{2}\right)$$

- By the definition of $\tilde{\mu}_Y$, $P(Y \leq \tilde{\mu}_Y) = 0.5$

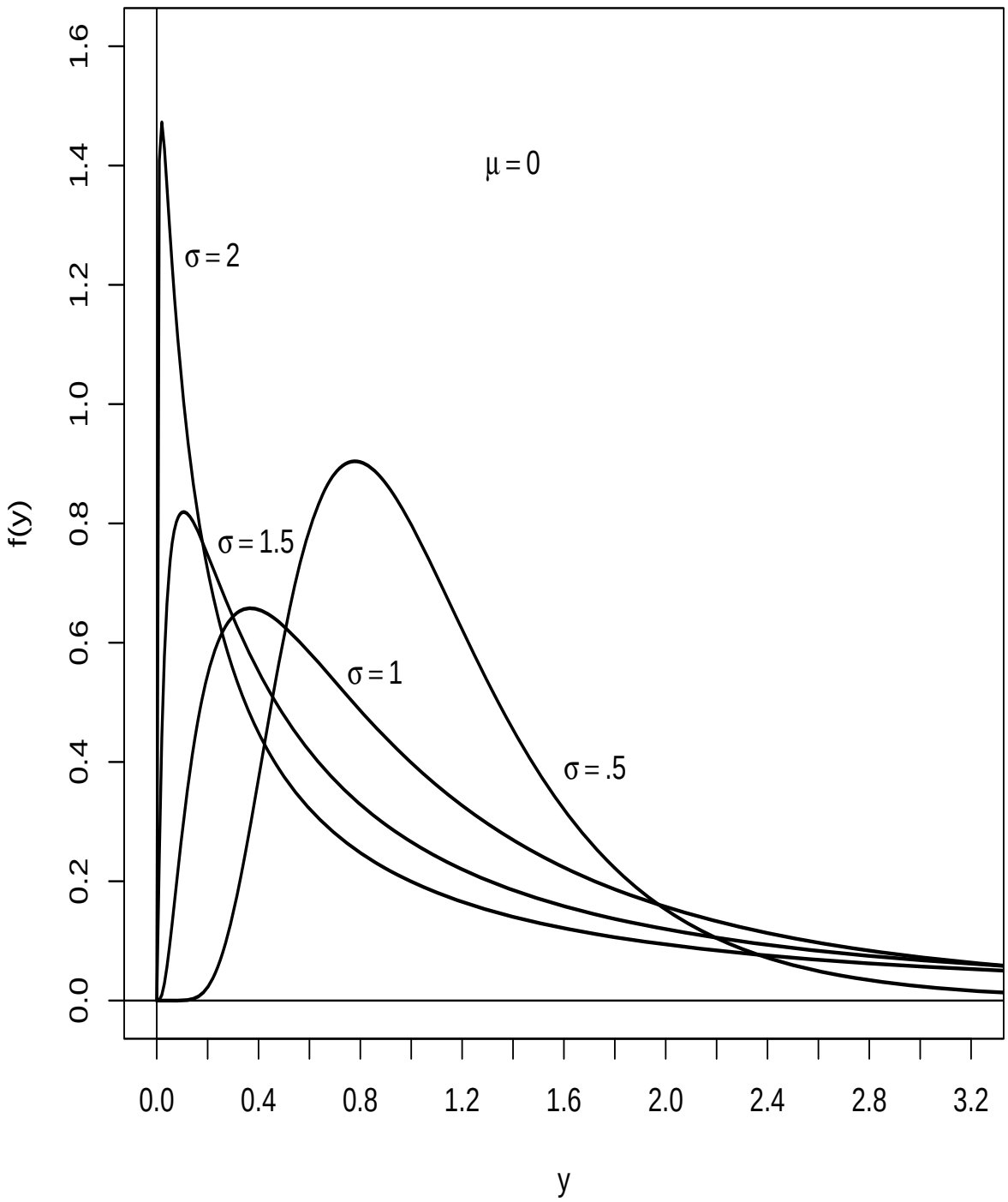
For the lognormal distribution with parameters, θ_1 and θ_2, the skewness parameter is given by

$$\beta_1 = \left(e^{\theta_2^2} - 1\right)\left(e^{\theta_2^2} + 2\right)^2$$

θ_2	β_1	μ_Y	$\tilde{\mu}_Y$	$P(Y \leq \mu_Y)$	$P(Y \leq \tilde{\mu}_Y)$
0.50	3.06e+00	1.1	1	0.599	0.5
0.75	1.06e+01	1.3	1	0.646	0.5
1.00	3.82e+01	1.6	1	0.691	0.5
1.25	1.72e+02	2.1	1	0.734	0.5
1.50	1.12e+03	3.0	1	0.773	0.5
1.75	1.11e+04	4.6	1	0.809	0.5
2.00	1.71e+05	7.3	1	0.841	0.5
2.25	4.01e+06	12.5	1	0.870	0.5
2.50	1.39e+08	22.7	1	0.894	0.5
2.75	7.14e+09	43.8	1	0.915	0.5
3.00	5.32e+11	90.0	1	0.933	0.5
3.25	5.77e+13	196.6	1	0.948	0.5
3.50	9.12e+15	457.1	1	0.960	0.5
3.75	2.09e+18	1131.4	1	0.970	0.5
4.00	7.01e+20	2980.9	1	0.977	0.5
4.25	3.41e+23	8360.3	1	0.983	0.5
4.50	2.41e+26	24959.2	1	0.988	0.5
4.75	2.49e+29	79320.3	1	0.991	0.5
5.00	3.73e+32	268337.2	1	0.994	0.5

From the table, we can observe

- For highly skewed distributions, the mean μ is not an appropriate representation of the center of the distribution. Even though there is only a small percent of the distribution in the tail of a highly skewed distribution, these few very large values (relative to the rest of the distribution), cause the mean of the distribution to be very large.
- The median does not have this flaw. Whereas, the mean is the weighted average of all values for the r.v., the median is not affected by the extreme values for the r.v.
- The median is the true middle value dividing the distribution into two equal parts irregardless of the size of the extremes in the population/process.
- The median $\tilde{\mu}$ is the more appropriate parameter in these situations for representing the center of the distribution or the typical value for the population/process.
- Suppose we have a population distribution which is highly right skewed. Based on a random sample of n observation, we want to estimate the population total, $T = \sum_{i=1}^{N} Y_i$. Should we use $\widehat{T} = N\widehat{\mu}$ or $\widehat{T} = N\tilde{\mu}$, mean or median as the "typical value" ?


LogNormal Density Functions
μ = 0
σ = 2
σ = 1.5
σ = 1
σ = .5
f(y)
y
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2

In general we will make use of all three quartiles when describing the "center" of a population distribution:

$$Q_1 = Q(.25) \qquad \tilde{\mu} = Q_2 = Q(.5) \qquad Q_3 = Q(.75)$$

Note that the middle 50% of the distribution falls between Q_1 and Q_3, with 25% between Q_1 and $\tilde{\mu}$ and 25% between $\tilde{\mu}$ and Q_3.

How is the median and quartiles related to the mean?

From our results from the lognormal distribution it would appear there is no relationship between the mean and median. However, we know that

$$\mu_Y = e^{\theta_1 + \frac{1}{2}\theta_2^2} = \tilde{\mu}_Y e^{\frac{\theta_2^2}{2}}$$

For the lognormal distribution with parameter $\theta_1 = 0$, as the lognormal distribution became more skewed the mean of the distribution increased to a very large value whereas as the median remained constant. However, the standard deviation of the lognormal also was increasing very rapidly as can be seen from

$$\sigma_Y = e^{\theta_1}\sqrt{e^{\theta_2^2} - 1}$$

In fact, we have shown that for Y distributed logNormal(θ_1, θ_2),

$$\mu_Y = \tilde{\mu}_Y e^{\theta_2^2/2}$$

For the lognormal distribution, the mean of Y is related to both the median and the θ_2 parameter.

We can establish the following relationship between any percentile $Q(p)$ and the pair (μ, σ) for any distribution which has $|\mu| < \infty$ and $\sigma < \infty$:

$$|Q(p) - \mu| \;\leq\; \sigma \max\left\{\sqrt{\frac{1-p}{p}},\quad \sqrt{\frac{p}{1-p}}\right\}.$$

In particular, the median (p=1/2) satisfies

$$|\tilde{\mu} - \mu| \leq \sigma$$

That is, for any distribution for which the mean and standard deviation exist, the median and mean differ by at most one standard deviation.

From this statement we cannot infer that the mean will necessarily be *close* to the median.

Why?

Recall the lognormal example in which σ can be very large for large values of θ_2.

Trimmed Mean

When a distribution is very heavy-tailed, extremes to the central values can occur with a reasonably high frequency (Cauchy, logistic, t with small df). In these situations, the mean μ can be drawn a considerable distance from the median $\tilde{\mu}$.

An alternative to the mean, other than the median, which reduces the influence of extremes in the distribution is the trimmed mean. In heavy-tailed distributions, we "trim" off the extreme values prior to averaging the values in the population/process.

Definition: The $\alpha-$Trimmed Mean $\mu_{(\alpha)}$ is the mean of the population after excluding the smallest and largest $100\alpha\%$ of the distribution.

Let Y have pdf f and quantile function Q. After trimming off the smallest and largest $100\alpha\%$ of the distribution, it is necessary to renormalize the area under the pdf so that the total area is 1.

This results in the following pdf for the $\alpha-$trimmed distribution:

$$f_{(\alpha)}(y) = \begin{cases} \frac{1}{1-2\alpha} f(y) & if \quad Q(\alpha) \leq y \leq Q(1-\alpha) \\ 0 & if \quad y < Q(\alpha) \\ 0 & if \quad y > Q(1-\alpha) \end{cases}$$

$\frac{1}{1-2\alpha}$ is a normalizing constant which guarantees that the total area under $f_{(\alpha)}(y)$ is 1.

Using this pdf for the trimmed distribution, we obtain the following formula for the $\alpha-$trimmed mean:

$$\mu_{(\alpha)} = \int_{-\infty}^{\infty} y f_{(\alpha)}(y) dy = \frac{1}{1-2\alpha} \int_{Q(\alpha)}^{Q(1-\alpha)} y f(y) dy.$$

In the limit we obtain both the mean and median from the $\alpha-$trimmed mean:
As the amount of trimming is reduced, the trimmed mean converges to the population mean:

$$\lim_{\alpha \to 0} \mu_{(\alpha)} = \mu$$

As the amount of trimming is increased to 50%, the trimmed mean converges to the population median. (See Assignment 4 for a proof).

$$\lim_{\alpha \to .5} \mu_{(\alpha)} = \tilde{\mu}$$

Measures of Dispersion/Variability in a Population/Process Distribution

In a manufacturing process, a product is produced having a nominal physical characteristic of size θ_o:

- Producing ball bearings with a nominal diameter of 5 cm ($\theta_o = 5cm$)
- Producing an antibiotic with a nominal weight of 10 mg ($\theta_o = 10mg$)
- Producing gasoline with a nominal amount of an additive of .8% ($\theta_o = .8\%$)

In such situations, we want to determine

1. If the "average" value of the physical characteristic equals the nominal value and
2. How near are the values of the physical characteristic in a daily production run to this "average" value.

For example, suppose our daily production of ball bearings comes from two processes:

> Process 1: $\mu = 5$cm with 15% of output less than 4.8 cm and 15% of output greater than 5.2 cm.
>
> Process 2: $\mu = 5.002$cm with .1% of output less than 4.8 cm and .12% of output greater than 5.2 cm.

Which process is doing a better job of producing ball bearings with a diameter of 5 cm?

We need to evaluate both the location and the dispersion of the measured characteristic in a population/process.

There are a number of possible methods to measure the dispersion of the population/process about a central value. We will now define a few of these parameters.

1. **Definition:** **The Range** is the distance from the smallest possible value to the largest possible value in the population.

   For most distributions, the range is not very useful because it is generally infinity. Also, it depends on just two values and does not describe to any degree the clumping or lack of clumping of values about the measure of the center of the distribution.

   normal distribution has Range $= 2\infty$, Weibull has Range $= \infty$,

   Beta on [0,1], has Range =1.

2. **Definition:** **The Standard Deviation** is a measure of the concentration of the values about the population/process mean μ:

$$\sigma = \sqrt{E\left[(Y-\mu)^2\right]} = \sqrt{\int_{-\infty}^{\infty}(y-\mu)^2 dF(y)}$$

   σ is the "weighted average" distance of values about μ

   For very heavy-tailed distributions μ and/or σ may not exist (Cauchy and t with df$\leq$ 2)

   Highly skewed and heavy-tailed distributions result in highly inflated values of σ which may not represent a high concentration of values about μ with just a small proportion of values extreme to μ.

3. **Definition:** **The Interquartile Range (IQR)** measures the distance to cover the middle 50% of distribution

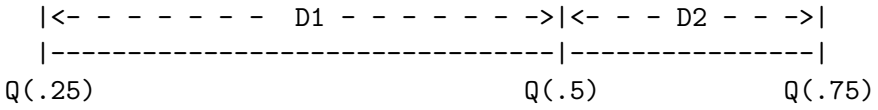
$$IQR = [(Q(.75) - Q(.25)]$$

4. **Definition:** **The Semi-Interquartile Range (SIQR)** measures the spread of the middle 50% of distribution about the median $\tilde{\mu}$:

$$SIQR = \frac{1}{2}[(Q(.75) - \tilde{\mu}) + (\tilde{\mu} - Q(.25)] = \frac{1}{2}(D_2 + D_1) = \frac{1}{2}IQR$$

   SIQR is the average distance from the lower quartile to the median, D_1 and from median to the upper quartile, D_2.

   For very skewed distributions and very heavy-tailed distributions, the pair $(\tilde{\mu}, SIQR)$ is often a better representation of center, variability than is (μ, σ). However, it only is dealing with the middle 50% of the data (from Q(.25) to Q(.75)) and thus is trimming 25% of the data in both tails prior to measuring spread. Too insensitive to tail behavior.



An alternative to the SIQR for very skewed distributions and very heavy-tailed distributions is **MAD**:

5. **Definition: The Median Absolute Deviation about the Median**

$$MAD = median\{|Y - \tilde{\mu}|\}$$

MAD measures the dispersion of the data about the median by taking the median absolute distance from the values in the population/process to the median of the population/process.

MAD is not affected by extremes in the population/process

Several versions of MAD

(a) Mean absolute deviation about the mean:

$$MAD_1 = E[|Y - \mu|]$$

Eliminates the problem in σ due to squaring distance for extreme observations but still is taking average over all values in population/process and still uses μ as the measure of center

(b) Mean absolute deviation about the median:

$$MAD_2 = E[|Y - \tilde{\mu}|]$$

Eliminates the problem of using μ as the measure of center but still is taking average over all values in population/process

(c) Median absolute deviation about the median:

$$MAD = Median|Y - \tilde{\mu}|$$

Eliminates the problem of using μ as the measure of center and eliminates problem of including the extremes of the population/process in computing the average distance

Some further observations about MAD.

1. Let Y be a r.v. with median $\tilde{\mu}$ and let $W = |Y - \tilde{\mu}|$.

   Then MAD is just the median of the distribution of W.

2. For a symmetric distribution with $\mu < \infty$ we have

$$\mu = \tilde{\mu} \quad \text{and} \quad MAD = SIQR$$

3. For a $N(\mu, \sigma^2)$ distribution,

$$\mu = \tilde{\mu} \quad \text{and} \quad MAD = 0.6745\sigma$$

   Normal distribution is symmetric $\Rightarrow$ MAD = SIQR

   $Q(.25) = \mu + Q_Z(.25)\sigma = \mu - .6745\sigma$

   $Q(.75) = \mu + Q_Z(.75)\sigma = \mu + .6745\sigma$

   MAD = SIQR = $\frac{1}{2}(Q(.75) - Q(.25)) = .6745\sigma$

4. Often MAD will be defined as

$$MAD = median|Y - \tilde{\mu}|/.6745$$

   For a normal distribution MAD and σ are the same. This is the definition used in R.

5. For the logNormal(θ_1, θ_2) distribution, $\sigma = e^{\theta_1}\sqrt{e^{\theta_2^2}\left[e^{\theta_2^2} - 1\right]}$

   For $\theta_1 = 0$ and $MAD = median(|Y - \tilde{\mu}|)/0.6745$, we have

	θ_2					
	.5	.75	1.0	1.5	2.0	2.5
σ	.604	.1.15	2.16	8.97	54.1	517.5
MAD	.485	.70	.89	1.17	1.34	1.42

Recommendations on the Selection of Measures of Center/Dispersion about Center

Why not just use the pair $(\tilde{\mu}, MAD)$ for all distributions?

1. When the population/process distribution is not heavily skewed or too heavily tailed (near normal in shape), (μ, σ) provide a more complete picture of the distribution.

   Furthermore, when we use data to estimate these parameters, the sample counterparts of (μ, σ) are more efficient estimators than are the the sample counterparts of $(\tilde{\mu}, MAD)$

2. When the population total $T = \sum_{i=1}^{N}$ is the parameter of interest to the researcher, the sample estimator $\hat{T} = N\hat{\mu}$ is a better estimator of T than is the estimator $\hat{T} = N\hat{\tilde{\mu}}$.

   When estimating the population total we will in most cases want the effect of extremes to be a part of the estimator.

   Example Let T = daily amount of pollutants discharged in a river from a chemical plant. The days in which there was a large discharge would have a large impact on the yearly total. Thus, in estimating the year total discharge, YT $= \sum_{i=1}^{365} T_i$ we would choose between 365 times the average daily discharge and 365 times the median daily discharge. However, $365\hat{\tilde{\mu}}_T << 365\hat{\mu}_T$. Thus, we could greatly underestimate YT if we used median in place of mean. However, if we wanted to obtain information on the "typical" daily discharge the median may be a more appropriate representative than the mean.

3. For highly skewed and/or heavy-tailed distributions use $(\tilde{\mu}, MAD)$.

   For Cauchy and t with $df \leq 2$, μ and σ do not exist but $(\tilde{\mu}, MAD)$ always exist.

4. If $(\theta_1,\ \theta_2)$ are location-scale parameters for a distribution, is the following true?

$$\mu = \theta_1 \quad \text{and} \quad \sigma = \theta_2$$

   a. If the distribution is Cauchy$(\theta_1,\ \theta_2)$, then μ and σ do not exist but $(\theta_1,\ \theta_2)$ are valid location-scale parameters

   b. If the distribution is Uniform on $(\theta + a,\ \theta + b)$ then θ is a location parameter but

$$\mu = \theta + \frac{a+b}{2} \neq \theta \quad \text{and} \quad \sigma = \sqrt{(b-a)^2/12}$$

   c. If the distribution is logistic$(\theta_1,\ \theta_2)$ then $(\theta_1,\ \theta_2)$ are valid location-scale parameters.

$$\mu = \theta_1 \quad \text{but} \quad \sigma = \frac{\pi\theta_2}{\sqrt{3}} \neq \theta_2$$

Measures of Association Amongst Vectors of R.V.s

Suppose we are dealing with a population/process where we measure several variables on each unit in the population/process or we observe a single characteristic of the unit across time or space. In these situations we want to determine the degree to which these variables are related.

Example 1: Suppose a company is developing a new production process for producing an alloy used in golf clubs. The company would be interested in measuring:

(a) X_1 - Rockwell hardness of alloy

(b) X_2 - Tensile strength of alloy

(c) X_3 - Smoothness of surface of alloy

(d) X_4 - % Carbon in alloy

Example 2: Suppose a new therapy of treating patients with high blood pressure is under evaluation. The medical doctors would be interested in measuring on each patient:

(a) W_1 - Age

(b) W_2 - Blood pressure prior to starting therapy

(c) W_3 - Blood pressure at conclusion of therapy

(d) W_4 - BMI

(e) W_5 - Cholestrol level

(f) W_6 - Hours per week of exercise

(g) W_7 - Yes or No for a stroke

Example 3: Suppose we are investigating how long a carcinogen in an industrial discharge remains in the atmosphere

D_0 - Amount of carcinogen in air sample immediately prior to discharge

D_1 - Amount of carcinogen in air sample 1 minute after discharge

D_2 - Amount of carcinogen in air sample 2 minutes after discharge

.

.

.

D_{150} - Amount of carcinogen in air sample 150 minutes after discharge

In each of the three examples the researchers would be interested in measuring the degree of association between the vector of r.v.s.

Definition: **The Correlation** between two random variables Y and W having means and standard deviations: $\mu_Y, \mu_W, \sigma_Y, \sigma_W$, respectively, is given by

$$\rho_{Y,W} = Corr(Y,W) = \frac{E[(Y-\mu_Y)(W-\mu_W)]}{\sigma_Y \sigma_W} = \frac{Cov(Y,W)}{\sigma_Y \sigma_W}$$

* $Cov(Y,W) = E(YW) - E(Y)E(W)$

1. Correlation is a unit-free measure of the linear relationship between the two variables.
2. $-1 \leq \rho_{Y,W} \leq 1$
3. $\rho_{Y,W} = \pm 1$ implies $Y = \beta_o + \beta_1 W$ where the sign of β_1 is the same as the sign of $\rho_{Y,W}$
4. $\rho_{Y,W}$ has the limitation of only measuring linear relationship.

   Thus, higher order relationships may not be detected

   Example: Let X have a symmetric distribution with mean μ and variance σ^2.

   Let $Y = (X-\mu)^2$. Thus, $E[Y] = \sigma^2$.

   $Corr(X,Y) = 0$ because

   $E[(X-\mu)(Y-\sigma^2)] = E[(X-\mu)^3] - \sigma^2 E[(X-\mu)] = 0 - 0 = 0$

   Thus, Y and X are uncorrelated but they are perfectly related by $Y = (X-\mu)^2$, a nonlinear relationship.

5. $\rho_{Y,W}$ only measures linear relationships between two of the many variables under study.

   Thus, may fail to detect nonlinear relationships that exist between several of the variables simultaneously. For example, $W_3 = W_1 e^{W_2}$

6. If X and Y are independent, $Corr(X,Y) = 0$:

   $E[(X-\mu_X)(Y-\mu_Y)] = E[(X-\mu_X)]E[(Y-\mu_Y)] = (E(X)-\mu_X)(E(Y)-\mu_Y)] = (0)(0)$

- The converse is not true. That is, $Corr(X,Y) = 0$ does NOT imply that X and Y are independent.

  Counter Example

  Let X have a pdf $f(x-\theta)$ which is symmetric distribution about 0 with $\mu_X = E[X] = \theta$

  Let $Y = I(|X-\theta| < 2)$, then X and Y are not independent:

  $$P[X > \theta + 3] \ > 0 \ \text{ but } \ P[X > \theta + 3 | Y = 1] = 0$$

  However,

  $\mu_Y = E[Y] = P[|X-\theta| < 2] = \int_{\theta-2}^{\theta+2} f(x-\theta)dx = \int_{-2}^{2} f(t)dt$

$$\begin{aligned} E[XY] = \int_{-\infty}^{\infty} xI(|x-\theta| < 2)f(x-\theta)dx &= \int_{\theta-2}^{\theta+2} xf(x-\theta)dx \\ &= \int_{-2}^{+2} (t+\theta)f(t)dt \\ &= \int_{-2}^{+2} tf(t)dt + \theta\int_{-2}^{+2} f(t)dt \\ &= 0 + \theta E[Y] = E[X]E[Y] \end{aligned}$$

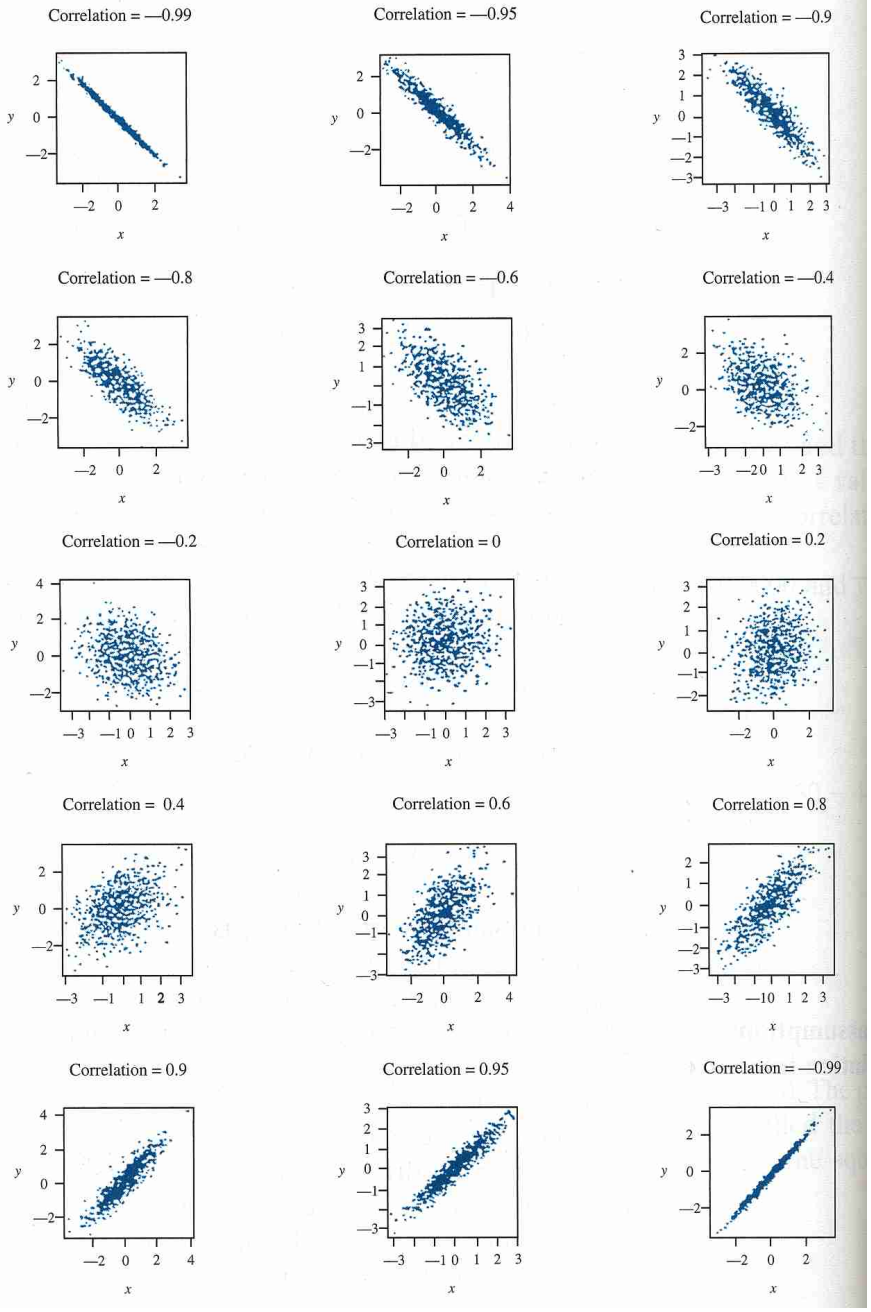
We used the fact that $f(x-\theta)$ was symmetric about 0 to conclude that $\int_{-2}^{+2} tf(t)dt = 0$
Therefore, $E[XY] = E[X]E[Y]$ which implies $Cov(X,Y) = 0$ which implies $Corr(X,Y) = 0$
However, X and $Y = I(|X-\theta| < 2)$ are obviously not independent.

Special Case: If (X,Y) have a bivariate normal distribution, then

$Corr(X,Y) = 0$ implies X and Y are independent

The following scatter plots from *An Introduction to Statistical Methods and Data Analysis* will give an indication of what correlation is measuring.

FIGURE 11.20
Samples of size 1,000 from the bivariate normal distribution



START CN: 9/24/21 (Friday)

Time Series Data

When we are observing a physical characteristic over time, we are interested in the degree to which these measurements are associated. One measure of this association is the autocorrelation:

Definition: **The AutoCorrelation of Order k** in a series of stationary random variables: $X_t : t = 1, 2, 3, \ldots$ having the same mean μ and standard deviation σ is given by

$$\rho_k = \frac{E[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2} \quad \text{for} \quad k = 1, 2, \ldots$$

↳ correlation btwn obs at time t & obs at time k.

ρ_k measures the degree of linear relationship in the random variable X over time or space.

ρ_1 the 1st order autocorrelation is the most widely used of these correlations.

A very simple, but widely used model for correlated over time or space observations is the $AR(1)$ model:

$$X_t = \theta + \rho X_{t-1} + e_t,$$

where $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma_e^2$, $e_t s$ are independent of the $X_t s$ and $|\rho| < 1$.

Under this model, we can show that:

- $X_t s$ have mean:

$$\mu = E[X_t] = \theta + \rho E[X_{t-1}] + E[e_t] = \theta + \rho\mu + 0 \quad \Rightarrow \quad \mu = \frac{\theta}{1-\rho} \; *$$

- $X_t s$ have variance:

$$\sigma_X^2 = Var(X_t) = Var(\theta + \rho X_{t-1} + e_t) = \rho^2 Var(X_{t-1}) + Var(e_t) = \rho^2 \sigma_X^2 + \sigma_e^2 \Rightarrow \sigma_X^2 = \sigma_e^2/(1-\rho^2) \; *$$

- $X_t s$ are not independent:

$$\begin{aligned} Cov(X_t, X_{t-1}) &= E[(X_t - \mu)(X_{t-1} - \mu)] \\ &= E[(\theta + \rho X_{t-1} + e_t - \mu)(X_{t-1} - \mu)] \\ &= \theta E[X_{t-1} - \mu)] + \rho E[(X_{t-1})^2] - \rho\mu E[X_{t-1}] + E[(e_t - \mu)(X_{t-1} - \mu)] \\ &= 0 + \rho(\sigma^2 + \mu^2) - \rho\mu^2 + E[e_t - \mu]E[X_{t-1} - \mu] \\ &= 0 + \rho(\sigma^2 + \mu^2) - \rho\mu^2 + 0 \\ &= \rho\sigma^2 > 0 \;\Rightarrow\; X_t \text{ and } X_{t-1} \text{ are Not independent} \end{aligned}$$

- $\rho_k = Corr(X_t, X_{t-k}) = \rho^k \to 0$ as $k \to \infty$ because $|\rho| < 1$

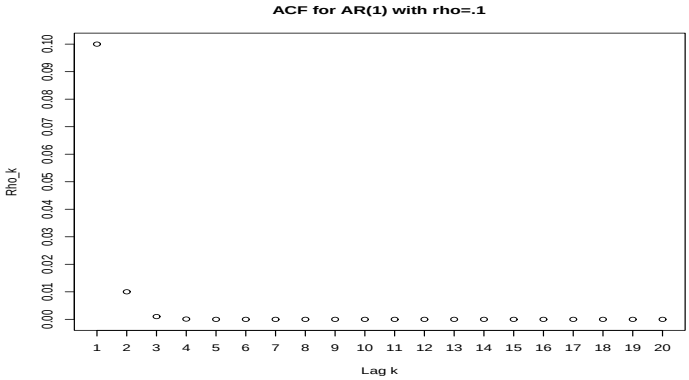
The plots on the next pages display the autocorrelation function (acf) for an AR(1)

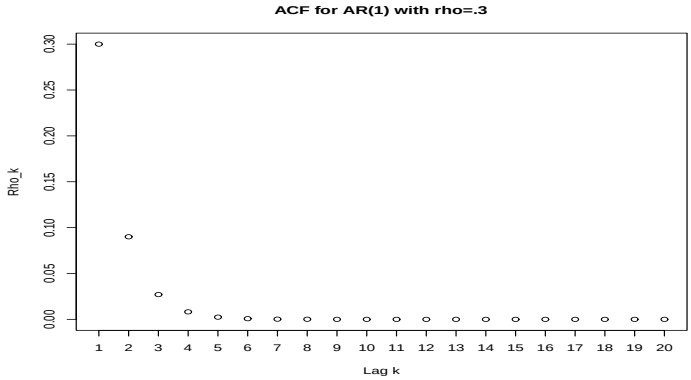
with $\rho = .9, .6, .3, .1$.

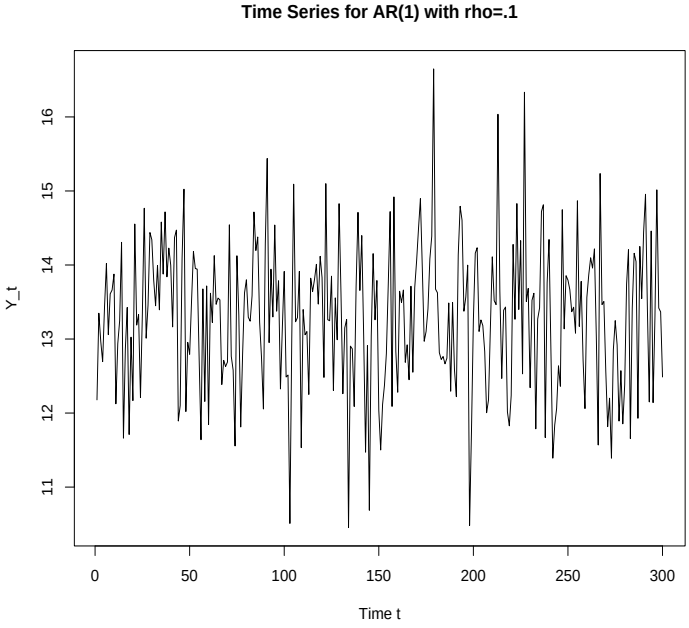
Also, there are corresponding plots of the time series obtained by simulating 300 observations from an AR(1)

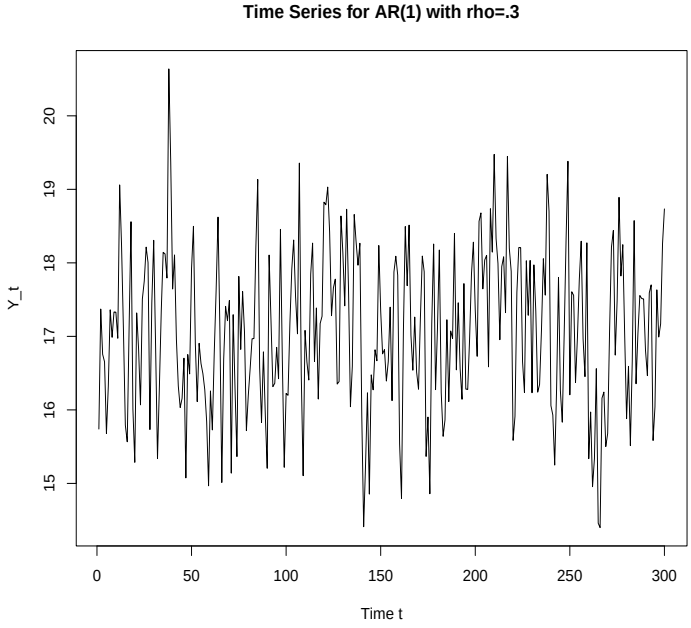
with $\rho = .9, .6, .3, .1$ using the model

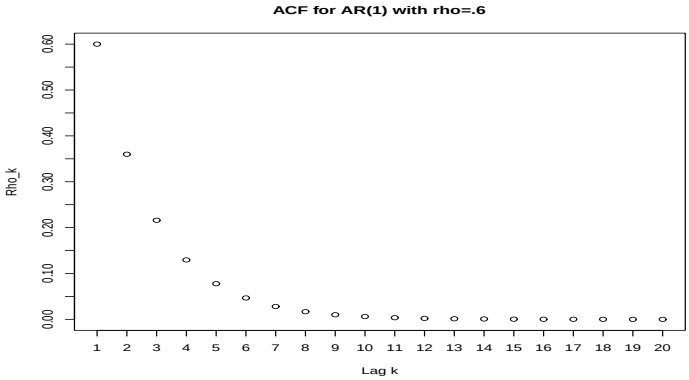
$$X_t = 12 + \rho X_{t-1} + e_t.$$

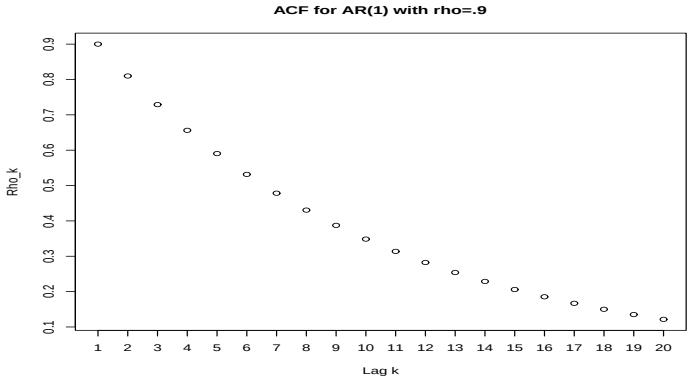

ACF for AR(1) with rho=.1
Rho_k
Lag k

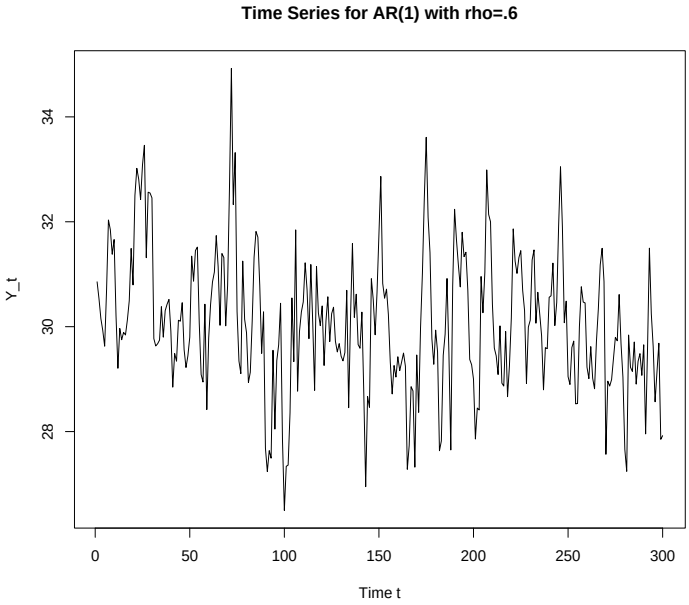

ACF for AR(1) with rho=.3
Rho_k
Lag k

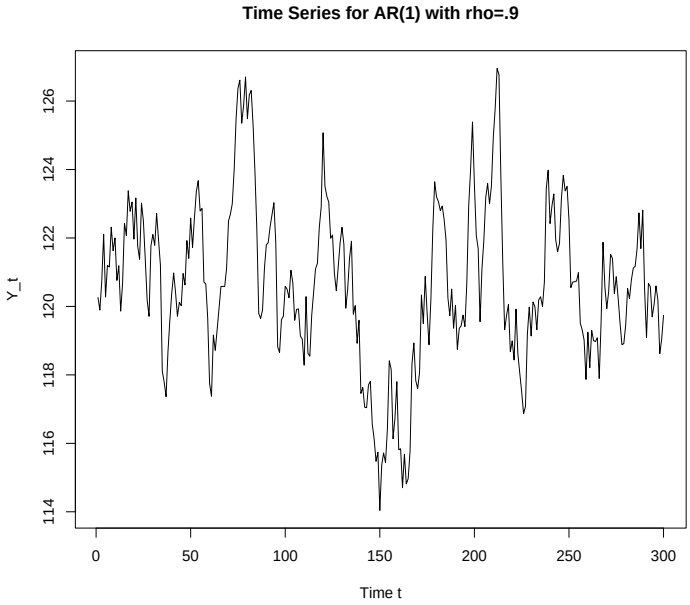

Time Series for AR(1) with rho=.1
Y_t
Time t


Time Series for AR(1) with rho=.3
Y_t
Time t


ACF for AR(1) with rho=.6
Rho_k
Lag k


ACF for AR(1) with rho=.9
Rho_k
Lag k


Time Series for AR(1) with rho=.6
Y_t
Time t


Time Series for AR(1) with rho=.9
Y_t
Time t

The plots on the previous pages were of stationary time series, that is, time series in which the mean and variance remained constant over time.

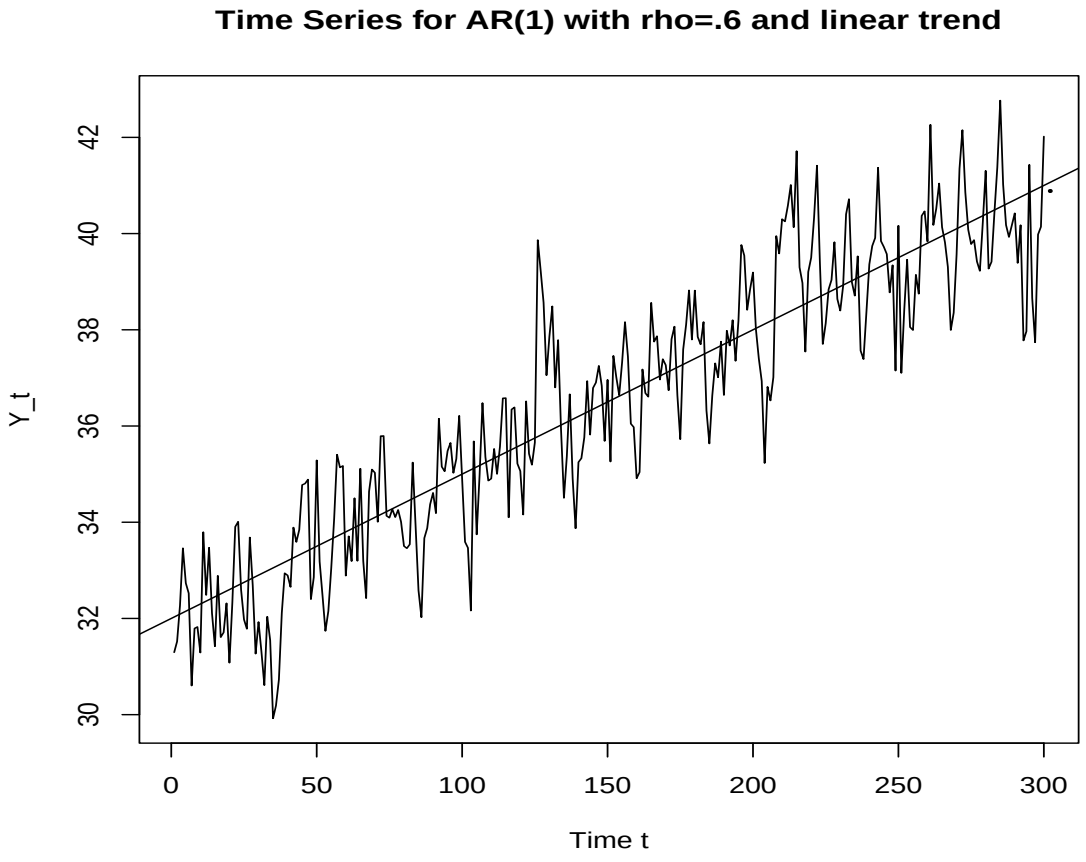
$$\mu_t = E[X_t] = \mu \quad \text{and} \quad \sigma_t^2 = Var(X_t) = \sigma^2 \quad \text{for} \quad t = 1, 2, 3, 4, \ldots$$

In many applications, the time series will not be stationary but will have a trend in its mean and/or variance.

In the plot given below we can observe that there is an increase in the mean of X_t with increasing t. This type of behavior often occurs when we are studying such measures as monthly sales or temperature or many other physical processes.

The process generating the data is given by

$$Y_t = 2 + .03t + .6Y_{t-1} + e_t \quad \text{where } e_t \text{ are iid N(0,1) r.v.s}$$



Finished Friday 9/24/21

START CN: Friday 9/24/21

HANDOUT #6: NUMERICAL SUMMARIES OF DATA

ESTIMATORS FOR PARAMETRIC FAMILIES

1. **Graphical Estimators of Location-Scale Parameters**
2. **Method of Moment Estimators (MOM)**
3. **Maximum Likelihood Estimators (MLE)**
4. **Pdf Based Estimators of Summary Parameters**
5. **Distribution-Free Summaries**
   (a) **Estimators of Measures of Location**
      i. Sample Mean ($\hat{\mu} = \bar{Y}$)
      ii. Sample Median ($\widetilde{Y} = \widehat{Q}_2$)
      iii. Sample Quartiles ($\widehat{Q}_1, \widehat{Q}_3$)
      iv. Trimmed Mean ($\hat{\mu}_{(\alpha)}$)
   (b) **Estimators of Measures of Level of Dispersion**
      i. Range (R)
      ii. Semi-interquartile Range ($SIQR$)
      iii. Sample Standard Deviation ($\hat{\sigma} = S$)
      iv. Mean Absolute Deviation (MAD)
   (c) **Five Number Summary of Data Set:**

      Minimum, $\hat{Q}(.25)$, Median, $\hat{Q}(.75)$, Maximum
   (d) **Estimators of Shape of PDF**
      i. Sample Skewness: $\hat{\beta}_1$
      ii. Sample Kurtosis: $\hat{\beta}_2$
   (e) **Estimators of Measures of Correlation**
      i. Sample Correlation Coefficient: Pearson and Spearman ($\hat{\rho}$, $\hat{\rho}_{sp}$)
      ii. Sample Autocorrelation Coefficient of Lag k ($\hat{\rho}_k$)

Supplemental Reading:

- Chapters 2 and 4 in the Tamhane/Dunlop book

Planning a Comparison of Several Populations/Processes

The researcher must carefully design the study by addressing the following questions:

1. What is the specific population or process that is of interest. Carefully specify the particular conditions and limitations associated with the process or population.

- Environmental conditions in Lab, Differences in Technicians, Differences in Equipment

2. What characteristics of the population/process need to be measured?

- Blood pressure, severity of disease, reduction in pollution after new equipment is installed

3. How much data needs to be collected?

- Based on how much accuracy is needed, how much risk of erroneous conclusions is acceptable, how variability in population

4. How will the data be collected?

- Sampling Design: simple random sample, stratified sample, cluster sampling, observational study, historical data

- Experimental Design: completely randomized design, randomized block design, split plot design

5. Is the data independent or correlated temporally/spatially?

- Collected over time
- multiple measurements in close proximity
- observations over a grid on a potential oil field

6. How will the data be summarized?

   Graphically Numerically

7. What comparisons need to be made?
8. To what degree of accuracy do we need to make the comparisons?

- $\widehat{\mu} \pm \Delta$
- How large a difference in $\widehat{\mu}_1 - \widehat{\mu}_2$ is a practical difference?

SUMMARIES FOR PARAMETRIC FAMILIES

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: $\theta_1, \theta_2, \cdots, \theta_k$. Suppose we want to estimate certain population parameters. There are a number of possible methods for obtaining the estimators of the parameters. We will consider three such approaches. Once the estimators of the pdf's parameters are obtained, the population summary parameters are obtained by replacing the unknown parameters involved in these summaries with their sample estimators.

We will illustrate these ideas with an example:

Suppose $Y_1, Y_2, \ldots, Y_n$ be the times to failure of a random sample of n cell phones which are produced in a newly designed production facility. From historical reliability records, failure times for this type of phone have a Weibull cdf $F(y) = 1 - e^{(y/\alpha)^\gamma}$ but both α and γ are unknown due to the changes in the production process.

The mean and standard deviation of Y are given by

$$\mu = \alpha\,\Gamma\left(1 + \frac{1}{\gamma}\right) \qquad \sigma = \sqrt{\alpha^2\left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right]}$$

where the gamma function is defined as follows: $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\mathrm{d}y$

First we use the data to obtain estimators of α and γ, $\widehat{\alpha}$ and $\widehat{\gamma}$, respectively, then we substitute these estimators into the formulas for μ and σ to obtain the corresponding estimators of the mean and standard deviation of the Weibull distribution:

$$\widehat{\mu} = \widehat{\alpha}\,\Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) \qquad \widehat{\sigma} = \sqrt{\hat{\alpha}^2\left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]}$$

We will now discuss several methods for obtaining the point estimators of the unknown parameters in the cdf.

Method 1: Graphical Estimators of Location-Scale Parameters

If the specified cdf is a location-scale family of cdfs, for example,

1. Normal(μ, σ^2)
2. Logistic(μ, β)
3. Exponential(β)
4. Weibull(γ, α)
5. Cauchyl(θ_1, θ_2)

then we can use a graphical procedure which use reference distribution plots to obtain *rough* estimators of the location and scale parameters.

Let $Q_o(u)$ be the quantile function of the standard member of the family and $\widehat{Q}(u)$ be the sample quantile for the n data values $Y_1, Y_2, \ldots, Y_n$.

Plot $\widehat{Q}(u_i)$ versus $Q_o(u_i)$ for $u_i = \frac{i-.5}{n}$; $\quad i = 1, 2, \ldots, n$.

If the n plotted points are reasonably close to a straight-line then **graphical estimators** of θ_1 and θ_2 are given by

$\widehat{\theta_1}$ = Y-intercept of fitted line

$\widehat{\theta_2}$ = Slope of fitted line

We will discuss in detail **Reference Distribution Plots** in Handout 8.

EXAMPLE The time to failure, in 100 hours, of a random sample of 25 newly designed fuel pumps are recorded as follows:

15.321	9.008	20.104	7.729	45.154	8.404	5.332	0.577	4.305
4.517	12.594	6.829	3.291	37.175	0.841	1.317	7.613	20.582
2.030	10.001	4.666	12.933	0.591	39.454	8.875		

The researcher states that from previous studies that the Weibull distribution was a good approximation to cdf of the r.v. Y, Time to Failure. The cdf of Y is given by

$$F_Y(y) = 1 - e^{-(y/\alpha)^\gamma}.$$

From the form of the cdf of Y, it can be observed that α is a scale parameter but γ is a shape parameter. Therefore, the Weibull family of cdf's is not a location-scale family.

However, a transformation of the data to $W = log(Y)$ yields the following results:

$W = log(Y)$ has cdf given by

$$\begin{aligned} F_W(w) = P[W \le w] = P[log(Y) \le w] = P[Y \le e^w] &= 1 - e^{-(e^w/\alpha)^\gamma} \\ &= 1 - e^{-(e^{\gamma w}/\alpha^\gamma)} \\ &= 1 - e^{-e^{\gamma w - log(\alpha^\gamma)}} \\ &= 1 - e^{-e^{\gamma(w - log(\alpha))}} \\ &= 1 - e^{-e^{(w-log(\alpha))/\frac{1}{\gamma}}} \end{aligned}$$

From the above we can conclude that the family of cdf's for W = log(Y) is a location-scale family with

$$\text{location parameter, } \theta_1 = log(\alpha) \quad \text{and scale parameter,} \quad \theta_2 = \tfrac{1}{\gamma}$$

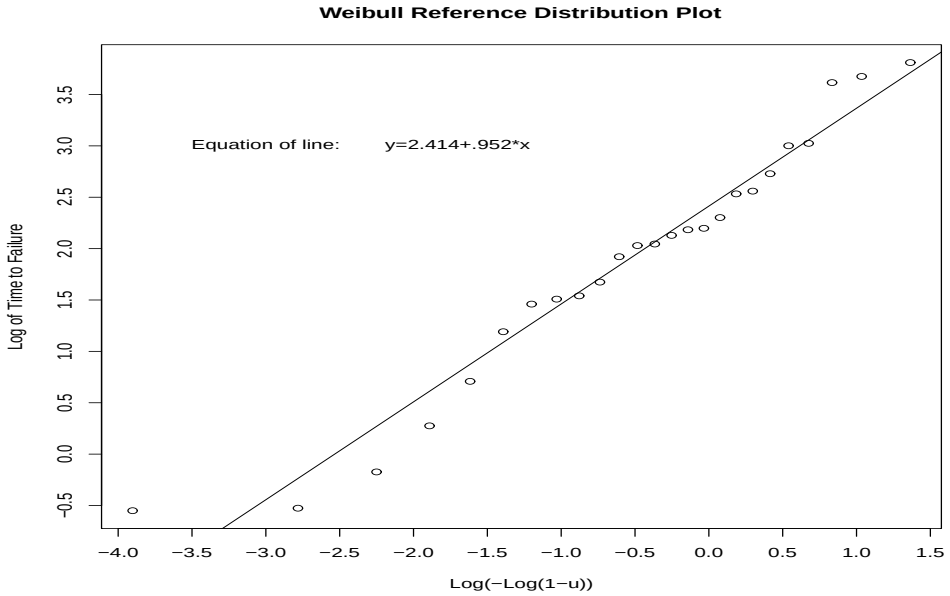
Plot the data on a Weibull Reference Distribution Plot with the sample quantile function for $W_i = log(Y_i)$, $\widehat{Q}_W(u_i)$ on the vertical axis and the standard member,

$\theta_1 = 0,\ \theta_2 = 1$, of the W family of quantiles, $Q_o(u_i)$, on the horizontal axis.

$\theta_1 = 0 \ \Rightarrow\ log(\alpha) = 0 \ \Rightarrow\ \alpha = 1$ and $\theta_2 = 1 \ \Rightarrow \gamma = 1$

Next, determine $Q_o(u) = F_o^{-1}(u)$, in the following manner:

$u = F_o(w_u) = 1 - e^{-e^{(w_u - 0)/1}} \ \Rightarrow\ Q_o(u) = w_u = log(-log(1-u))$



The 25 data values are relatively close to the fitted line:

$$\widehat{Q}_W(u_i) = 2.414 + .952 Q_o(u_i)$$

Thus, we can conclude that graphical estimators of θ_1 and θ_2 are given by

$$\widehat{\theta}_1 = 2.414 \quad \text{and} \quad \widehat{\theta}_2 = .952$$

From these estimators we can then compute:

$$\widehat{\gamma} = \frac{1}{\widehat{\theta}_2} = \frac{1}{.952} = 1.05042 \approx 1 \text{ (which implies Exponential cdf) and } \widehat{\alpha} = e^{\widehat{\theta}_1} = e^{2.414} = 11.178586$$

From these values, we can then estimate μ and σ:

$$\widehat{\mu} = \widehat{\alpha}\, \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) = (11.178586)\, \Gamma\left(1 + \frac{1}{1.05042}\right) = 10.96$$

$$\widehat{\sigma} = \sqrt{\widehat{\alpha}^2 \left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]}$$

$$\widehat{\sigma} = \sqrt{(11.178586)^2 \left[\Gamma\left(1 + \frac{2}{1.05042}\right) - \Gamma^2\left(1 + \frac{1}{1.05042}\right)\right]} = 10.44$$

From the data we have the distribution-free estimators:

$$\widehat{\mu} = \bar{Y} = 11.57 \qquad\qquad \widehat{\sigma} = S = 12.29$$

Both of these methods are very crude estimators of the population mean and standard deviation. When we know the underlying population distributions, there are much more accurate estimators.

Method of Moments (MOM) Estimators:

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a population or n iid realizations from a process having cdf which depends on k unknown parameters: $\theta_1, \theta_2, \ldots, \theta_k$. The population moments $m_i = E(Y^i)$ depend on the k θs:

$$m_i = E(Y^i) = \int_{-\infty}^{\infty} y^i dF(y) = g_i(\theta_1, \theta_2, \ldots, \theta_k) \quad \text{for} \quad i = 1, 2, 3, 4$$

We obtain sample estimators of these $m_i s$ by replacing F with the edf $\widehat{F}$:

$$\widehat{m}_i = \int_{-\infty}^{\infty} y^i d\widehat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} Y_j^i$$

To obtain estimators of the θs, we just equate the sample moments to the population moments (with θ_is replaced with $\widehat{\theta}_i$s and solve for the $\widehat{\theta}_i$s:

$$\widehat{m}_i = g_i(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k) \quad \text{for} \quad i = 1, 2, \ldots, k$$

We have k equations in k unknowns.

EXAMPLE #1 Suppose F is a $N(\theta_1, \theta_2^2)$ cdf and we have a random sample:

$Y_1, Y_2, \ldots, Y_n$ iid $N(\theta_1, \theta_2^2)$. Find $\widehat{\theta}_1$ and $\widehat{\theta}_2$

$$m_1 = E[Y] = \theta_1 = g_1(\theta_1, \theta_2) \quad m_2 = E[Y^2] = Var(Y) + (E[Y])^2 = \theta_2^2 + \theta_1^2 = g_2(\theta_1, \theta_2)$$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n} \sum_{1}^{n} Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\theta}_1, \widehat{\theta}_2) = \widehat{\theta}_1$$

$$(2) \quad \frac{1}{n} \sum_{1}^{n} Y_i^2 = \widehat{m}_2 = g_2(\widehat{\theta}_1, \widehat{\theta}_2) = \widehat{\theta}_2^2 + \widehat{\theta}_1^2$$

Solving equations (1) and (2) we obtain

$$(1) \Rightarrow \widehat{\theta}_1 = \bar{Y} \qquad (2) \Rightarrow \widehat{\theta}_2^2 = \widehat{m}_2 - \widehat{\theta}_1^2 = \frac{1}{n} \sum_{1}^{n} Y_i^2 - \bar{Y}^2 = \frac{1}{n} \sum_{1}^{n} (Y_i - \bar{Y})^2$$

Therefore, we obtain:

$$\widehat{\theta}_1 = \bar{Y} \qquad \widehat{\theta}_2 = \sqrt{\frac{1}{n} \sum_{1}^{n} (Y_i - \bar{Y})^2}$$

EXAMPLE #2 Suppose F is a $Gamma(\alpha, \beta)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Gamma(\alpha, \beta)$. Find $\widehat{\alpha}$ and $\widehat{\beta}$.

$m_1 = E[Y] = \alpha\beta = g_1(\alpha, \beta) \quad m_2 = E[Y^2] = Var(Y) + (E[Y])^2 = \alpha\beta^2 + (\alpha\beta)^2 = g_2(\alpha, \beta)$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n}\sum_1^n Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\alpha}, \widehat{\beta}) = \widehat{\alpha}\widehat{\beta}$$

$$(2) \quad \frac{1}{n}\sum_1^n Y_i^2 = \widehat{m}_2 = g_2(\widehat{\alpha}, \widehat{\beta}) = \widehat{\alpha}\widehat{\beta}^2 + \widehat{\alpha}^2\widehat{\beta}^2$$

Solving equations (1) and (2) we obtain

$$(1) \Rightarrow \quad \widehat{\alpha} = \bar{Y}/\widehat{\beta} \qquad\qquad (2) \Rightarrow \quad \widehat{m}_2 = \bar{Y}\widehat{\beta} + \bar{Y}^2 \quad \Rightarrow \quad \widehat{\beta} = (\widehat{m}_2 - \bar{Y}^2)/\bar{Y}$$

Therefore, we obtain:

$$\widehat{\alpha} = \bar{Y}^2/(\widehat{m}_2 - \bar{Y}^2) \qquad\qquad \widehat{\beta} = (\widehat{m}_2 - \bar{Y}^2)/\bar{Y}$$

EXAMPLE #3 Suppose F is a $Weibull(\gamma, \alpha)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Weibull(\gamma, \alpha)$. Find $\widehat{\gamma}$ and $\widehat{\alpha}$.

$$m_1 = \alpha\ \Gamma\left(1 + \frac{1}{\gamma}\right) = g_1(\gamma, \alpha), \quad \text{where} \ \ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

$$m_2 = \alpha^2\ \Gamma\left(1 + \frac{2}{\gamma}\right) = g_2(\gamma, \alpha)$$

Next, we equate sample moments to population moments:

$$(1) \quad \frac{1}{n}\sum_1^n Y_i = \bar{Y} = \widehat{m}_1 = g_1(\widehat{\gamma}, \widehat{\alpha}) = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right)$$

$$(2) \quad \frac{1}{n}\sum_1^n Y_i^2 = \widehat{m}_2 = g_2(\widehat{\gamma}, \widehat{\alpha}) = \widehat{\alpha}^2\ \Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right)$$

The two equations are then solved numerically, closed form solutions are not possible.

END: Friday 9/24/21

START: Monday 9/27/21

Maximum Likelihood Estimation (MLE)

MOM's only used the first few moments of the distribution and hence do not use the full knowledge of the structure of the population distribution. The MLE's will directly use the pdf in obtaining the estimates of the unknown parameters.

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, where θs are elements of a parameter space Θ.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, and define the likelihood function as

$$L(\theta_1, \theta_2, \cdots, \theta_k; \mathbf{y}) = f(y_1, y_2, \ldots, y_n; \theta_1, \theta_2, \cdots, \theta_k) \quad \text{joint pdf of} \quad Y_1, Y_2, \ldots, Y_n$$

If Y_i has a discrete pdf, then

$$L(\theta_1, \theta_2, \cdots, \theta_k; \mathbf{y}) = P[Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n]$$

Because the $Y_i s$ are iid we have

$$L(\theta_1, \theta_2, \cdots, \theta_k; \mathbf{y}) = f(y_1; \boldsymbol{\theta}) f(y_2; \boldsymbol{\theta}) \cdots f(y_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})$$

The MLEs of the θ_is is that vector $(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k)$ which maximizes the likelihood function:

$$L(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k; \mathbf{y}) = \max_{\theta \epsilon \Theta} L(\theta_1, \theta_2, \cdots, \theta_k; \mathbf{y})$$

Example #1 Exponential pdf:

Suppose F is an Exponential (β) cdf and $T_1, T_2, \ldots, T_n$ is a random sample from F. Find the MLE of β : $\widehat{\beta}$.

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(t_i; \beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-t_i/\beta} \\ &= \left(\frac{1}{\beta}\right)^n e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i} \end{aligned}$$

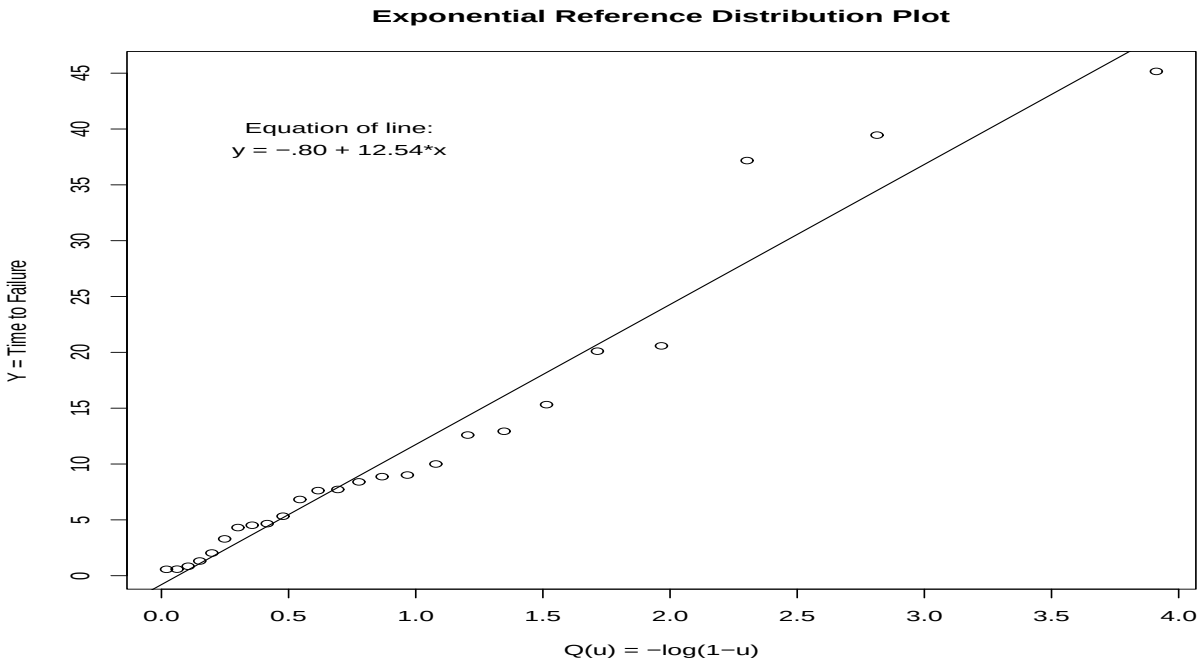
We will demonstrate the derivation of MLE's for the time to failure example.

The time to failure, in 100 hours, of a random sample of 25 newly designed fuel pumps are recorded as follows:

15.321	9.008	20.104	7.729	45.154	8.404	5.332	0.577	4.305
4.517	12.594	6.829	3.291	37.175	0.841	1.317	7.613	20.582
2.030	10.001	4.666	12.933	0.591	39.454	8.875		

From the data, compute $\sum_{i=1}^{25} T_i = 289.243$ and $\bar{T} = 11.570$

From the Exponential Distribution Plot given below, we conclude that the times to failure are adequately modeled by an Exponential Distribution.



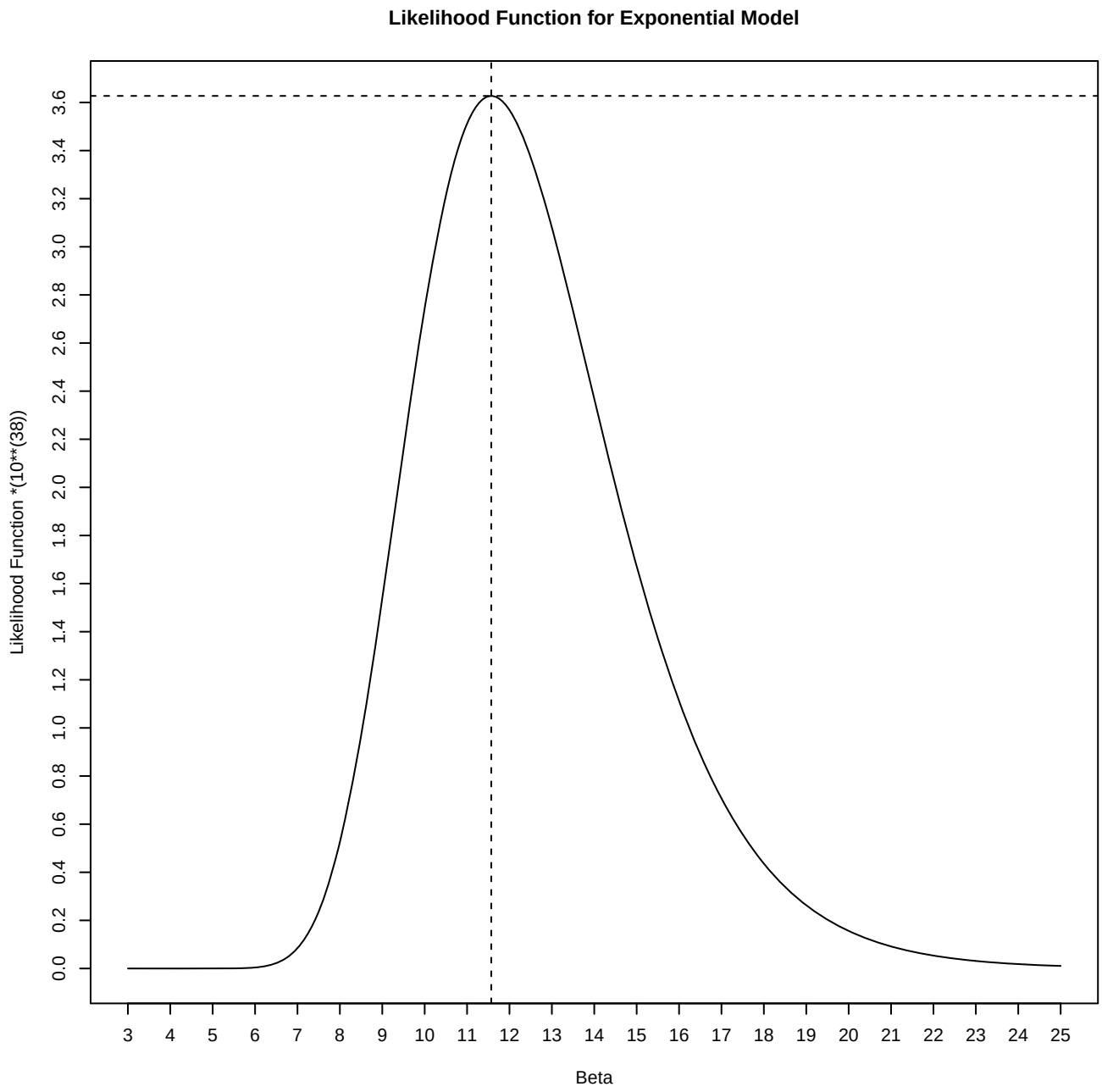
We next need to estimate the parameter β in the exponential distribution:

$$f(t) = \frac{1}{\beta} e^{-t/\beta}$$

We will use Maximum Likelihood techniques, that is, find the value of β which maximizes the likelihood function:

$$L(\beta) = \prod_{i=1}^{n} f(T_i) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-T_i/\beta} = \beta^{-n} e^{-\sum_{i=1}^{n} T_i/\beta} = \beta^{-25} e^{-289.243/\beta}.$$

Next, we plot the likelihood function, $L(\beta)$ and determine the value of β which maximizes this function.



From the plot of the likelihood function, we can observe that the maximum occurs at

$$\beta \approx 11.57$$

.

R code used to produce Exponential Reference plot and plot of the likelihood function

```
# program name: mle_exp.R

t = c(15.321,9.008,20.104,7.729,45.154,8.404,5.332,0.577,4.305,4.517,12.594,
     6.829,3.291,37.175,0.841,1.317,7.613,20.582,2.030,10.001,4.666,12.933,
     0.591,39.454,8.875)
t = sort(t)
n = length(t)
i = seq(1,n,1)
u = (i-.5)/n
x =  -log(1-u)


plot(x,t, xlab="-Log(1-u)",ylab="Time to Failure",lab=c(13,11,7),
main="Exponential Reference Distribution Plot")
abline(lm(t~x))
text(.7,40,"Equation of line:")
text(.7,38,"y = -.80 + 12.54*x")

b = seq(3,30,.01)
LK = (b^(-25))*exp(-289.243/b)*(10)^38
out =  cbind(b,LK)
LKmax  =  max(LK)
bmax  =  which(LK==max(LK))
bmax
MLE  =  b[bmax]
MLE
par(cex=.65)
plot(b,LK, type="l",lab=c(30,16,7))
par(cex=.99)
title("Likelihood Function for Exponential Model",xlab="Beta",
ylab="Likelihood Function *(10**(38))")
abline("v"=b[bmax],lty=2)
abline("h"=max(LK),lty=2)


library(MASS)
mle_exp=fitdistr(t,"exponential")
mle_weibull=fitdistr(t,"weibull",lower=c(0,0))
```

easier way to obtain MLE

The statement "lower=c(0,0)" avoids having an error statement appear in the R output.

Using the last 3 lines of code in the R program, given on the previous page:

```
library(MASS)
mle_exp=fitdistr(t,"exponential")
```

we can obtain the MLE estimates of the parameters in an Exponential and Weibull model:
Output from R code:

```
> fitdistr(t,"exponential")
      rate
  0.08643252
 (0.01728650)
```

Note that the estimate in the exponential model is given in terms of $1/\beta$, called the "rate" in R: $\hat{\beta}$ =1/rate = 1/.08643252 = 11.5697 which is the value we determined.

In fact, we can show that the MLE of β in the exponential distribution occurs at $\widehat{\beta} = \bar{T}$:

Suppose F is a Exponential (β) cdf and we have a random sample: $T_1, T_2, \ldots, T_n$ from F, i.e., iid Exponential (β). Find the MLE's : $\widehat{\beta}$.

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f(t_i; \beta) \\ &= \prod_{i=1}^{n} \beta^{-1} e^{-t_i/\beta} \\ &= \beta^{-n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\beta; y) = -nlog(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} t_i$$

Take the partial derivate of the log-likelihood wrt β, set derivate equal to 0, and then solve for $\widehat{\beta}$:

$$\frac{\partial l(\beta; y)}{\partial \beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} t_i \quad \Rightarrow \quad \widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Note that the second partial of the log-likelihood is negative at $\widehat{\beta}$:

$$\frac{\partial^2 l(\beta; y)}{\partial^2 \beta} \text{ at } (\beta = \widehat{\beta}) = \frac{n}{\widehat{\beta}^2} - \frac{2}{\widehat{\beta}^3}\sum_{i=1}^{n} t_i = -\frac{n}{\widehat{\beta}^2} < 0$$

Thus, the likelihood function is a maximum at $\widehat{\beta}$.

In our data set, we had $\bar{T} = 11.5692$ which to roundoff error is the value we obtained for the MLE of β using 1/rate = = 1/.08643252 = 11.5697

EXAMPLE #2 Suppose F is a $N(\theta_1, \theta_2^2)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $N(\theta_1, \theta_2^2)$. Find the MLE's : $\widehat{\theta_1}$ and $\widehat{\theta_2}$

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(y_i; \theta_1, \theta_2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}(y_i - \theta_1)^2} \\ &= \frac{1}{(2\pi)^{n/2}(\theta_2)^n} e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i - \theta_1)^2} \end{aligned}$$

Taking the natural log of both sides of the equation yields with $l(\theta_1, \theta_2; y) = log(L(\theta_1, \theta_2); y)$:

$$l(\theta_1, \theta_2; y) = -\frac{n}{2} log(2\pi) - n log(\theta_2) - \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (y_i - \theta_1)^2$$

Take the partial derivates wrt θ_1 and θ_2, set derivates equal to 0, and then solve for $\widehat{\theta_1}$ and $\widehat{\theta_2}$.

$$\frac{\partial l(\theta_1, \theta_2; y)}{\partial \theta_1} = \frac{1}{\theta_2^2} \sum_{i=1}^{n} (y_i - \theta_1) \quad \Rightarrow \widehat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{\partial l(\theta_1, \theta_2; y)}{\partial \theta_2} = \frac{-n}{\theta_2} + \frac{1}{\theta_2^3} \sum_{i=1}^{n} (y_i - \theta_1)^2 \quad \Rightarrow \widehat{\theta}_2 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\theta}_1)^2}$$

The same answers that we obtained using MOM.

To be complete, we would need to verify that these solutions in fact yield the maximums in the likelihood function and not the minimums. Examine the second derivatives and the mixed derivatives of the likelihood function evaluated at $(\hat{\theta}_1,\ \hat{\theta}_2)$

$$\text{Let A } = \frac{\partial^2 l(\theta_1, \theta_2; y)}{\partial \theta_1^2}, \qquad \text{B } = \frac{\partial^2 l(\theta_1, \theta_2; y)}{\partial \theta_1 \partial \theta_2}, \qquad \text{C } = \frac{\partial^2 l(\theta_1, \theta_2; y)}{\partial \theta_2^2}$$

There are four cases:

1. If $B^2 - AC < 0$ and $A + C < 0$ then $(\hat{\theta}_1,\ \hat{\theta}_2)$ is a relative maximum
2. If $B^2 - AC < 0$ and $A + C > 0$ then $(\hat{\theta}_1,\ \hat{\theta}_2)$ is a relative minimum
3. If $B^2 - AC > 0$ then $(\hat{\theta}_1,\ \hat{\theta}_2)$ is a saddle point
4. If $B^2 - AC = 0$ then $(\hat{\theta}_1,\ \hat{\theta}_2)$ is indeterminate

EXAMPLE #3 Suppose F is a $Gamma(\alpha, \beta)$ cdf and we have a random sample: $Y_1, Y_2, \ldots, Y_n$ iid $Gamma(\alpha, \beta)$. Find the MLE's : $\widehat{\alpha}$ and $\widehat{\beta}$.

$$\begin{aligned} L(\alpha, \beta; y) &= \prod_{i=1}^{n} f(y_i; \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta} \\ &= \frac{1}{\Gamma^n(\alpha)\beta^{n\alpha}} (\prod_{i=1}^{n} y_i)^{\alpha-1} e^{-\frac{1}{\beta}\sum_{i=1}^{n} y_i} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\alpha, \beta; y) = -nlog(\Gamma(\alpha)) - n\alpha log(\beta) + (\alpha - 1)\sum_{i=1}^{n} log(y_i) - \frac{1}{\beta}\sum_{i=1}^{n} y_i$$

Take the partial derivates wrt α and β, set derivates equal to 0, and then solve for $\widehat{\alpha}$ and $\widehat{\beta}$.

$$\frac{\partial l(\alpha, \beta; y)}{\partial \alpha} = -\frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} - nlog(\beta) + \sum_{i=1}^{n} log(y_i) \quad \Rightarrow \quad -\frac{n\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} - nlog(\widehat{\beta}) + \sum_{i=1}^{n} log(y_i) = 0$$

$$\frac{\partial l(\alpha, \beta; y)}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} y_i \quad \Rightarrow \quad -\frac{n\widehat{\alpha}}{\widehat{\beta}} + \frac{1}{\widehat{\beta}^2}\sum_{i=1}^{n} y_i = 0$$

Need to verify that the solutions to these two equations $\widehat{\alpha}$ and $\widehat{\beta}$ are maximums of the log-likelihood and not minimums.

For the gamma family of pdfs, there is no closed form solution to the equations because of the terms involving :

$$\Gamma(c) = \int_0^{\infty} y^{c-1} e^{-y} dy \quad \text{and its derivative} \quad \Gamma'(c)$$

.

Thus, we would need to obtain a numerical solution to the equations.

Using the following R code we obtain the MLE's:

```
y =  c(y1,y2,...,yn) - DATA

library(MASS)
fitdistr(y,"gamma")
```

The estimators will be outputted as "Shape" and "Rate"

Then we obtain

$$\widehat{\beta} = \frac{1}{\text{Rate}} \quad \text{and} \quad \widehat{\alpha} = \text{Shape}$$

EXAMPLE #4 Suppose F is a $Weibull(\gamma, \alpha)$ cdf and we have a random sample:
$T_1, T_2, \ldots, T_n$ iid $Weibull(\gamma, \alpha)$.
Find the MLE's : $\widehat{\gamma}$ and $\widehat{\alpha}$.

$$\begin{aligned} L(\gamma, \alpha; \mathbf{t}) &= \prod_{i=1}^{n} f(t_i; \gamma, \alpha) \\ &= \prod_{i=1}^{n} \frac{\gamma}{\alpha} \left(\frac{t_i}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\gamma}} \\ &= \left(\frac{\gamma}{\alpha}\right)^{n} \left(\prod_{i=1}^{n} \left(\frac{t_i}{\alpha}\right)^{\gamma-1}\right) e^{-\frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma}} \end{aligned}$$

Taking logarithms of both sides of the equation yields:

$$l(\gamma, \alpha; \mathbf{t}) = log(L(\gamma, \alpha; \mathbf{t})) = nlog(\gamma) - nlog(\alpha) - n(\gamma - 1)log(\alpha) + (\gamma - 1)\sum_{i=1}^{n} log(t_i) - \frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma}$$

Take the partial derivates wrt γ and α, set derivates equal to 0, and then solve for $\widehat{\gamma}$ and $\widehat{\alpha}$:

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \alpha} = -\frac{n\gamma}{\alpha} + \frac{\gamma}{\alpha^{\gamma+1}} \sum_{i=1}^{n} t_i^{\gamma}$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \alpha} = 0 \quad \Rightarrow \quad \widehat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i^{\widehat{\gamma}}\right)^{1/\widehat{\gamma}}$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \gamma} = \frac{n}{\gamma} - nlog(\alpha) + \sum_{i=1}^{n} log(t_i) + \frac{log(\alpha)}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma} - \frac{1}{\alpha^{\gamma}} \sum_{i=1}^{n} t_i^{\gamma} log(t_i)$$

$$\frac{\partial l(\gamma, \alpha; y)}{\partial \gamma} = 0 \quad \Rightarrow \quad \frac{\sum_{i=1}^{n} t_i^{\widehat{\gamma}} log(t_i)}{\sum_{i=1}^{n} t_i^{\widehat{\gamma}}} - \frac{1}{\widehat{\gamma}} = \frac{1}{n} \sum_{i=1}^{n} log(t_i)$$

Numerical techniques would be required to find the solutions to these two equations. R and SAS have a routines for estimating the parameters in the Weibull distribution.

We will illustrate the code using the data from the Weibull example used earlier in this handout to illustrate the graphical estimation of parameters.

Recall that the Weibull Reference distribution plot on page 5 indicated that a Weibull distribution would be an appropriate model for the data. From the reference distribution plot we obtained the following graphical estimates of the scale and shape parameters:

$$\widehat{\gamma} = 1.05 \qquad \text{and} \qquad \widehat{\alpha} = 11.18$$

Using the following R code we obtain the MLE's:

```
y = c(15.321, 9.008, 20.104, 7.729, 45.154, 8.404, 5.332, 0.577, 4.305, 4.517,
12.594, 6.829, 3.291, 37.175, 0.841, 1.317, 7.613, 20.582, 2.030, 10.001,
4.666, 12.933, 0.591, 39.454,  8.875)

library(MASS)
fitdistr(y,"weibull")


OUTPUT from R:

     shape         scale
  0.9839245    11.4852981
( 0.1512936) ( 2.4660607)
```

The values in parentheses are the standard errors of the estimators.

Recall, that we also used the exponential model for this data set and obtained $\hat{\beta} = 11.57$.

Furthermore, $E(Y) = \beta = \sqrt{Var(Y)}$.

Thus, we have $\hat{\mu} = 11.57 = \hat{\sigma}$.

How would these values change if modeled the data with the Weibull distribution?

$$\widehat{\mu} = \widehat{\alpha}\ \Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) = (11.4852981)\Gamma\left(1 + \frac{1}{.9839245}\right) = 11.5659$$

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\widehat{\alpha}^2\left[\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)\right]} \\ &= \sqrt{(11.4852981)^2\left[\Gamma\left(1 + \frac{2}{0.9839245}\right) - \Gamma^2\left(1 + \frac{1}{0.9839245}\right)\right]} = 11.75532\end{aligned}$$

Note: $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ which can be obtain using the R function **gamma(x)**

From the data we compute $\bar{Y} = 11.57 \qquad S = 12.284.$

Thus, the two estimates of the mean are closely matched but there is some difference in the two estimates of the standard deviation.

The next page contains the SAS code needed to obtain the MLS's for the parameters in the Weibull model.

```
SAS code for estimating parameters from a Weibull Distribution:
weibmle_fuelpumps.SAS

option ls=75 ps=55 nocenter nodate;
title 'Weibull MLE Estimation of Fuel Pump Data Data';
data cords;
input F @@;
label F  = 'Time to Failure of Pumps';
cards;
15.321   9.008    20.104    7.729    45.154    8.404    5.332    0.577     4.305
 4.517  12.594     6.829    3.291    37.175    0.841    1.317    7.613    20.582
 2.030  10.001     4.666   12.933     0.591   39.454    8.875
run;
proc print;
proc lifereg data=cords;
 model  F = /dist=weibull covb;
run;

OUTPUT

Weibull MLE Estimation of Fuel Pump Data Data                            11

The LIFEREG Procedure

                         Model Information

Data Set                      WORK.CORDS
Dependent Variable                Log(F)    Time to Failure of Pumps
Number of Observations                25
Noncensored Values                    25
Right Censored Values                  0
Left Censored Values                   0
Interval Censored Values               0
Name of Distribution             Weibull
Log Likelihood              -39.33978867


Number of Observations Read          25
Number of Observations Used          25

                   Analysis of Parameter Estimates

                           Standard   95% Confidence     Chi-
Parameter     DF Estimate    Error       Limits       Square Pr > ChiSq

Intercept      1   2.4411   0.2147   2.0202   2.8619  129.25     <.0001
Scale          1   1.0163   0.1563   0.7519   1.3738
Weibull Scale  1  11.4853   2.4661   7.5401  17.4947
Weibull Shape  1   0.9839   0.1513   0.7279   1.3300


     Estimated Covariance Matrix

              Intercept         Scale

Intercept      0.046102     -0.010810
Scale         -0.010810      0.024423
```

Note the estimates, Weibull Scale, $\hat{\alpha}$ and Weibull Shape, $\hat{\gamma}$ are identical to the values obtained from R on the previous page.

Caved in STAT 638

Bayesian Inference

Alternatives to MLE procedures for estimating parameters in a distribution are Bayesian procedures. In Bayesian procedures, the unknown parameters are assumed to be random parameters and not fixed but unknown population parameters as in MLE procedures.

Suppose θ is the unknown parameter in a pdf f for which information is sought. A random sample (iid) observations (data) are obtained from f and then the likelihood function is formulated, as was done in the MLE procedure.

This is where the Bayesian approach and MLE diverge. In the Bayesian procedure, the parameter θ is a random variable with a *prior* distribution g which is specified before the data is collected. The prior g is a summary of the researcher's knowledge about the random nature of θ. This prior distribution is then updated using the collected data to obtain a *posterior* distribution, h:

$$h(\theta|data) = \frac{Likelihood\ x\ Prior}{\sum Likelihood\ x\ Prior} = \frac{f(data|\theta)g(\theta)}{\sum_{\text{all }\theta} f(data|\theta)g(\theta)}$$

The mode of the distribution is then the "estimate" of θ and probability intervals, the Bayesian version of confidence intervals, can be calculated using the posterior distribution. Bayesian analysis is somewhat problematic because the validity of the results depend on the validity of the selection of the prior distribution. This validation can be difficult to assess statistically. Although, often a flat prior can be used when the researcher has very little knowledge about θ.

Bayesian Analysis is widely used in many applied fields and many statisticians conduct research to develop new methodology for Bayesian Analyses.

The course STAT 638, **Introduction to Bayesian Methods** is offered during the fall semesters both locally and online.

DISTRIBUTION-FREE SUMMARIES

Let $Y_1, Y_2, \ldots, Y_n$ a random sample (or iid observations) from a population in which the pdf $f(y)$ is not specified. Suppose estimators of population summaries are desired. There are a number of possible methods to obtain the estimators. In general, we can simply replace the the population distribution function (cdf) with the sample distribution function (edf) in the definition of the summary parameter.

Estimators Based on Population Moments

Suppose we have a random sample (iid r.v.s) $Y_1, Y_2, \ldots, Y_n$ with cdf F which is unknown. We want to estimate the various population summaries of location:

In the definition of $\mu_i = \int_{-\infty}^{\infty}(y-\mu)^i dF(y)$ replace $F(y)$ with the edf

$$\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y)$$

We then obtain

$$\widehat{\mu} = \int_{-\infty}^{\infty} y d\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} Y_{(i)} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

and

$$\widehat{\mu}_k = \int_{-\infty}^{\infty} (y-\widehat{\mu})^k d\widehat{F}(y) = \frac{1}{n}\sum_{i=1}^{n}(Y_{(i)}-\widehat{\mu})^k = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^k$$

Thus, we have the following estimators of the population standard deviation σ, population skewness β_1, population kurtosis β_2, and trimmed mean $\mu_{(\alpha)}$:

1. **Sample Standard Deviation**

$$\widehat{\sigma} = \sqrt{\widehat{\mu}_2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

2. **Sample Skewness**

$$\widehat{\beta}_1 = \frac{\widehat{\mu}_3}{(\widehat{\mu}_2)^{3/2}} = \frac{\widehat{\mu}_3}{(\widehat{\sigma})^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^3}{(\widehat{\sigma})^3}$$

3. **Sample Kurtosis**

$$\widehat{\beta}_2 = \frac{\widehat{\mu}_4}{(\widehat{\mu}_2)^2} = \frac{\widehat{\mu}_4}{(\widehat{\sigma})^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^4}{(\widehat{\sigma})^4}$$

Some software packages, e.g., SAS, report the excess kurtosis defined as $\hat{\beta}_2 - 3$. The word excess referring to the difference from the kurtosis for the normal distribution.

4. **Sample α-Trimmed Mean**

Recall,

$$\mu_{(\alpha)} = \frac{1}{1-2\alpha}\int_{Q(\alpha)}^{Q(1-\alpha)} y\ dF(y)$$

thus using our idea of replacing the quantile function with its sample estimator we have

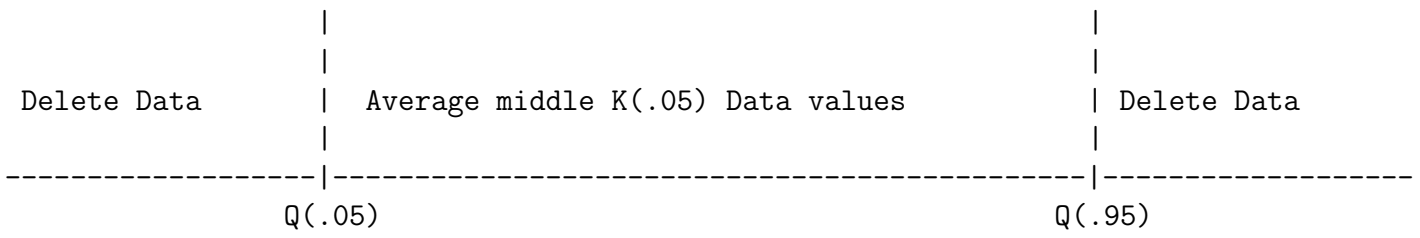
$$\widehat{\mu}_{(\alpha)} = \frac{1}{1-2\alpha}\int_{\widehat{Q}(\alpha)}^{\widehat{Q}(1-\alpha)} y\ d\widehat{F}(y) = \frac{1}{K(\alpha)/n}\sum_{i=[n\alpha+1]}^{n-[n\alpha]} Y_{(i)}\frac{1}{n} = \frac{1}{K(\alpha)}\sum_{i=[n\alpha+1]}^{n-[n\alpha]} Y_{(i)},$$

where $[A]$ is the largest integer less than or equal to the real number A and

$$K(\alpha) = n - [n\alpha] - [n\alpha+1] + 1$$

is the number of data values between $\widehat{Q}(\alpha)$ and $\widehat{Q}(1-\alpha)$.

Thus, $\widehat{\mu}_{(\alpha)}$ is the average of the data values that remain after removing the $[n\alpha]$ smallest and $[n\alpha]$ largest values in the data set. We are trimming exactly $100\alpha\%$ of the data from both tails when $n\alpha$ is an integer and slightly less than $100\alpha\%$ when $n\alpha$ is not an integer.



For example, suppose $n = 30$, $\alpha = .05$ then $n\alpha = (30)(.05) = 1.5$

$$K(\alpha) = n - [n\alpha] - [n\alpha+1] + 1 = 30 - [1.5] - [2.5] + 1 = 28$$

Thus, we would trim $(1 - \frac{28}{30})/2 = .033$ from the left and right tails, not $\alpha = .05$.

If $n\alpha$ is an integer, then exactly $100\alpha\%$ is trimmed from both tails.

For example, suppose $n = 20$, $\alpha = .10$ then $n\alpha = (20)(.1) = 2$

$$K(\alpha) = n - [n\alpha] - [n\alpha+1] + 1 = 20 - [2] - [3] + 1 = 16$$

Thus, we would trim $(1 - \frac{16}{20})/2 = .10$ from the left and right tails and this would be exactly $\alpha = .1$.

In R, use **mean(y,trim=.1)**

Modified Estimators

In practice, the previously defined estimators are often modified so that the estimators are **unbiased** estimators of the specified parameter when the data are from a normal distribution. In particular, many computer packages use the following quantities.

1. **Unbiased Estimator of μ:**

$$\text{Let} \quad \hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{then} \qquad E[\bar{Y}] = \mu$$

- That is, $\bar{Y}$ is an unbiased estimator of $\hat{\mu}$ for any distribution for which $|\mu| < \infty$

2. **Unbiased Estimator of σ^2:**

$$\text{Let} \quad \widehat{\sigma}^{2*} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \text{then} \qquad E[S^2] = \sigma^2$$

- That is, S^2 is an unbiased estimator of σ^2 for any distribution for which $\sigma^2 < \infty$
- However, S is a biased estimator of σ, that is, $E[S] \neq \sigma$.

  In Handout 10, we will prove the results in 1. and 2.

3. **Modified Estimator of Skewness Parameter β_1:**

$$\text{Let} \quad \widehat{\beta}_1^* = \left(\frac{n^2}{(n-1)(n-2)}\right) \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^3}{(\widehat{\sigma}^*)^3} = \left(\frac{n^2}{(n-1)(n-2)}\right) \frac{\hat{\mu}_3}{(S^2)^{3/2}}$$

- If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then $\beta_1 = 0$ and $E[\widehat{\beta}_1^*] = 0$, therefore, $\widehat{\beta}_1^*$ is an unbiased estimator of β_1. This result does not necessarily hold when the data is from a non-normal distribution.

4. **Modified Estimator of Excess Kurtosis Parameter $\beta_2 - 3$:**

$$\text{Let} \quad \widehat{\beta}_2^* - 3 = \frac{(n+1)n^2}{(n-1)(n-2)(n-3)} \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^4}{(\widehat{\sigma}^*)^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

- If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then $\beta_2 = 3$ and $E[\widehat{\beta}_2^*] = 3$ so $\widehat{\beta}_2^*$ is an unbiased estimator of β_2. This result does not hold necessarily hold when the data is from a non-normal distribution.

These are the forms of the estimators used in most software packages. However, not in all. You must check the definitions if you want to know what exactly the software program is computing.

Estimators Based on Quantiles

To obtain estimators of population parameters which are defined in terms of the population quantile $Q(u)$, we will just replace the population quantile with the sample quantile in the definition of the population parameter:

1. **Median** The population median $\tilde{\mu} = Q(.5)$ is estimated by

$$\widehat{Q}(.5) = \begin{cases} Y_{((n+1)/2)} & if \quad n \text{ is odd} \\ \left(Y_{(n/2)} + Y_{(n/2+1)}\right)/2 & if \quad n \text{ is even} \end{cases}$$

   In R, use **quantile(y,.5)**

2. **Quartiles** The lower and upper quartiles $Q_1 = Q(.25)$ and $Q_3 = Q(.75)$ are estimated by their corresponding sample quartiles in most instances: $\widehat{Q}_1 = \widehat{Q}(.25)$ and $\widehat{Q}_3 = \widehat{Q}(.75)$.

   In R, use **u=c(.25,.5,.75), quantile(y,u)** to obtain $\widehat{Q}(.25), \widehat{Q}(.5), \widehat{Q}(.75)$

   However, in some software packages and textbooks an alternative definition is given:

   Divide the data set into two equal halves:

   If n is even:

   Set 1: $Y_{(1)}, \cdots, Y_{(n/2)}$ and

   Set 2: $Y_{(n/2+1)}, \cdots, Y_{(n)}$

   Then $\widehat{Q}_1$ is the median of Set 1 and $\widehat{Q}_3$ is the median of Set 2

   If n is odd:

   Set 1: $Y_{(1)}, \cdots, Y_{((n+1)/2)}$ and

   Set 2: $Y_{((n+1)/2)}, \cdots, Y_{(n)}$

   Then $\widehat{Q}_1$ is the median of Set 1 and $\widehat{Q}_3$ is the median of Set 2

   For small n, these values for $\widehat{Q}_1$ and $\widehat{Q}_3$ may differ from the values obtained from $\widehat{Q}(.25)$ and $\widehat{Q}(.75)$

3. **Five Number Summary of Data:** Produced in R using the function **quantile(y)**

$$\left[Min = Y_{(1)}; \qquad Q_1; \qquad Q_2; \qquad Q_3; \qquad Max = Y_{(n)}\right]$$

   In R, the function **summary(y)** provides the 5 number summary of data in y but it also provides the sample mean $\bar{y}$.

4. **Interquartile Range (IQR)** The sample estimator of the population

   $IQR = Q(.75) - Q(.25)$ is just $\widehat{IQR} = \hat{Q}(.75) - \hat{Q}(.25)$

5. **Range**: The population range is $Q(1) - Q(0) = R$. The sample estimator is taken to be

   $\widehat{R} = Y_{(n)} - Y_{(1)}$

6. **MAD**: $MAD = Median\{|Y - Median\{Y\}|\}/.6745 = Median\{|Y - Q_Y(.5)|\}/.6745$

   Therefore, $\widehat{MAD} = Median\{|Y - \widehat{Q}_Y(.5)|\}/.6745$

   To compute $\widehat{MAD}$

   (a) Find $\widehat{Q}_Y(.5)$ from $Y_1, Y_2, \ldots, Y_n$

   (b) Compute $W_i = |Y_i - \widehat{Q}_Y(.5)|$ for $i = 1, \ldots, n$

   (c) Find $\widehat{Q}_W(.5)$ from $W_1, W_2, \ldots, W_n$

   (d) $\widehat{MAD} = \widehat{Q}_W(.5)/.6745$

7. **m-estimator of Location** A modification of the $\alpha-$trimmed mean is the m-estimator. In place of deleting observations the m-estimator $\hat{m}$ assigns weights w_i to each data value such that the more extreme a data value is from the "center" of the data, the smaller the weight. Thus, in place of deleting extreme data values as we did with the $\alpha-$trimmed mean, we just reduce their influence.

   The m-estimator is defined as

$$\widehat{m} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i Y_i$$

   The m-estimator involves three parameters:

   (a) v a robust measure of dispersion, $\widehat{MAD}$, for example.

   (b) t a "tuning constant"

   The tuning constant is generally taken to be a value in (1.345, 1.5). Smaller values of t place a larger penalty on data values for being extreme to the center of the data.

   (c) w_i a weight assigned to each data value Y_i with

$$w_j = \begin{cases} -\frac{tv}{Y_j - \hat{m}} & if \quad Y_j < \hat{m} - tv \\ 1 & if \quad \hat{m} - tv \le Y_j \le \hat{m} + tv \\ \frac{tv}{Y_j - \hat{m}} & if \quad Y_j > \hat{m} + tv \end{cases}$$

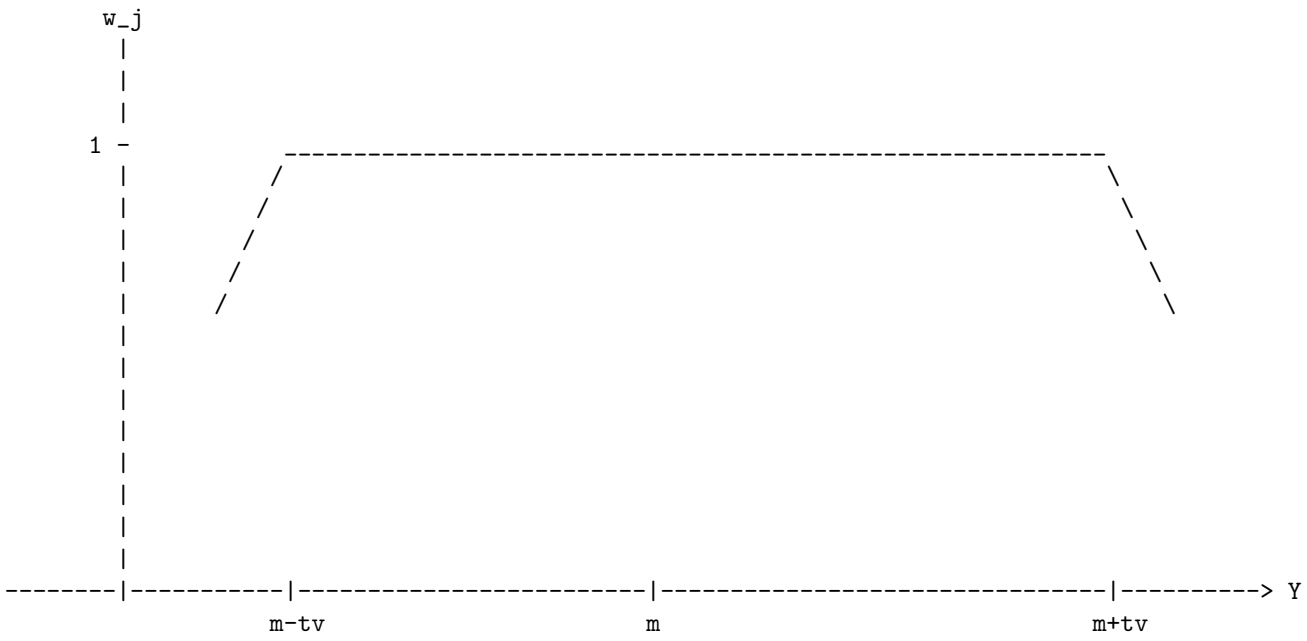
Note that $0 \leq w_j \leq 1$

For $\widehat{m} - tv \leq Y_j \leq \widehat{m} + tv$ we have $w_j = 1$, thus there is no downweighting of the data value Y_j:

For $Y_j < \widehat{m} - tv$ or $Y_j > \widehat{m} + tv$, then $w_j < 1$ and the data value Y_j receives a greater downweighting as it moves further from the center, i.e., w_j decreases as $|Y_j - \widehat{m}|$ increases

The computation of $\hat{m}$ is iterative:

1. Select an initial value for $\widehat{m}$, for example, $\widehat{m}_1 = \widehat{Q}(.5)$.
2. Select value for t (e.g. $t = 1.345$)
3. Select number of iterations OR
4. Select a level of relative accuracy RA $= |\frac{\widehat{m}_{k+1} - \widehat{m}_k}{\widehat{m}_k}| < \epsilon$ (a stopping point)
5. Select a robust estimator of dispersion ($v = \widehat{MAD}$)
6. Calculate $w_j s$ using $\widehat{m}_1$
7. Calculate $\widehat{m}_2$
8. Calculate RA: If RA$< \epsilon$ STOP calculations and output $\hat{m} = \hat{m}_2$
9. If RA$\geq \epsilon$, return to step 6
10. continue the above loop until either RA is achieved or the specified number of iterations occurs

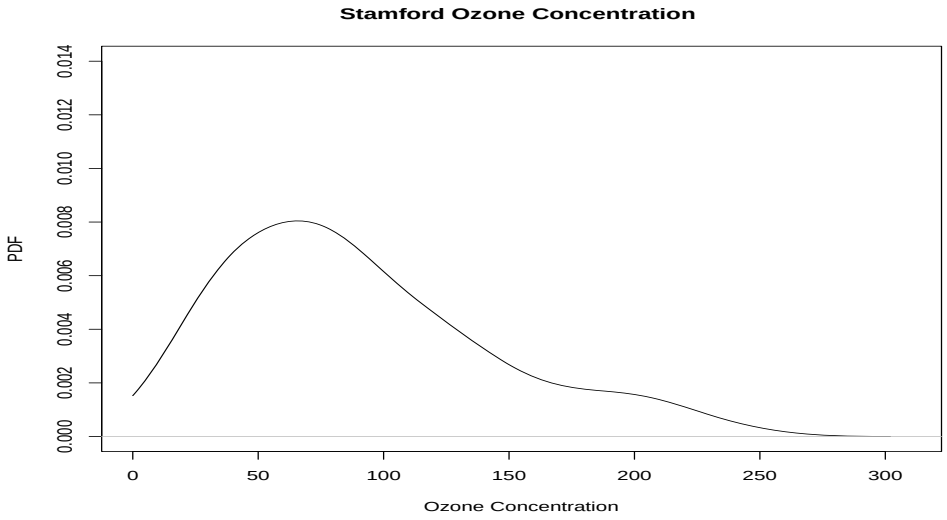


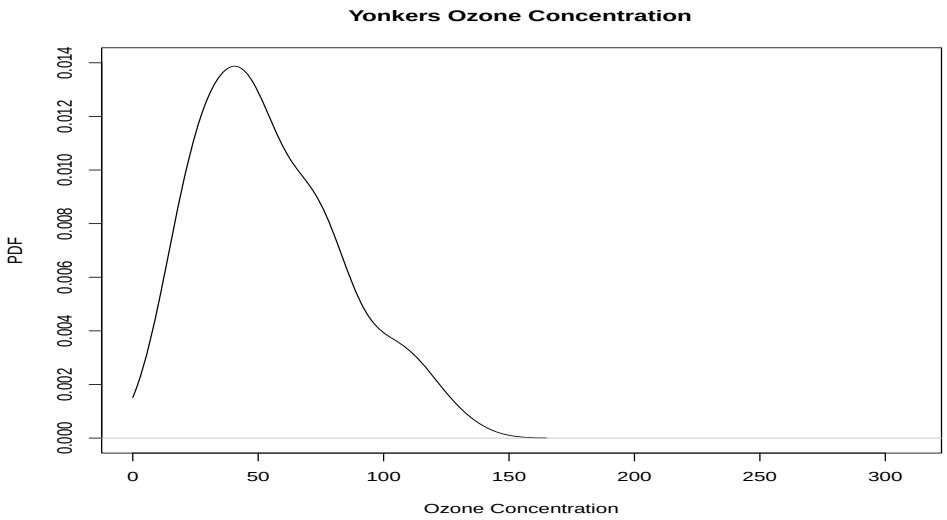
The computation of the sample statistics will be illustrated using the Ozone Data:

Maximum Daily Ozone Concentrations

Daily maximum ozone concentrations in parts per billion (ppb) at ground level recorded between May 1 and September 30, 1974 at sites in Stamford,Connecticut and Yonkers, New York are given below. (There are 17 missing days ofdata at Stamford and 5 at Yonkers due to equipment malfunction.) The current federal standard for ozone states that the concentration should not exceed 120 ppb more than one day per year at any particular location. A day with ozone concentration above 220 ppb is regarded as heavily polluted.

May		June		July		August		September	
Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs	Stmf	Ykrs
66	47	61	36	152	76	80	66	113	66
52	37	47	24	201	108	68	82	38	18
—	27	—	52	134	85	24	47	38	25
—	37	196	88	206	96	24	28	28	14
—	38	131	111	92	48	82	44	52	27
—	—	173	117	101	60	100	55	14	9
49	45	37	31	119	54	55	34	38	16
64	52	47	37	124	71	91	60	94	67
68	51	215	93	133	—	87	70	89	74
26	22	230	106	83	50	64	41	99	74
86	27	—	49	—	27	—	67	150	75
52	25	69	64	60	37	—	127	146	74
43	—	98	83	124	47	170	96	113	42
75	55	125	97	142	71	—	56	38	—
87	72	94	79	124	46	86	54	66	38
188	132	72	36	64	41	202	100	38	23
118	—	72	51	75	49	71	44	80	50
103	106	125	75	103	59	85	44	80	34
82	42	143	104	—	53	122	75	99	58
71	45	192	107	46	25	155	86	71	35
103	80	—	56	68	45	80	70	42	24
240	107	122	68	—	78	71	53	52	27
31	21	32	19	87	40	28	36	33	17
40	50	114	67	27	13	212	117	38	21
47	31	32	20	—	25	80	43	24	14
51	37	23	35	73	46	24	27	61	32
31	19	71	30	59	62	80	77	108	51
47	33	38	31	119	80	169	75	38	15
14	22	136	81	64	39	174	87	28	21
—	67	169	119	—	70	141	47	—	18
71	45			111	74	202	114		


Stamford Ozone Concentration
PDF
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0
50
100
150
200
250
300
Ozone Concentration


Yonkers Ozone Concentration
PDF
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0
50
100
150
200
250
300
Ozone Concentration

We will now analyze the ozone data using the following SAS and R code:

```
#The following R code generates various summary statistics for the
#Ozone data. The ozone data is in the files ozone1.DAT and ozone2.DAT
#The following file can be found in ~longneck/meth1/Rfiles/ozonesum.R
#------------------------------------------------------------------------------

#input the data from data files:

          y1  =  scan("u:/meth1/Rfiles/ozone1.DAT")
          y2  =  scan("u:/meth1/Rfiles/ozone2.DAT")
          y1p  =  scan("u:/meth1/Rfiles/ozone1+.DAT")
          y2p  =  scan("u:/meth1/Rfiles/ozone2+.DAT")

#compute summary statistics for Ozone data:

    MeanYkrs  =  mean(y1)
    MeanStmf  =  mean(y2)
    VarYkrs   =  var(y1)
    VarStmf   =  var(y2)
    StDevYkrs  =  sd(y1)
    StDevStmf  =  sd(y2)
    StErrMeanYkrs  =  sqrt(mean(y1)/length(y1))
    StErrMeanStmf  =  sqrt(mean(y2)/length(y2))
    MedianYkrs  =  median(y1)
    MedianStmf  =  median(y2)
    MinYkrs  =  min(y1)
    MinStmf  =  min(y2)
    MaxYkrs  =  max(y1)
    MaxStmf  =  max(y2)
    RangeYkrs  =  max(y1) - min(y1)
    RangeStmf  =  max(y2) - min(y2)
    Q.25Ykrs  =  quantile(y1,.25)
    Q.25Stmf  =  quantile(y2,.25)
    Q.75Ykrs  =  quantile(y1,.75)
    Q.75Stmf  =  quantile(y2,.75)
    IQRYkrs  =  Q.75Ykrs-Q.25Ykrs
    IQRStmf  =  Q.75Stmf-Q.25Stmf
    MadStmf  =  mad(y2)
    MadYkrs  =  mad(y1)

#compute summary statistics for Ozone data with 5 outliers added (1000, 1200, 1500, 2000, 2500):

    MeanYkrsp  =  mean(y1p)
    MeanStmfp  =  mean(y2p)
    VarYkrsp   =  var(y1p)
    VarStmfp   =  var(y2p)
    StDevYkrsp  =  sqrt(var(y1p))
    StDevStmfp  =  sqrt(var(y2p))
    StErrMeanYkrsp  =  sqrt(mean(y1p)/length(y1p))
    StErrMeanStmfp  =  sqrt(mean(y2p)/length(y2p))
    MedianYkrsp  =  median(y1p)
    MedianStmfp  =  median(y2p)
    MinYkrsp  =  min(y2p)
    MinStmfp  =  min(y1p)
    MaxYkrsp  =  max(y1p)
    MaxStmfp  =  max(y2p)
    RangeYkrsp  =  max(y1p) - min(y1p)
    RangeStmfp  =  max(y2p) - min(y2p)
    Q.25Ykrsp  =  quantile(y1p,.25)
    Q.25Stmfp  =  quantile(y2p,.25)
    Q.75Ykrsp  =  quantile(y1p,.75)
    Q.75Stmfp  =  quantile(y2p,.75)
    IQRYkrsp  =  Q.75Ykrsp-Q.25Ykrsp
    IQRStmfp  =  Q.75Stmfp-Q.25Stmfp
    MadStmfp  =  mad(y2p)
    MadYkrsp  =  mad(y1p)
```

```
  SumStat  =  c(MeanYkrs,MeanStmf,StDevYkrs,StDevStmf,MedianYkrs,MedianStmf,
                MinYkrs,MinStmf,MaxYkrs,MaxStmf,Q.25Ykrs,Q.25Stmf,Q.75Ykrs,
                Q.75Stmf,IQRYkrs,IQRStmf,MadYkrs,MadStmf)
  SumStatName  =  c("MeanYkrs","MeanStmf","StDevYkrs","StDevStmf","MedianYkrs",
                   "MedianStmf","MinYkrs","MinStmf","MaxYkrs","MaxStmf",
                   "Q.25Ykrs","Q.25Stmf","Q.75Ykrs","Q.75Stmf","IQRYkrs",
                   "IQRStmf","MadYkrs","MadStmf")
   SumStatp  =  c(MeanYkrsp,MeanStmfp,StDevYkrsp,StDevStmfp,MedianYkrsp,
                MedianStmfp,MinYkrsp,MinStmfp,MaxYkrsp,MaxStmfp,Q.25Ykrsp,
                Q.25Stmfp,Q.75Ykrsp,Q.75Stmfp,IQRYkrsp,IQRStmfp,MadYkrsp,
                MadStmfp)

Round summary statistics to 2 decimal points:

  SumStat  =  round(SumStat,2)

  SumStatp  =  round(SumStatp,2)

  SumStat  =  cbind(SumStatName,SumStat,SumStatp)


 #Output summary statistics to file named SumOzone:

  sink("SumOzone")

  SumStat

  sink()

#     SumStatName  SumStat SumStatp
# ---------------------------------
#     "MeanYkrs"   "89.67"  "144.65"
#     "MeanStmf"   "54.69"  "106.5"
#     "StDevYkrs"  "52.11"  "309.95"
#     "StDevStmf"  "28.11"  "300.92"
#     "MedianYkrs" "80"     "80"
#     "MedianStmf" "49.5"   "50"
#     "MinYkrs"    "14"     "14"
#     "MinStmf"    "9"      "9"
#     "MaxYkrs"    "240"    "2500"
#     "MaxStmf"    "132"    "2500"
#     "Q.25Ykrs"   "48.5"   "51"
#     "Q.25Stmf"   "33.75"  "34"
#     "Q.75Ykrs"   "119.75" "124"
#     "Q.75Stmf"   "74"     "75"
#     "IQRYkrs"    "71.25"  "73"
#     "IQRStmf"    "40.25"  "41"
#     "MadYkrs"    "30.39"  "31.13"
#     "MadStmf"    "49.67"  "54.86"
```

```
* SAS program to obtain summary statistics for ozone data. ;


option ls=75 ps=55 nocenter nodate;
title 'Ozone Concentration';
data OZONE1;
   infile 'C:\sasdata\ozone1.DAT';           * input data;
      input ozone1 @@;
data OZONE2;
   infile 'C:\sasdata\ozone2.DAT';           * input data;
      input ozone2 @@;
data ozone;
      merge ozone1 ozone2;                              * combine data sets;

   label OZONE1='Ozone Level in Stamford'
         OZONE2='Ozone Level in Yonkers';

run;
proc print;
run;

* generates various summary statistics and plots;

proc univariate plot normal def=5;
   var ozone1 ozone2;

* create data set with ozone values in one column
  and city name in second column;

data ozonebox1;
set ozone1;
ozone=ozone1;drop ozone1;
run;
data ozonebox2;
set ozone2;
ozone=ozone2;drop ozone2;
run;
data ozonebox;
set ozonebox1 ozonebox2;
if _n_<137 then city="Stamford";
if _n_>=137 then city="Yonkers";drop obs;
run;

* creates side-by-side box plots;

proc boxplot;
   plot ozone*city/boxstyle=schematic;
run;
```

The UNIVARIATE Procedure
Variable: ozone1 (Ozone Level in Stamford)

Moments

N	136	Sum Weights	136
Mean	89.6691176	Sum Observations	12195
Std Deviation	52.1074467	Variance	2715.186
Skewness	0.88432102	Kurtosis	0.18080786
Uncorrected SS	1460065	Corrected SS	366550.11
Coeff Variation	58.1108057	Std Error Mean	4.46817669

Location		Variability		Quantile	Estimate
		100% Max	240.0		
Mean	89.66912	Std Deviation	52.10745	75% Q3	120.5
Median	80.00000	Variance	2715	50% Median	80.0
Mode	38.00000	Range	226.00000	25% Q1	48.0
		Interquartile Range	72.50000	0% Min	14.0

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.928464	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.116014	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.422695	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.754894	Pr > A-Sq	<0.0050

```
  Stem Leaf                     #  Boxplot
    24 0                        1     0
    23 0                        1     0
    22
    21 25                       2     |
    20 1226                     4     |
    19 26                       2     |
    18 8                        1     |
    17 034                      3     |has has
    16 99                       2     |
    15 025                      3     |
    14 1236                     4     |
    13 1346                     4     |
    12 2244455                  7  +-----+
    11 1334899                  7  |     |
    10 013338                   6  |     |
     9 1244899                  7  |  +  |
     8 0000002235667779        16  *-----*
     7 11111122355             11  |     |
     6 0114444668889           13  |     |
     5 1222259                  7  |     |
     4 023677779                9  +-----+
     3 11223788888888          14     |
     2 3444467888              10     |
     1 44                       2     |
       ----+----+----+----+
   Multiply Stem.Leaf by 10**+1
```

```
The UNIVARIATE Procedure
Variable:  ozone2  (Ozone Level in Yonkers)

                          Moments

N                         148    Sum Weights               148
Mean                 54.6891892    Sum Observations         8094
Std Deviation        28.1148885    Variance            790.446957
Skewness             0.65601648    Kurtosis            -0.2458541
Uncorrected SS           558850    Corrected SS        116195.703
Coeff Variation       51.408494    Std Error Mean       2.3110296

    Location                    Variability              Quantile      Estimate

Mean      54.68919     Std Deviation 28.11489            100% Max         132.0
Median    49.50000     Variance               790.44696   75% Q3            74.0
Mode      27.00000     Range                  123.00000   50% Median        49.5
                       Interquartile Range     40.50000   25% Q1            33.5
                              0% Min             9.0

                  Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.952488    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.092739    Pr > D     <0.0100
Cramer-von Mises      W-Sq  0.296356    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  1.927373    Pr > A-Sq  <0.0050

   Stem Leaf                     #  Boxplot
     13 2                        1     |
     12 7                        1     |
     12                          0     |
     11 779                      3     |
     11 14                       2     |
     10 66778                    5     |
     10 04                       2     |
      9 667                      3     |
      9 3                        1     |
      8 5678                     4     |
      8 00123                    5     |
      7 55556789                 8     |
      7 0001124444              10  +-----+
      6 6677778                  7  |     |
      6 0024                     4  |     |
      5 556689                   6  |  +  |
      5 000111223344            12  *-----*
      4 5555667777899           13  |     |
      4 011223444                9  |     |
      3 5566677777889           13  |     |
      3 01112344                 8  +-----+
      2 55557777778             11     |
      2 011122344                9     |
      1 5678899                  7     |
      1 344                      3     |
      0 9                        1     |
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+1
```

```
#The following R code will be used to
#calculate the m-estimator of a location parameter using 30 iterations

# mest1.R:
x  =  scan("u:/meth1/ozone1.DAT")
t  =  1.345
mx  =  median(x)
y  =  abs(x-mx)
my  =  median(y)
v  =  my/.6745
n  =  length(x)
k  =  30
r  =  k+1
mest1  =  matrix(0,r,1)
s =  matrix(0,r,1)
p =  matrix(0,r,1)
mest1[1]  =  median(x)
w  =  matrix(0,n,1)
for(j in 1:k)
{
m  =  mest1[j]
for(i in 1:n)
{
if               (x[i] < m-t*v)  w[i]  =  -t*v/(x[i]-m)
if (m-t*v <= x[i] && x[i] <= m+t*v)      w[i]  =  1
if               (x[i] > m+t*v)  w[i]  =  t*v/(x[i]-m)
}
s[j]  =  sum(w)
p[j]  =  t(x)%*%w
mest1[j+1]  =  p[j]/s[j]
}



#-------------------------------------------------------
#
#The following R code will be used to
#calculate the m-estimator of a location parameter using iterations
#until a specified (1.e-8) degree of precision is achieved.
#-----------------------------------------------------------------------
#mest2.s:

x  =  scan("u:/meth1/ozone1.DAT")
t  =  1.345
mx  =  median(x)
y  =  abs(x-mx)
mad  =  median(y)
v  =  mad/.6745
n  =  length(x)
mest2  =  median(x)
m  =  mest2
lastm  =  2*m
w  =  matrix(0,n,1)
nit  =  1
while(abs((mest2-lastm)/lastm)>1.e-8)
{
lastm  =  mest2
m  =  mest2
for(i in 1:n)
{
if               (x[i] < m-t*v)  w[i]  =  -t*v/(x[i]-m)
if (m-t*v <= x[i] && x[i] <= m+t*v)      w[i]  =  1
if               (x[i] > m+t*v)  w[i]  =  t*v/(x[i]-m)
}
s  =  sum(w)
p  =  t(x)%*%w
mest2  =  p/s
```

```
nit  =  nit+1

}
```

OUTPUT FROM mest1.s and mest2.s

```
         Stamford Ozone                          Yonkers Ozone

        mest1       mest2                   mest1           mest2
 [1,] 80.00000     84.39501           [1,] 49.50000      52.88504
 [2,] 83.93129      nit=9             [2,] 52.56577        nit=9
 [3,] 84.34605                        [3,] 52.85504
 [4,] 84.38983                        [4,] 52.88221
 [5,] 84.39446                        [5,] 52.88477
 [6,] 84.39495                        [6,] 52.88501
 [7,] 84.39501                        [7,] 52.88503
 [8,] 84.39501                        [8,] 52.88504
 [9,] 84.39501                        [9,] 52.88504
[10,] 84.39501                       [10,] 52.88504
...............                      [29,] ........
[30,] 84.39501                       [30,] 52.88504
```

With the outliers(1000, 1200, 1500, 2000, 2500) added to the data,
we obtain:

```
        Stamford     Yonkers
------------------------------
mest1   88.3884      54.61086
mest2   88.3884      54.61086
nit       9            9
```

mest1 was specified to quit after 30 iterations
mest2 took 9 iterations to achieve convergence

START: Wednesday 9/29/21

Comparison of Summary Statistics for Ozone Data

w/o outliers
w/ outliers

Estimators of Center of Data Set				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Sample Size: n	148	153	136	141
Mean: $\bar{Y}$	54.69	106.50	89.67	144.65
$\%Y_{i's} \leq \bar{Y}$	58.1%	90.2%	67.6%	82.3%
Median: M	49.0	50.0	80.0	80.0
$\%Y_{i's} \leq M$	50.0%	50.3%	52.9%	51.1%
5% Trimmed Mean: $T_{.05}$	53.75	55.66	86.4	91.17
$\%Y_{i's} \leq T_{.05}$	56.7%	58.8%	57.3%	58.9%
M-EST: $MEST$	52.89	54.61	84.33	88.40
$\%Y_{i's} \leq MEST$	55.4%	56.2%	55.1%	57.4%
Estimators of Level of Dispersion in Data Set				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Sample Size: n	148	153	136	141
Range: R	123	2491	226	2486
Semi-Interquarile Range: $SIQR$	20.38	20.50	36.75	36.50
Standard Deviation: S	28.11	300.90	52.11	310.00
MAD_3	30.39	31.13	49.67	54.86
$Skewness$	0.66	6.16	0.88	5.76
$Kurtosis$	-0.25	39.94	0.18	35.53
Five Number Summaries of Data				
	Yonkers		Stamford	
Data Set	Without	With	Without	With
Minimum	9	9	14	14
Q(.25)	33.75	34	48.5	51
Q(.5)	49.5	50	80	80
Q(.75)	74	75	119.75	124
Maximum	132	2500	240	2500

The Without Column refers to the original data set

The With Column data set with 5 large values added: 1000, 1200, 1500, 2000, 2500

Selecting a Measure of Center and Dispersion about Center

Why not just use the pair $(\widehat{Q(.5)}, \widehat{MAD})$ for all data sets?

When the data is not heavily skewed or has only a few outliers (near normal in shape), $(\widehat{\mu}, \widehat{\sigma})$ provide a more complete picture of the distribution. Furthermore, $(\widehat{\mu}, \widehat{\sigma})$ are more efficient estimators than are $(\widehat{Q(.5)}, \widehat{MAD})$. We will discuss these comments in a later handout and in STAT 611. We will now consider some recommendations on how to select an estimator for a broad class of distributions.

Robust Estimation of Location and Scale

A robust estimator should be relatively unaffected by two types of anomalies that are often encountered in data sets:

1. A few outliers - values that are large relative to the other data values
2. Many relatively small deviations in the data which may occur due to rounding or grouping of the data

There are three broad classes of robust estimators of location and scale:

1. R-Estimators: Estimators based on linear combinations of the ranks of the data values - Rank based regressions
2. M-Estimators: Estimators which minimize an objective function - Least Squares Estimators
3. L-Estimators: Estimators which are linear combinations of the order statistics

A comparison of a number of L-Estimators is given in the book, *Understanding Robust and Exploratory Data Analysis*, by D. Hoaglin, F. Mosteller, and J. Tukey. All of the following estimators have symmetric coefficients and hence are unbiased estimators of a location parameter.

1. α-Trimmed Mean:

$$T(\alpha) = \frac{1}{n(1-2\alpha)}\left(pY_{([n\alpha]+1)} + pY_{(n-[n\alpha])} + \sum_{i=[n\alpha+2]}^{n-[n\alpha]-1} Y_{(i)}\right)$$

where $p = 1 + [n\alpha] - n\alpha$. This definition is slightly modified from our definition in order to better approximate trimming exactly $100\alpha\%$ from each tail of the data. When $n\alpha$ is an integer, the definitions agree. When $n\alpha$ is not an integer, we are only using $100p\%$ of the smallest and largest untrimmed data values in the average.

Example Suppose $n = 30,\ \alpha = .05,\ n\alpha = 1.5,\ p = 1 + [1.5] - 1.5 = .5$. Thus, trim $Y_{(1)}$ and $Y_{(30)}$ and partially trim $Y_{(2)}$ and $Y_{(29)}$ yielding

$$T(.05) = \frac{1}{27}\left(.5Y_{(2)} + .5Y_{(29)} + \sum_{i=3}^{28} Y_{(i)}\right)$$

2. Mean: The average of all n data values, a 0% trimmed mean.

3. MidMean: The average of the central half of the order statistics. This is just $T(.25)$.

4. Median (M): A variably trimmed mean with trimming proportion equal to $\alpha_n = \frac{1}{2} - \frac{1}{2n}$

5. Trimean (TRI): $TRI = \frac{1}{4}\left(\widehat{Q}(.25) + 2M + \widehat{Q}(.75)\right)$

6. Best Linear Unbiased Estimator (BLUE): The linear combination of the order statistics (L-estimator) which is unbiased and has smallest variance among all L-estimators.

The goal is to evaluate the above estimators of location over a broad class of symmetric distributions. In order to compare the mass in the tails of the distributions relative to the normal distribution, we will define the following index of tail weight in a distribution:

Let F be a symmetric distribution with quantile function Q. Let Q_o be the standard normal quantile function. The Tail Weight Index of a distribution is defined as

$$\tau(F) = \frac{Q(.99) - Q(.5)}{Q(.75) - Q(.5)} \Big/ \frac{Q_o(.99) - Q_o(.5)}{Q_o(.75) - Q_o(.5)}$$

The value of $\tau(F)$ for various distributions is given in the following table:

Dist.	Uniform	Triangular	Normal	CN(.05;3)	Logistic	D-Exp	CN(.05,10)	Slash	Cauchy
$\tau(F)$	.57	.86	1.00	1.20	1.21	1.63	3.42	7.85	9.22

1. CN(α,k) is a contaminated normal distribution: $F(x) = (1-\alpha)\Phi(x) + \alpha F_2(x)$ where Φ is $N(0,1)$ and F_2 is $N(0,k^2)$

2. D-Exp is the Double Exponential Distribution

3. Slash: $F(x) = \Phi(x) - xf(x)$ with $f(x) = \frac{1-e^{-x^2/2}}{x^2\sqrt{2\pi}}$ for $x \neq 0$ and $f(0) = \frac{1}{2\sqrt{2\pi}}$

Two other sampling schemes will be considered in comparing the estimators.

1. One-Out is a sample with one observation from a $N(0,3^2)$ distribution and $n-1$ observations from a $N(0,1)$

2. One-Wild is a sample with one observation from a $N(0,10^2)$ distribution and $n-1$ observations from a $N(0,1)$

The following tables display the variance of the estimators for the above distributions. For each of the following distributions, the preferred estimator is the one having smallest variance.

Table 1: Variance of Estimators based on a sample of size 10

	Distributions and Sampling Situations						
Estimator	Normal	One-Out	Logistic	One-Wild	D-Exp	Slash	Cauchy
Best Trim	.1000	.1295	.0940	.1416	.1403	.6995	.3362
(Trim %)	(0%)	(11%)	(13%)	(16%)	(34%)	(38%)	(40%)
Mean(0%)	.1000	.1800	.1000	1.090	.2000	∞	∞
T(10%)	.1053	.1296	.0943	.1432	.1617	∞	∞
T(20%)	.1133	.1339	.0962	.1433	.1463	.9649	.5377
TriMean	.1136	.1348	.0971	.1448	.1503	1.114	.6348
MidMean(25%)	.1164	.1366	.0975	.1454	.1424	.8389	.4498
T(30%)	.1238	.1438	.1019	.1521	.1408	.7252	.3672
BMED(35%)	.1270	.1472	.1039	.1554	.1404	.7063	.3500
Median(40%)	.1383	.1596	.1120	.1679	.1452	.7048	.3362
BLUE	.1000	*	.0935	*	.1399	*	.3263

Table 2: Variance of Estimators based on a sample of size 20

	Distributions and Sampling Situations						
Estimator	Normal	One-Out	One-Wild	Logistic	Slash	D-Exp	Cauchy
Best Trim	.0500	.0578	.0610	.0464	.3039	.0638	.1357
(Trim %)	(0%)	(6%)	(9%)	(12%)	(34%)	(37%)	(39%)
Mean(0%)	.0500	.0700	.2975	.0500	∞	.1000	∞
T(5%)	.0512	.0578	.0619	.0472	∞	.0854	∞
T(10%)	.0528	.0583	.0611	.0465	.6964	.0778	.4141
T(20%)	.0568	.0616	.0637	.0474	.3579	.0690	.1874
TriMean	.0576	.0625	.0646	.0482	.3849	.0702	.2045
MidMean(25%)	.0593	.0640	.0659	.0486	.3221	.0664	.1591
T(30%)	.0621	.0668	.0686	.0502	.3071	.0647	.1444
BMED(37.5%)	.0664	.0713	.0731	.0530	.3048	.0638	.1361
T(40%)	.0689	.0739	.0757	.0547	.3092	.0644	.1358
Median(45%)	.0734	.0787	.0806	.0581	.3229	.0666	.1395
BLUE	.0500	*	*	.0462	*	*	.1257

For the group of distributions Normal, One-Out, Logistic, One-Wild, Slash, the best overall estimator is the MidMean for n=10 and $T(20\%)$ for n=20.

For the group of distributions One-Out, Logistic, One-Wild, Slash, the best overall estimator is the $T(30\%)$ for n=10 and MidMean(25%) for n=20.

For the group of distributions Normal, One-Out, Logistic, One-Wild, the best overall estimator is the $T(10\%)$ for n=10 and T(5%) for n=20.

Measures of Association Amongst Vectors of R.V.s

We will now define the sample estimators of the correlation coefficient and autocorrelation function:

Pearson Correlation Coefficient

$$Corr(Y,W) = \frac{E[(Y-\mu_Y)(W-\mu_W)}{\sigma_Y \sigma_W}$$

Definition: Estimator of Pearson Correlation Coefficient is based on having n independent pairs (Y_i, W_i) of observations on possibly correlated random variables Y and W.

$$r_{Y,W} = \widehat{Corr}(Y,W) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(W_i-\bar{W})}{(s_Y)(s_W)} = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(W_i-\bar{W})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\sqrt{\sum_{i=1}^{n}(W_i-\bar{W})^2}}$$

where s_Y and s_W are the sample standard deviations

1. The Pearson correlation coefficient is a unit-free measure of the linear relationship between the two variables.
2. $-1 \leq r_{Y,W} \leq 1$
3. $r_{Y,W} = \pm 1$ implies $Y_i = \beta_o + \beta_1 W_i$ where the sign of β_1 is the same as the sign of $r_{Y,W}$
4. $r_{Y,W}$ has the limitation of only measuring linear relationship. Thus, higher order relationships may not be detected
5. $r_{Y,W}$ only measures linear relationships between two of the many variables under study. Thus, may fail to detect linear/nonlinear relationships that exist between several of the variables simultaneously.

FIGURE 11.20
Samples of size 1,000 from the bivariate normal distribution

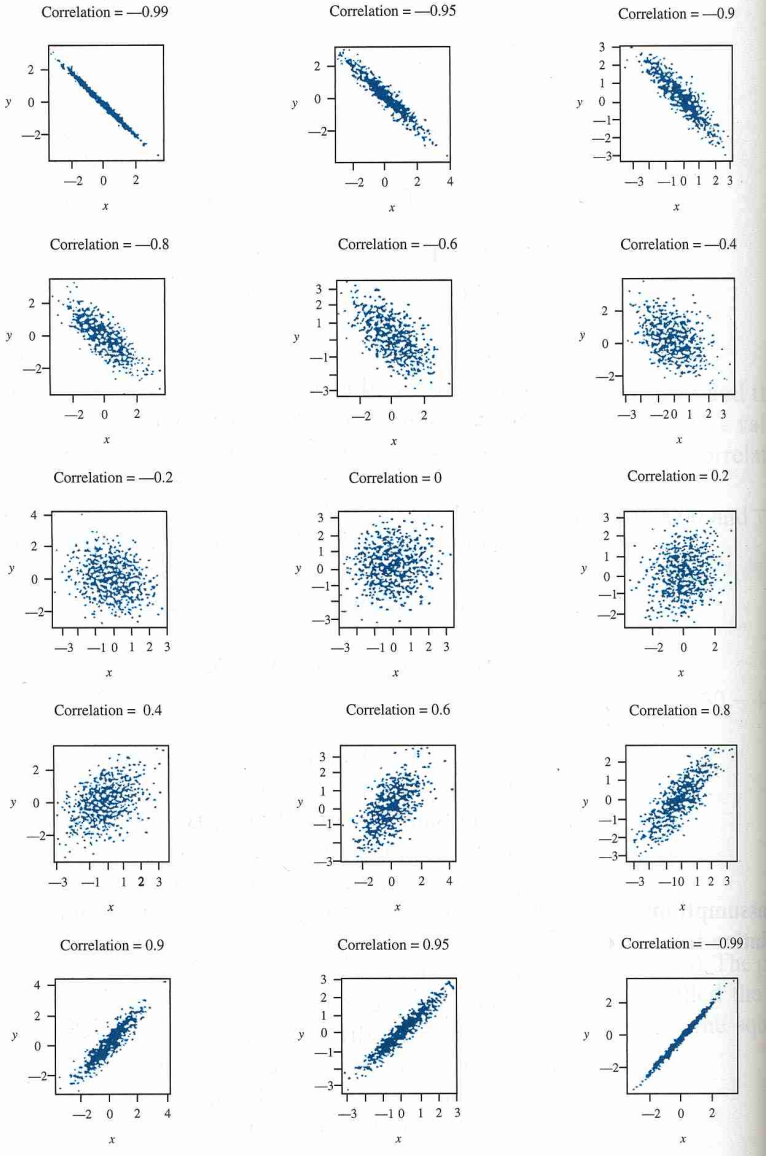


TABLE 3.10 Data for epilepsy study: successive 2-week seizure counts for 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = Progabide), 8-week baseline seizure counts, and age (in years)

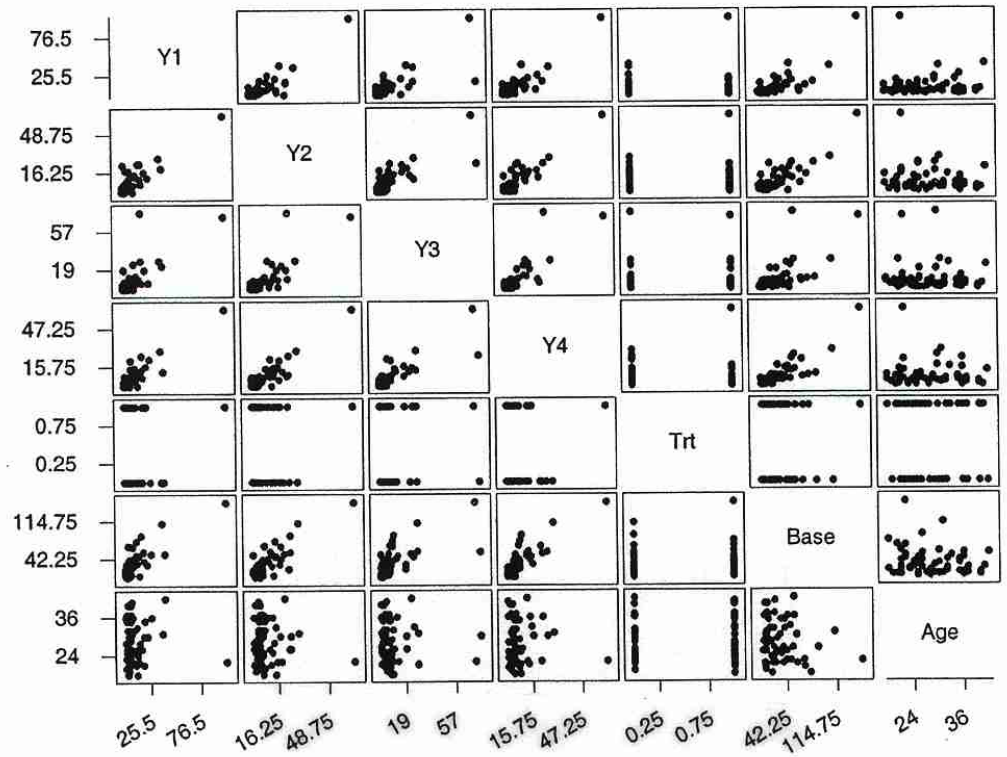
ID	y_1	y_2	y_3	y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

Correlations: Y1, Y2, Y3, Y4, Base, Age

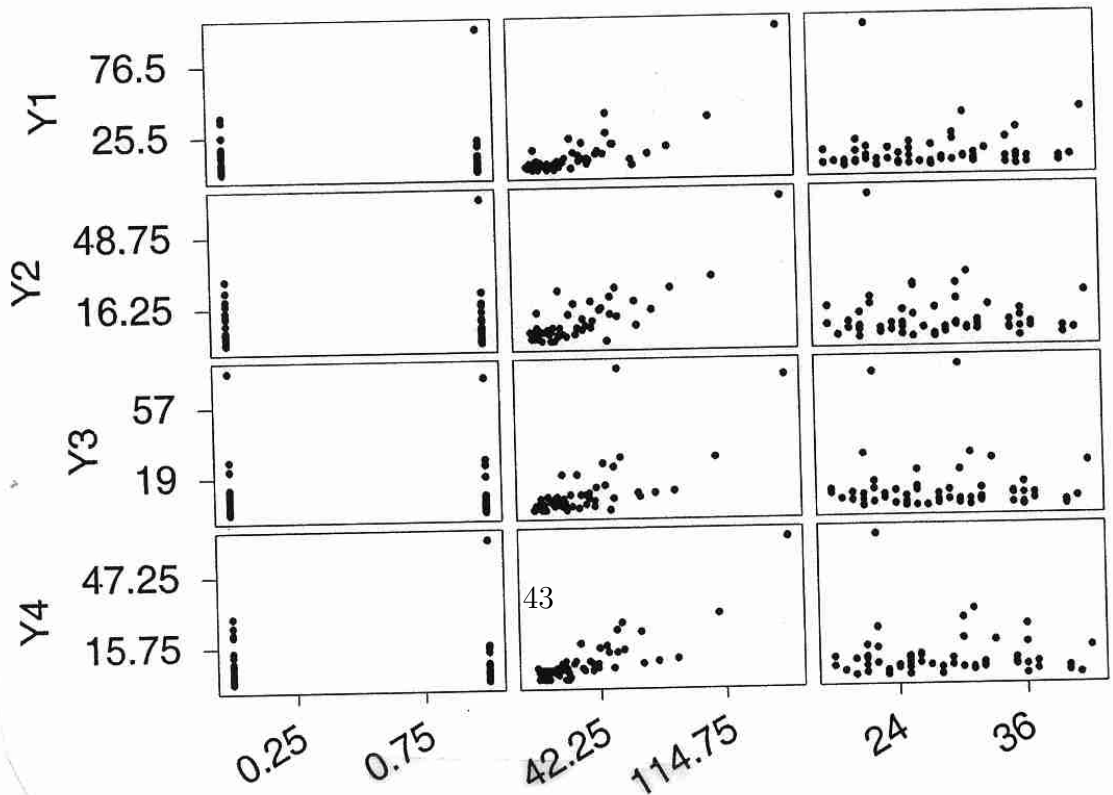
	Y1	Y2	Y3	Y4	Base
Y2	0.871				
	0.000				
Y3	0.738	0.802			
	0.000	0.000			
Y4	0.892	0.895	0.824		
	0.000	0.000	0.000		
Base	0.796	0.831	0.672	0.843	
	0.000	0.000	0.000	0.000	
Age	0.008	-0.116	-0.049	-0.077	-0.181
	0.955	0.384	0.714	0.563	0.171

Cell Contents: Pearson correlation
P-Value

Matrix Plot of Epilpsy Data



Draftsman Plot of Epilpsy Data



Spearman Rank Correlation

When the relationship between the observations on two variables Y and X has a monotone relationship which is not linear, the Pearson correlation coefficient may not reflect the strength of the relationship between Y and X. Also, when there are extreme values in the data set with respect to the x-variable, called influential observations, the value of the correlation coefficient can be inflated. These two cases are illustrated in the following two graphs.

Figure 1: Nonlinear Relationship Y vs X

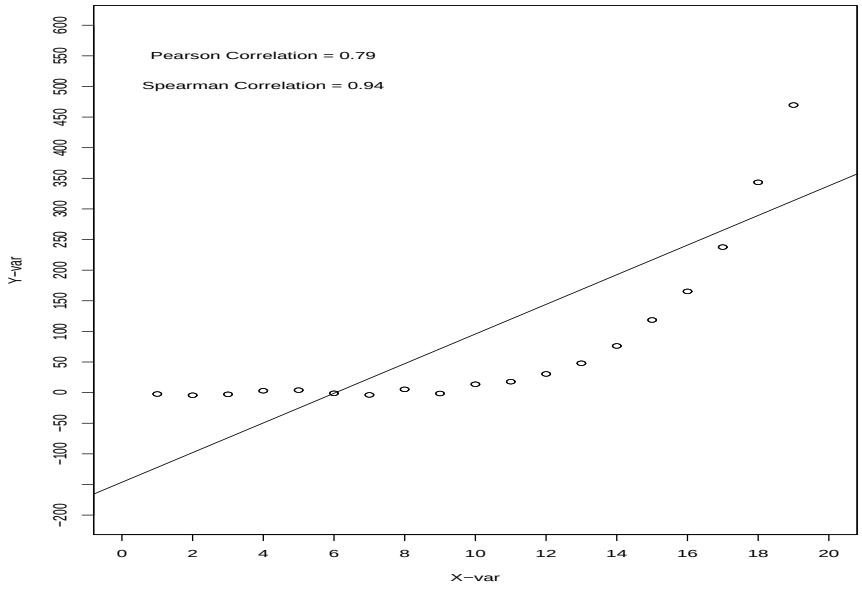
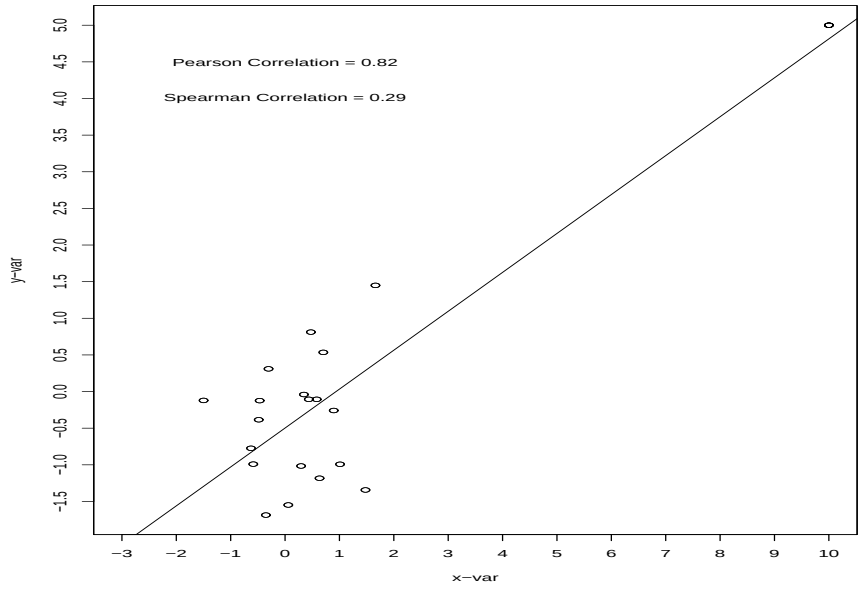


Figure 2: One Influential X Value



Spearman Correlation Coefficient An alternative to the Pearson correlation coefficient is the Spearman Rank Correlation coefficient.

Suppose we have n pairs of observation $(X_i, Y_i), \quad i = 1, \ldots, n$. Let R_i denote the rank of X_i among the Xs and S_i denote the rank of Y_i among the Ys. If there are tied X values, the average rank is assigned to the tied observations. A similar assignment is done for the Ys. The Spearman rank correlation, denoted by r_{sp} is obtained by applying the formula for the Pearson correlation coefficent to the pairs of ranks $(R_i, S_i), \quad i = 1, \ldots, n$.

Definition The Spearman Correlation Coefficient: let $(X_i, Y_i) \;\; i = 1, \ldots, n$ be independent pairs of random variables. Let $(R_i, S_i), \quad i = 1, \ldots, n$ be the corresponding ranks of (X_i, Y_i). The Spearman Correlation Coefficient is given by

$$r_{sp} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{s_R s_S} = \frac{12\sum_1^n R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

If there are no ties in the data or if all the values of R_i and S_i are integers, then we have the following simplification to r_{sp}:

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad \text{where} \quad d_i = R_i - S_i$$

The values of r_{sp} is now compared to r for two data sets given above.

Note that in the case of a nonlinear relationship between X and Y, Figure 1, the Pearson Correlation Coefficient underestimates the strength of the relationship, $r = .79$. The Spearman Correlation Coefficient identifies the strong non-linear relationship between X and Y, $r_{sp} = .94$.

In Figure 2, there is no relationship between X and Y but there is one data value which is very extreme to the other data values in the X-direction. This results in a relatively large value for the Pearson Correlation Coefficient, $r = .82$ whereas the Spearman Correlation Coefficient is small, $r_{sp} = .29$. Thus, reflecting the lack of a relationship between X and Y.

Sample Autocorrelation Function

When we are observing a physical characteristic over time, we are interested in the degree to which these measurements are associated. One measure of this association is the autocorrelation:

Definition: **The AutoCorrelation of Order k** in a series of stationary random variables: $X_t : t = 1, 2, 3, \ldots$ having the same mean μ and standard deviation σ is given by

$$\rho_k = \frac{E[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2} \quad \text{for} \quad k = 1, 2, \ldots$$

ρ_k measures the degree of linear relationship in the random variable X over time or space. ρ_1 the 1st order autocorrelation is the most widely used of these correlations.

A very simple, but widely used model for correlated over time or space observations is the $AR(1)$ model:

$$X_t = \mu + \rho X_{t-1} + e_t,$$

where $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma^2$, $e_t s$ are independent of the $X_t s$ and $|\rho| < 1$.

Under this model, we can show that the $X_t s$ are not independent and
$\rho_k = \rho^k \to 0$ as $k \to \infty$

The computation of the sample autocorrelation function of lag k is given by

$$\hat{\rho_k} = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

A plot of the time series and autocorrelation functions is given on the next pages. These plots and the computation of the sample autocorrelation function are obtained using the following R code:

```
#The following R code produces plots which take into account the
#time element in the ozone data. The file is named ozone_ts.R
#-------------------------------------------------------------------------
#input data:

          y1 = scan("u:/meth1/Rfiles/ozone1.DAT")
          y2 = scan("u:/meth1/Rfiles/ozone2.DAT")

          y1na = scan("u:/meth1/Rfiles/ozone1,na.DAT")
          y2na = scan("u:/meth1/Rfiles/ozone2,na.DAT")

          t1 = c(1:136)
          t2 = c(1:148)


         plot.ts(y1na,type="b",ylab="Ozone Conc-Stamford (ppb)",xlab="DAY",
          main="Time Series Plot of Stamford Data",cex=.9)
         abline(h=90)
```

```
plot.ts(y2na,type="b",ylab="Ozone Conc-Yonkers (ppb)",xlab="DAY",
 main="Time Series Plot of Yonkers Data",cex=.9)
 abline(h=55)


acf_S=acf(y1,main="ACF for Stamford Ozone Concentration")

acf_Y=acf(y2,main="ACF for Yonkers Ozone Concentration")

sink("u:/meth1/psfiles/autocorr")
acf_S = acf(y1,plot=F)
acf_S
acf_Y = acf(y2,plot=F)
acf_Y
sink()
```

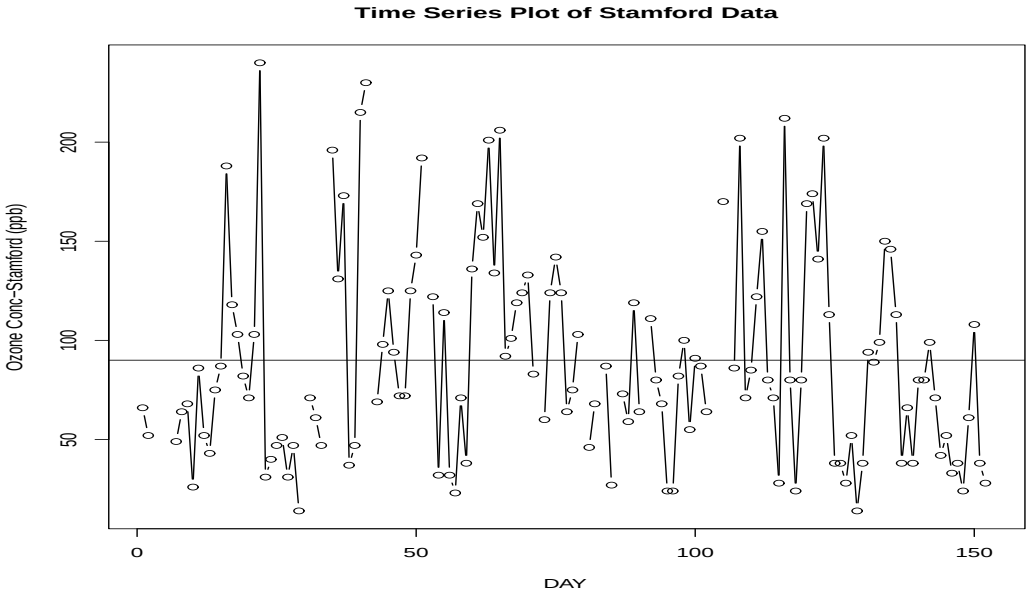
The sample estimators of ρ_k are given in the following table:

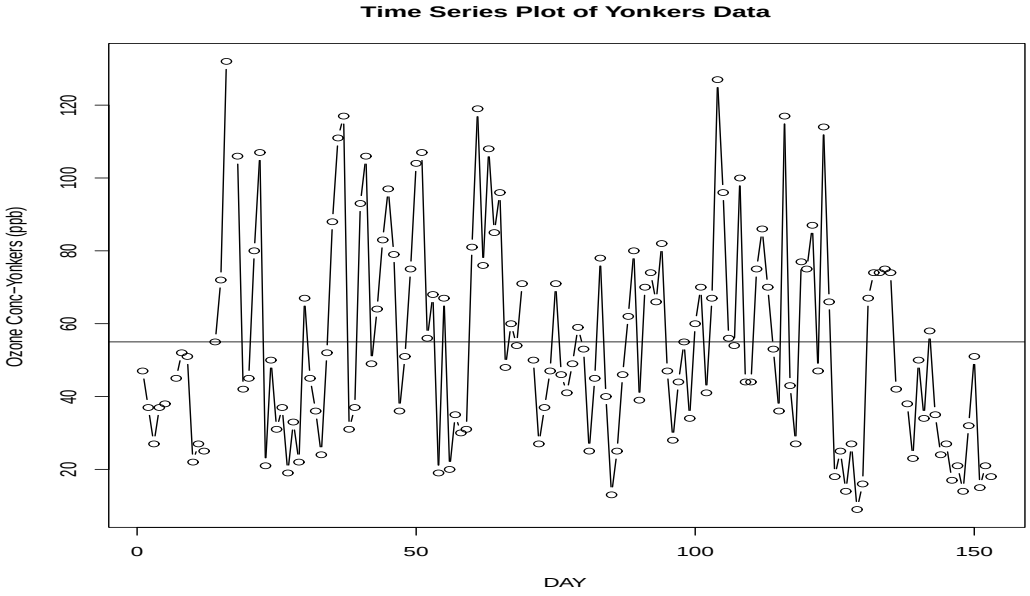
AUTOCORRELATION MATRIX:

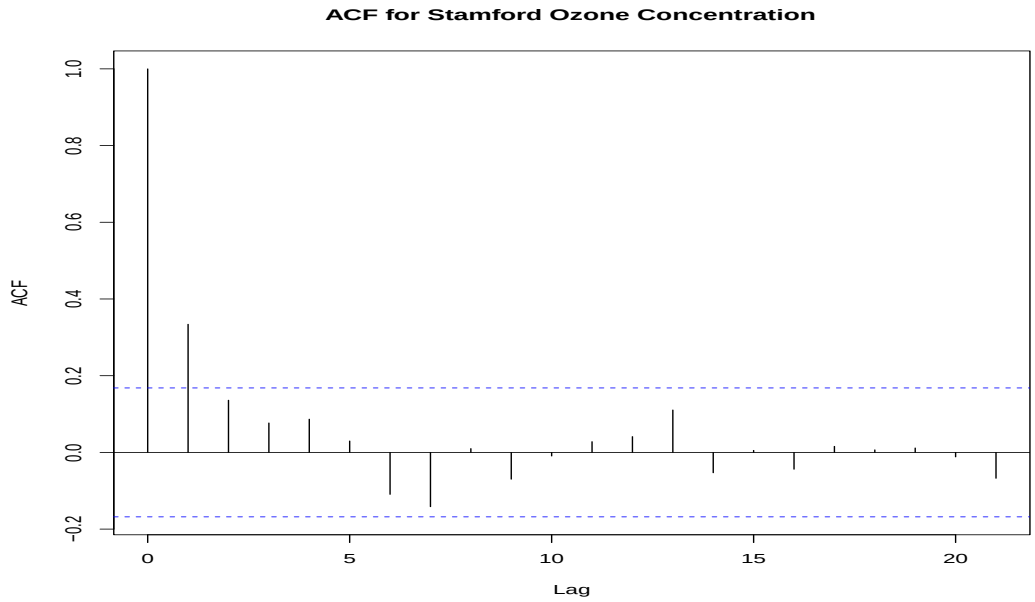
LAG k	YONKERS OZONE rho_k
0	1.0000
1	0.4342
2	0.1352
3	0.0805
4	0.1828
5	0.0621
6	-0.0993
7	-0.0694
8	-0.0130
9	-0.0237
10	0.0008
11	-0.0138
12	0.0385
13	0.0251
14	0.0651
15	0.1280
16	0.0144
17	-0.0690
18	0.0583
19	0.1566
20	0.0553
21	-0.0617

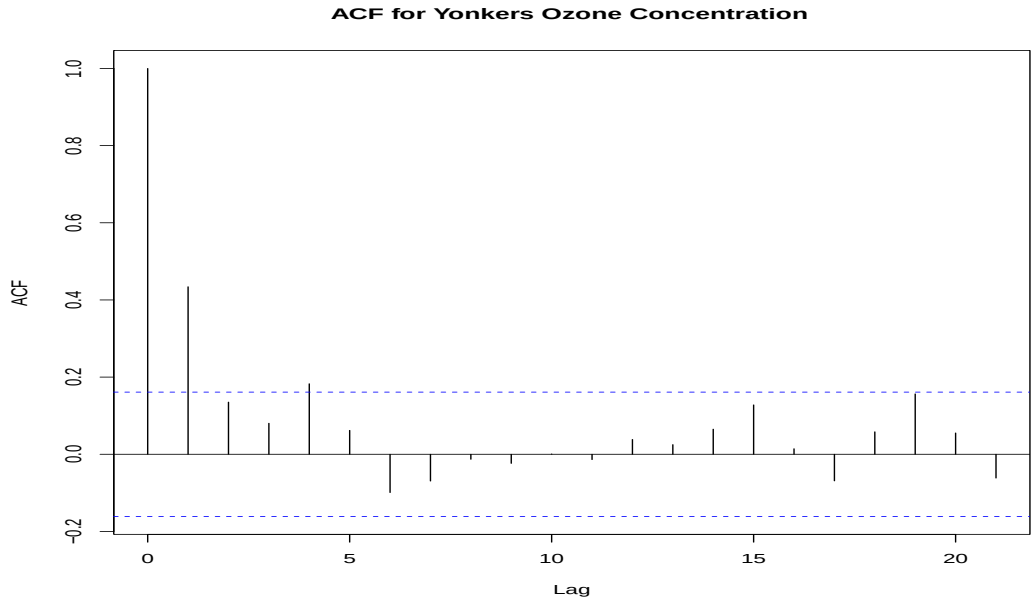
AUTOCORRELATION MATRIX:

LAG k	STAMFORD OZONE rho_k
0	1.0000
1	0.3342
2	0.1361
3	0.0768
4	0.0868
5	0.0298
6	-0.1095
7	-0.1417
8	0.0101
9	-0.0700
10	-0.0095
11	0.0281
12	0.0413
13	0.1106
14	-0.0532
15	0.0054
16	-0.0440
17	0.0159
18	0.0067
19	0.0116
20	-0.0118
21	-0.0676


Time Series Plot of Stamford Data
Ozone Conc–Stamford (ppb)
50
100
150
200
0
50
100
150
DAY


Time Series Plot of Yonkers Data
Ozone Conc–Yonkers (ppb)
20
40
60
80
100
120
0
50
100
150
DAY


ACF for Stamford Ozone Concentration
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag


ACF for Yonkers Ozone Concentration
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag

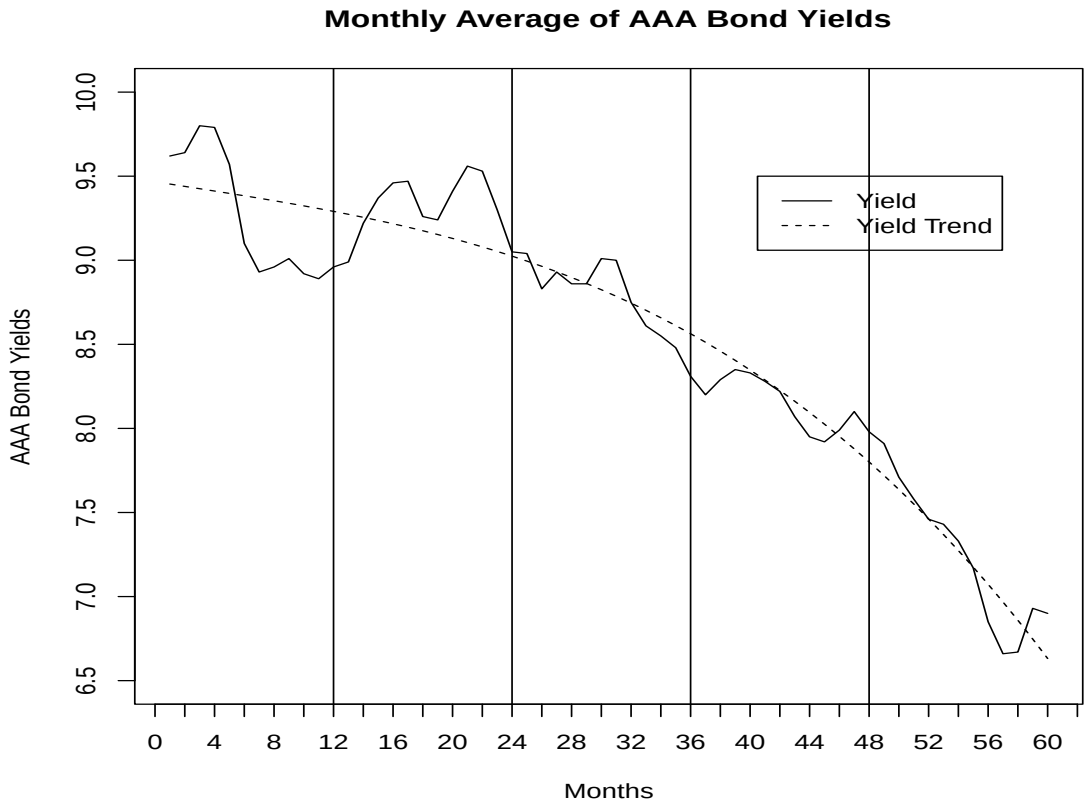
The time series plots for the Stamford and Yonkers ozone data appears to be relatively stationary with no trends over the 5 month summer period.

The following data is the monthly average of daily yields of Moody's AAA bonds for the years 1989 to 1993.

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1989	9.62	9.64	9.80	9.79	9.57	9.10	8.93	8.96	9.01	8.92	8.89	8.96
1990	8.99	9.22	9.37	9.46	9.47	9.26	9.24	9.41	9.56	9.53	9.30	9.05
1991	9.04	8.83	8.93	8.86	8.86	9.01	9.00	8.75	8.61	8.55	8.48	8.31
1992	8.20	8.29	8.35	8.33	8.28	8.22	8.07	7.95	7.92	7.99	8.10	7.98
1993	7.91	7.71	7.58	7.46	7.43	7.33	7.17	6.85	6.66	6.67	6.93	6.90

The researcher was interested in determining if there was a trend and how the mean and variance changed over time.

- Create a time series plot of the data.



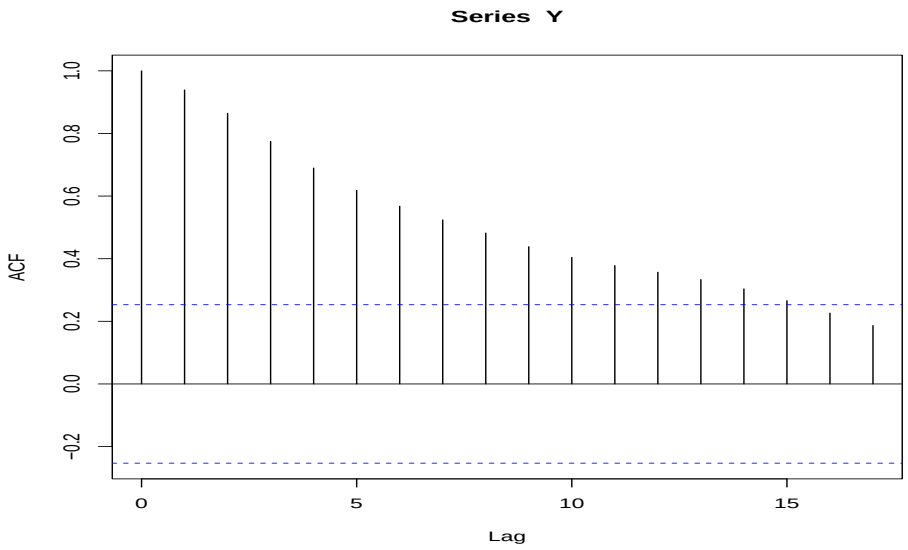
There is a downward trend in the bond yields which can be modeled as

$$Y_i = 9.467 - 0.01388 * M_i + 0.00005455 * M_i^2 - .00001018 * M_i^3$$

where Y_i is the bond yield for Month M_i, $\quad i = 1, 2, \ldots, 60$.

- To evaluate if there is correlation in the data, calculate the values of ρ_k, the autocorrelation coefficients. The lag k autocorrelations are given below. Based on these correlations and the plot it would appear that the adjacent monthly sales have a strong positive correlation. There appears to be a very slow decline in the autocorrelations with a pattern such as $\rho_k = (\rho_1)^k = (.939)^k$ for $k = 1, 2, \cdots, 17$, as would be seen in an AR(1) model. This would indicate that the monthly average yields of the AAA bonds are strongly correlated.

i	0	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}_i$	1.000	0.939	0.864	0.775	0.690	0.618	0.568	0.524	0.482	0.438	0.404
i	11	12	13	14	15	16	17				
$\hat{\rho}_i$	0.378	0.357	0.333	0.304	0.266	0.226	.187				



- The mean and standard deviations of the monthly sales over the five years are given in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Mean	8.752	8.738	8.806	8.780	8.722	8.584	8.482	8.384	8.352	8.332	8.340	8.240
St.Dev.	0.690	0.760	0.871	0.927	0.889	0.808	0.857	1.007	1.119	1.085	0.907	0.872

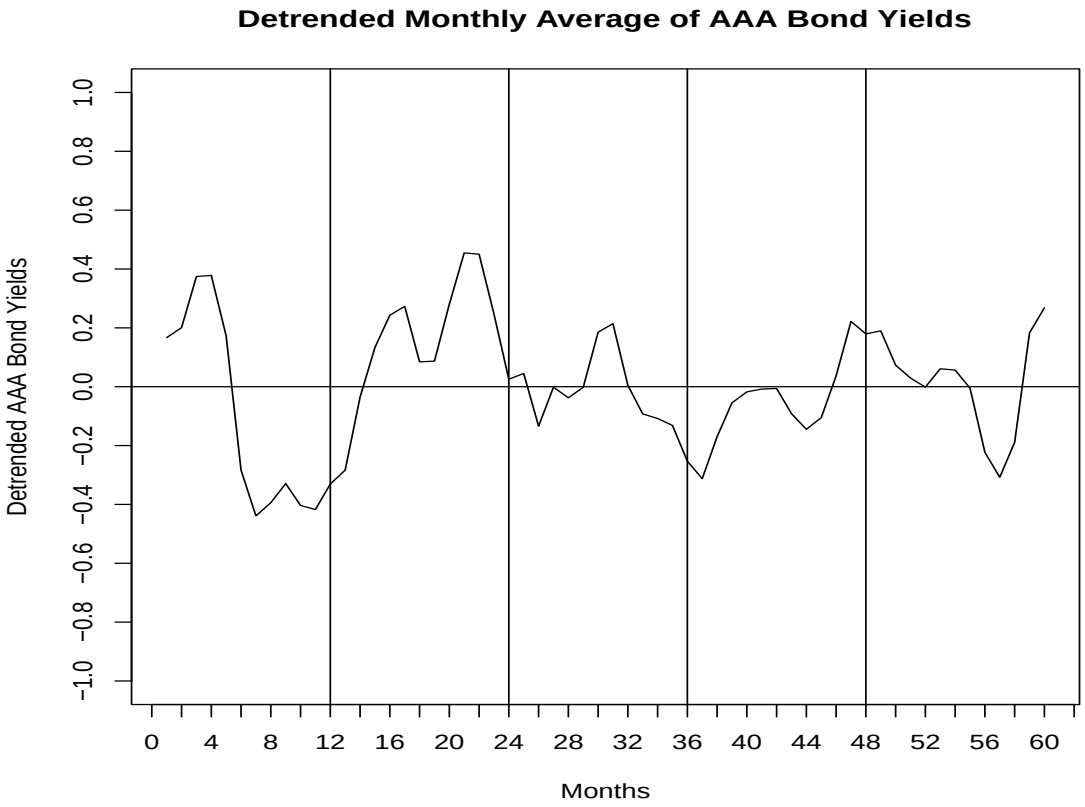
The sales appear non-stationary with an overall decline in yields along with a somewhat cyclic behavior over the five years. However, the monthly means and standard deviations over the five years are somewhat stable with a pattern of higher values for January through May and then lower values through the remaining months.

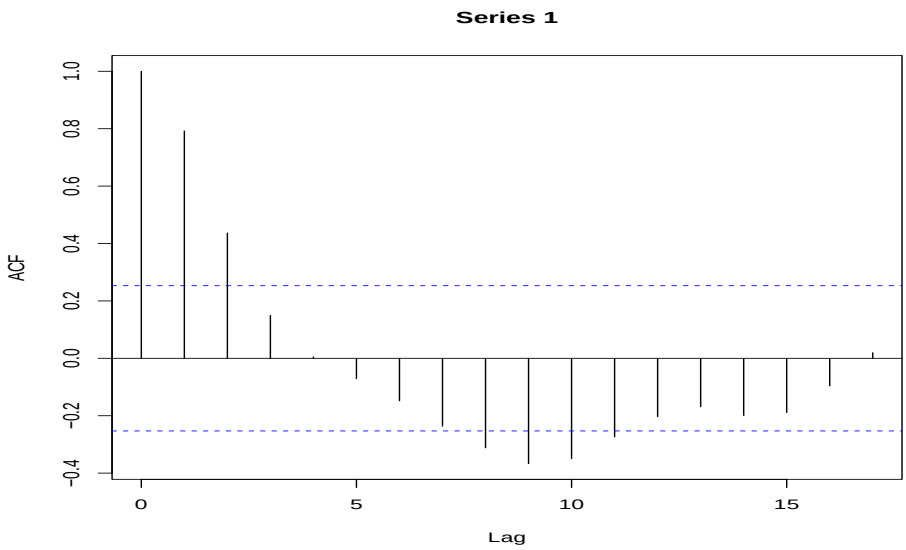
The analysis of the monthly yields is somewhat misleading if the trend is not taken into account.

If the data is detrended by defining

$$Z_i = Y_i - (9.467 - 0.01388 * M_i + 0.00005455 * M_i^2 - .00001018 * M_i^3)$$

The plot of Z_i and the autocorrelation function for the Z_i are given below:





i	0	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}_i$	1.000	0.792	0.437	0.150	0.005	-0.071	-0.148	-0.237	-0.312	-0.367	-0.350
i	11	12	13	14	15	16	17				
$\hat{\rho}_i$	-0.274	-0.204	-0.169	-0.199	-0.189	-0.096	0.020				

Not surprising, that there is still a strong correlation in the detrended data with $\rho_1 = 0.792$ and significant values for $\rho_1, \rho_2, \rho_8, \rho_9, \rho_{10}, \rho_{11}$.

- The mean and standard deviations of the detrended monthly sales over the five years are given in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Mean	-0.039	-.013	0.096	0.113	0.099	0.007	-0.047	-0.095	-0.076	-0.043	-0.021	-0.022
St.Dev.	0.243	0.152	0.170	0.187	0.121	0.177	0.247	0.255	0.316	0.318	0.289	0.263

The detrended monthly sales now appear to be relatively stationary over the five years. The regression line relating Z_i to M_i is $Z_i = .00024 + .000005 * M_i$ that is essentially correlated noise about a horizontal line through 0. Also, the monthly means and standard deviations over the five years are relatively stable considering that the values are based on only 5 data values each.

The difficult phase of the analysis of time series data is the modeling of the correlation. There are many different types of models autoregressive (AR), moving averages (MA), a combination of AR and MA (ARMA), and many more. I refer you to STAT 626 which is taught every summer.

STARTED on 9/29/21 (midway through lecture)

NOT Missing but Incomplete

Partial data

HANDOUT #7: CENSORED DATA

where you have partial information about some of your data points.

1. Type I Censoring - Fixed Censoring Time
2. Type II Censoring - Fixed Number of Observed Failures
3. Random Censoring
4. Many other Types of Censoring
5. Form of Censored Data
6. Parametric Estimation When Data is Censored
7. Distribution-Free Estimation When Data is Censored
8. Example - Using SAS and R

- Usually time to event data

- Ex: when you only know the event was at least a particular #.

Supplemental Reading

- Problem 4.51, Problem 6.34, Problem 15.10 in Tamhane/Dunlop book

CENSORED DATA

In some situations, the observations from a population or the outcomes from a process are not a complete set of data. Some of the experimental units on whom we planned to make observations or take measurements are removed from the experiment (study) prior to the end of the experiment. There may be partial information about these experimental units but not complete information.

Suppose we have n randomly selected experimental units from a population whose times to occurrence of an event are to be observed:

- Time to failure of an electrical device
- Time to occurrence of a tumor in a laboratory animal injected with a toxic chemical
- Time until patient's blood pressure reaches a specified level after receiving a treatment.
- Amount of stress at which alloy specimen fractures

Several forms of censoring will now be defined.

Suppose n units are placed on test and we want to measure the variable T_i on unit $i = 1, \ldots, n$.

1. **No Censoring - Complete Data Set**

   If we observe $T_1, \ldots, T_n$ on all n units and hence we have a complete data set. Standard methods of estimation are then used to estimate population parameters and make inferences about the population.

2. **Right Censoring - Incomplete Data Set**

   If we observe the value of T_i for $i = 1, \cdots, m$ but only know that the remaining $n - m$ units have values of T_i greater than their recorded values then we have right censored data. We only have a lower bound on the value of T_i.

- Type I and Type II censoring are special cases of Right Censoring.

Ex: Someone leaves hospital before they die. Don't know true until death, but know that it is at least up until they left the hospital

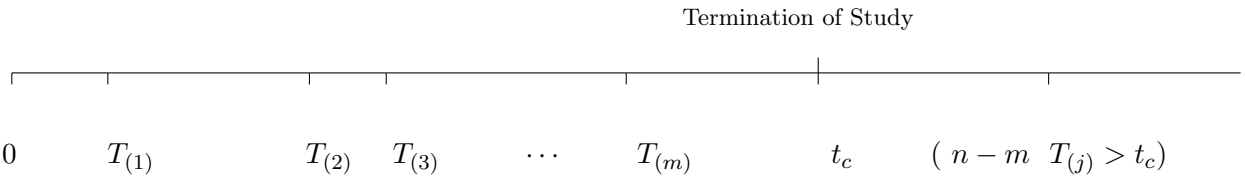
Type I Censoring - Stop Experiment at Time t_c

(The textbook refers to this type of censoring as Type II, see pages 235-236 in textbook)

Suppose n units are placed on test and $T_1, T_2, \ldots, T_n$ are their times to failure. Suppose the reliability study terminates at a preselected time, t_c. All experimental units which have not failed before time t_c have **Type I** censoring.

The experiment has n units on test and the experiment is terminated at a preselected time t_c. Suppose m units have failed before the experiment terminates. Then the ordered times to failure will satisfy:

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)} \leq t_c < T_{(m+1)} \leq \cdots \leq T_{(n)}$$



Because the experiment is terminated at time t_c, we will only observe the times for m of the units. For the remaining $n - m$ units, we will only know that their times satisfy $T_j > t_c$.

In Type I censoring, the time to termination of the experiment, t_c is fixed but the number of observed times, m is random. Type II fixes this and.

A problem arises if t_c is too small in that nearly all the data values will be censored.

Type II Censoring - Stop Experiment When mth Unit Fails

(The textbook refers to this type of censoring as Type I)

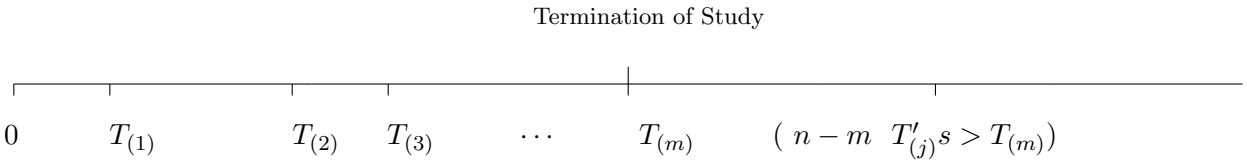
An alternatively experimental design has the experiment terminate when the mth (preselected) unit fails. This is referred to as **Type II censoring**.

The experiment has n units on test and the experiment is terminated when m units have failed.

The ordered times to failure will satisfy:

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)} < T_{(m+1)} \leq \cdots \leq T_{(n)}$$

With the times: $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)}$ observed but the times for the remaining $n - m$ units unobserved because the experiment was terminated before these units failed.



The remaining $n - m$ units will not have their failure times recorded. However, we will have the information that $T_{(j)} > T_{(m)}$ for $j = m + 1, \ldots, n$.

Type II censoring has the number of observed times, m fixed but the time to termination of the experiment, $T_{(m)}$ random.

If the device under test is very reliable, and m is too large, then it may take a long time to terminate the study, i.e., for m units to fail.

3. **Left Censoring - Incomplete Data Set**

If we observe the value of T_i for $i = 1, \cdots, m$ but only know that the remaining $n - m$ units have values of T_i less than their recorded values then we have left censored data. We only have an upper bound on the value of T_i.

- For example, we place n units on test and want to record their time to failure. The units are inspected at specified inspection times and some of the units fail prior to the first inspection time. The failure times of these units would be unknown except we know that they are less than the time of the first inspection, left censoring.

  Note, if there are units still functioning after the last inspection time then we would have right censoring.

- A second example would be if a measuring device is not sensitive enough to measure observations below a known threshold. For example, if we were measuring ultrasonic noise, the measuring device may not record a value for an sound having frequency less than 20KHz. Thus, any sound having less than 20KHz would not be recorded but we would know that the sound would have a value less than 20KHz.

  For example, monitoring communication between bats. We know there are sound waves being transmitted by observing the behaviour of the bats but the device fails to record any sounds.

4. **Random Censoring - Incomplete Data** (Anything not Left or Right censoring)

A third type of censoring is Random Censoring in which individual experimental units fail to have their time to the event recorded due to :

- Patient in study just stops returning to medical center
- Machine applying stress to specimens breaks down before specimen fractures
- Patient is removed from study due to side effects of drug
- Operator in Lab stops working
- Emergency budget cuts cause study to be reduced in size resulting in some patients dropping out of study
- Laboratory equipment fails while recording values for a given experimental unit

In this situation, we observe $min\{T_i, C_i\}$, where T_i is the time until event occurs for unit i and C_i is the time at which unit i leaves the experiment(study). Note that Type I censoring is equivalent to random censoring with $C_i \equiv t_c$ for all n units. In most applications, T and C are taken to be independent.

5. **Interval Censoring - Incomplete Data Set**

- Each of the censored units would have their measured value as being within an interval and no specific value recorded.
- In a failure time study with n units on test, the units are observed at specified inspection times. Therefore, the actual time to failure is unknown. The failure times are recorded as being between two inspection times.

6. **Other Types of Censoring or Incomplete Data Set**

- Arbitrary Censoring - Combinations of left and right censoring and interval censoring with overlapping intervals
- Truncated Data - Censoring occurs when there is a bound on all observations for which an exact value is not recorded: lower bound for observations censored on the right, upper bounds for observations censored on the left, and both upper and lower bounds for observations that are interval censored. Truncated data arises when even the existence of a potential observation would be unknown if its value were to lie in a certain range of values, either below a specified point τ_L or above a specified value, τ_U.

  Example from *Statistical Methods for Reliability Data*, **Meeker & Escobar** Ultrasonic inspection is used to detect flaws in titanium alloys. Ultrasonic signal amplitude is positively correlated with crack size. Titanium grain boundaries reflect signals as well as flaws. Thus below a specified threshold, τ_L it is impossible to be sure whether a signal is a flaw or grain boundary.

  In a lab test of the inspection process, specimens with flaws of known size are inspected. The signal's amplitude is measured only when it is above τ_L. All specimens in which a signal is not recorded would have left-censored observations.

  However, in a production inspection process, a flaw is not detected when the signal's amplitude is below τ_L. Thus the number of flaws that are present with signal amplitude below τ_L is unknown. The signal amplitudes recorded above τ_L are known as left-truncated observations, or observations from a left-truncated distribution.

Form of Censored Data

The data will have the following structure in each of four situations:

- Let $T_1, \ldots, T_n$ be the times to the event with $T_1, \ldots, T_n$ *iid* having cdf F.

  The values of $T_1, \ldots, T_n$ may or may not be observed depending on whether or not they have been censored.

- Let $Y_1, \ldots, Y_n$ be the observed data with $Y_1, \ldots, Y_n$ *iid* having cdf H.

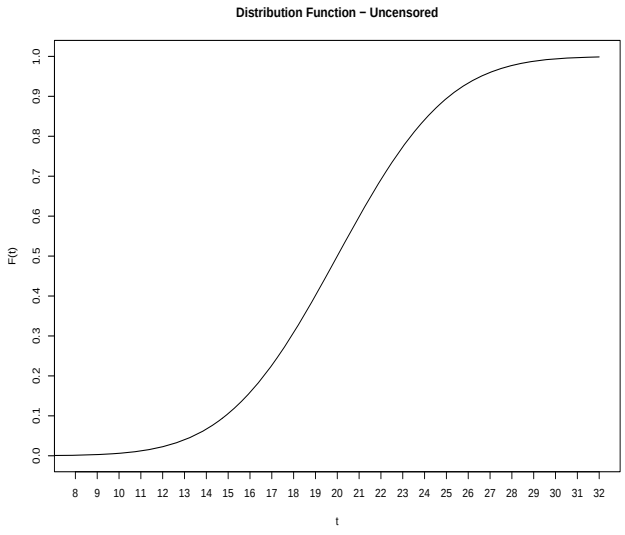
The form of Y_i and its cdf H will depend on the type of censoring in the experiment:

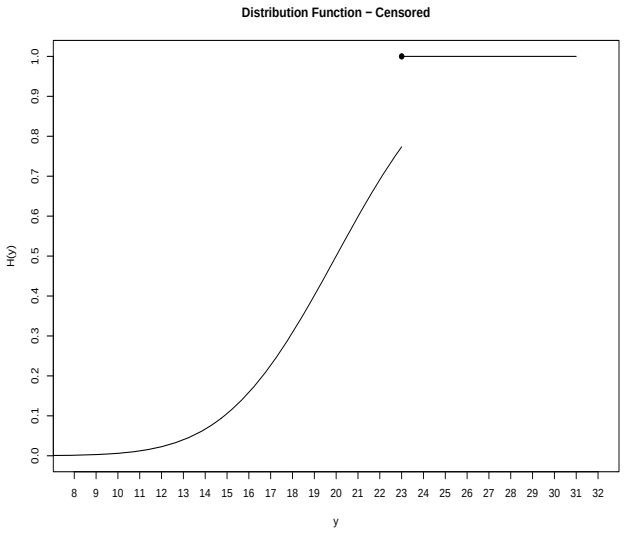
1. **No Censoring:** $Y_i = T_i$ for $i = 1, \ldots, n$ and $Y_1, \ldots, Y_n$ *iid* having cdf $H = F$.

2. **Type I Censoring:** For all $T_i > t_c$, data is censored, therefore we have

$$Y_i = \begin{cases} T_i & \text{if } T_i \leq t_c \\ t_c & \text{if } T_i > t_c \end{cases}$$

   The cdf of Y_i, H has a jump at t_c of height p where $p = P[Y_i = t_c] = P[T_i > t_c] = 1 - F(t_c) \neq 0$. This occurs because the distribution of Y_i is truncated at t_c.





3. **Type II Censoring:** The first m times to the event are observed and then the experiment is terminated. With $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$, we observe $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(m)}$ and then experiment is then terminated. Thus, the largest $n-m$ times to event are not observed. Therefore we have

$$Y_{(i)} = \begin{cases} T_{(i)} & \text{for } i = 1, \ldots, m \\ T_{(m)} & \text{for } i = m+1, \ldots, n \end{cases}$$

Using the properties of the order statistics, the joint pdf of the Y_i is

$$f(y_1, y_2, \ldots, y_n) = \binom{n}{m} m! f(y_1) f(y_2) \cdots f(y_m) [S(y_m)]^{n-m}$$

where $S(y) = 1 - F(y)$ is the survival function.

4. **Random Censoring:** Let $C_1, \ldots, C_n$ be the times at which the n units were censored with $C_i's$ iid with cdf $G(\cdot)$. Let $T_1, \ldots, T_n$ be the times to the occurrence of the event for the n units with $T_i's$ iid with cdf $F(\cdot)$. We observe $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ where $Y_i = min(T_i, C_i)$ and

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if Unit i is NOT censored} \\ 0 & \text{if Unit i IS censored} \end{cases}$$

With random censoring $Y_1, \ldots, Y_n$ are iid with cdf $H(\cdot)$.

The joint pdf of (Y, δ) is obtained using the independence of T and C:

If $\delta = 1$ (uncensored observation), then $T \leq C$ and $Y = T$, which yields

$$h(y, \delta) = f(y) P[C \geq y] = f(y)[1 - G(y)]$$

If $\delta = 0$ (censored observation), then $T \geq C$ and $Y = C$, which yields

$$h(y, \delta) = g(y) P[T \geq y] = g(y)[1 - F(y)] = g(y) S(y)$$

Using δ, we can combine the two situations into a single equation:

$$h(y, \delta) = [f(y)]^{\delta} [1 - G(y)]^{\delta} [g(y)]^{1-\delta} [S(y)]^{1-\delta}$$

For data having Type I, Type II, or Random Censoring our goal is to estimate the cdf F and parameters associated with F using the observed data $Y_1, \ldots, Y_n$.

STOP Wednesday 9/29/21

START Friday 10/1/21

Parametric Estimation

Suppose we know the family of distributions for which F is a member. For example, suppose the family is a logistic family of distributions with parameters $(\theta_1,\ \theta_2)$, that is, the family of distributions has pdf,

$$f(t;\theta_1,\ \theta_2) = \frac{1}{\theta_2}\frac{e^{-(t-\theta_1)/\theta_2}}{[1+e^{-(t-\theta_1)/\theta_2}]^2}$$

We want to estimate the parameters associated with F, θ_1 and θ_2. In order to use MLE techniques we need to specify the likelihood function for the data. The form of the likelihood will depend on the type of censoring:

1. **No Censoring:**

   Let $f(\cdot,\boldsymbol{\theta})$ be the pdf associated with $F(\cdot,\boldsymbol{\theta})$. The likelihood function for the observed data $Y_i's$ is given by

   $$L(\boldsymbol{\theta}) = f(Y_1, Y_2, \ldots, Y_n;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(T_i;\boldsymbol{\theta})$$

   Select the value of $\boldsymbol{\theta}$ to maximize $L(\boldsymbol{\theta})$.

2. **Random Censoring**

   Assume T_i and C_i are independent with $Y_i = min(T_i, C_i)$ and F the cdf of $T_i's$, H the cdf of $Y_i's$, and G the cdf of $C_i's$.

   (2a.) If Y has a discrete distribution, then the joint pdf for the pair (Y_i, δ_i) is given by

   $$P[Y_i = y, \delta = 1] = P[Y_i = y, C_i \geq T_i] = P[T_i = y, C_i \geq y] = P[T_i = y]P[C_i \geq y] = f(y)[1-G(y)]$$

   $$P[Y_i = y, \delta = 0] = P[Y_i = y, T_i \geq C_i] = P[C_i = y, T_i \geq y] = P[C_i = y]P[T_i \geq y] = g(y)[1-F(y)]$$

   (2b.) For continuous Y, the joint pdf for the pair (Y_i, δ_i) is given by

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = \begin{cases} f(Y_i;\boldsymbol{\theta})P[C_i \geq T_i] & \text{if } \delta_i = 1 \text{ (uncensored)} \\ g(Y_i)P[C_i < T_i] & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = \begin{cases} f(Y_i;\boldsymbol{\theta})[1-G(Y_i)] & \text{if } \delta_i = 1 \text{ (uncensored)} \\ g(Y_i)[1-F(Y_i;\boldsymbol{\theta})] & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

   $$L((Y_i,\delta_i);\boldsymbol{\theta}) = [f(Y_i;\boldsymbol{\theta})]^{\delta_i}[S(Y_i;\boldsymbol{\theta})]^{(1-\delta_i)}[1-G(Y_i)]^{\delta_i}[g(Y_i)]^{(1-\delta_i)},$$

   where $S(Y_i;\boldsymbol{\theta}) = 1 - F(Y_i;\boldsymbol{\theta})$ is the survival function for T_i.

   In the likelihood for the full sample we will designate terms involving $[1-G(Y_i)]$ and $g(Y_i)$ as $K(Y_i,\delta_i)$ because they do not involve the unknown parameters $\boldsymbol{\theta}$. Hence, the likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L(Y_i, \delta_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} [f(Y_i; \boldsymbol{\theta})]^{\delta_i} [S(Y_i; \boldsymbol{\theta})]^{(1-\delta_i)} \prod_{i=1}^{n} K(Y_i, \delta_i)$$

which yields

$$L(\boldsymbol{\theta}) = \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \prod_{i \epsilon C} S(C_i; \boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right]$$

where U is the set of indices for the uncensored units and C is the set of indices for the censored units.

3. **Type I Censoring**

The likelihood is obtained from the Random censoring case with $C_i \equiv t_c$ for all $i \epsilon C$:

$$L(\boldsymbol{\theta}) = \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \prod_{i \epsilon C} S(C_i; \boldsymbol{\theta}) \prod_{i=1}^{n} K(Y_i, \delta_i) = [S(t_c; \boldsymbol{\theta})]^{n-n_U} \prod_{i \epsilon U} f(T_i; \boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right]$$

4. **Type II Censoring**

Observe $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(m)}$ with remaining $n - m$ units having $Y_i \equiv T_{(m)}$. Therefore, the likelihood is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \binom{n}{m} \left[m! \prod_{i=1}^{m} f(Y_{(i)}; \boldsymbol{\theta}) \right] [S(Y_{(m)}, \boldsymbol{\theta})]^{n-m} \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right] \\ &= \frac{n!}{(n-m)!} \left[\prod_{i=1}^{m} f(T_{(i)}; \boldsymbol{\theta}) \right] [S(T_{(m)}, \boldsymbol{\theta})]^{n-m} \left[\prod_{i=1}^{n} K(Y_i, \delta_i) \right] \end{aligned}$$

Once we have established the likelihood function for the data, we then select $\hat{\boldsymbol{\theta}}$ to maximize $L(\boldsymbol{\theta})$ the likelihood function.

Example Let $T_1, \ldots, T_n$ be iid exponential: $F(t;\beta) = 1 - e^{-t/\beta}$

$$\Rightarrow f(t;\beta) = \frac{1}{\beta}e^{-t/\beta} \quad \text{and} \quad S(t;\beta) = 1 - F(t;\beta) = e^{-t/\beta}$$

Suppose we have **random censoring** and let n_U be the number of uncensored observations. The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i\epsilon U} f(T_i;\boldsymbol{\theta}) \prod_{i\epsilon C} S(C_i;\boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= \prod_{i\epsilon U} \frac{1}{\beta}e^{-T_i/\beta} \prod_{i\epsilon C} e^{-C_i/\beta} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= (\beta)^{-n_U} e^{-\frac{1}{\beta}\left(\sum_{i\epsilon U} T_i + \sum_{i\epsilon C} C_i\right)} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\ &= (\beta)^{-n_U} e^{-\frac{1}{\beta}\sum_{i=1}^{n} Y_i} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \end{aligned}$$

Thus, the log-likelihood is given by

$$l(\beta) = log(L(\beta)) = -n_U log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} Y_i + log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \quad \Rightarrow$$

$$\frac{d}{d\beta}l(\beta) = \frac{-n_U}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} Y_i + 0$$

$$\frac{d}{d\beta}l(\beta)|_{\beta=\hat{\beta}} = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{1}{n_U}\sum_{i=1}^{n} Y_i \quad \Rightarrow$$

$$\hat{\beta} = \frac{\text{Total Time n Units were Operating}}{\text{Total Number of Units Which reached Event}}$$

The same estimator is obtained for Type I censoring.

For Type II censoring, the estimator becomes

$$\hat{\beta} = \frac{1}{n_U}\sum_{i=1}^{n} Y_i = \frac{1}{n_U}\left[\sum_{i=1}^{n_U} T_{(i)} + (n - n_U)T_{(n_U)}\right]$$

$\hat{\beta}$ is referred to as the Winsorized Mean - replace all extreme values with the largest(smallest) non-extreme value.

Note that the naive estimator, $\hat{\beta}^*$ would satisfy

$$\hat{\beta}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \frac{1}{n}\sum_{i\epsilon U} T_i + \frac{1}{n}\sum_{i\epsilon C} C_i \;\leq\; \frac{1}{n}\sum_{i=1}^{n} T_i \;=\; \hat{\beta}, \text{ if no censoring occurred}$$

Thus, $\hat{\beta}^*$ would on the average underestimate β, because $\hat{\beta}$ is an unbiased estimator of β which would imply

$$E\left[\hat{\beta}^*\right] \;\leq\; E\left[\hat{\beta}\right] \;=\; \beta$$

Example Let $T_1, \ldots, T_n$ be iid with a Weibull distribution: $F(t;\gamma,\alpha) = 1 - e^{-(t/\alpha)^\gamma}$

$$\Rightarrow f(t;\gamma,\alpha) = \frac{\gamma}{\alpha}(t/\alpha)^{\gamma-1}e^{-(t/\alpha)^\gamma} \quad \text{and} \quad S(t;\gamma,\alpha) = 1 - F(t;\gamma,\alpha) = e^{-(t/\alpha)^\gamma}$$

Suppose we have **Type I censoring** and let $m = n_U$ be the number of uncensored observations. The likelihood function is given by

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i\epsilon U} f(T_i;\boldsymbol{\theta}) \prod_{i\epsilon C} S(C_i;\boldsymbol{\theta}) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \prod_{i\epsilon U} \frac{\gamma}{\alpha}(T_i/\alpha)^{\gamma-1}e^{-(T_i/\alpha)^\gamma} \prod_{i\epsilon C}\left[e^{-(t_c/\alpha)^\gamma}\right] \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \prod_{i\epsilon U} \frac{\gamma}{\alpha}(T_i/\alpha)^{\gamma-1}e^{-(T_i/\alpha)^\gamma} \left[e^{-(t_c/\alpha)^\gamma}\right]^{n-m} \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= \\
&= \left(\frac{\gamma}{\alpha}\right)^m \left(\frac{1}{\alpha^{\gamma-1}}\right)^m \left(\prod_{i\epsilon U} T_i^{\gamma-1}\right) \left(e^{-\frac{1}{\alpha^\gamma}\sum_{i\epsilon U} T_i^\gamma}\right) \left(e^{-(n-m)(t_c/\alpha)^\gamma}\right) \left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]
\end{aligned}$$

The log-likelihood function is given by

$$l(\gamma,\beta) = log(L(\gamma,\beta))$$

$$
\begin{aligned}
l(\gamma,\beta) &= log(L(\gamma,\beta)) \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(T_i) - \frac{1}{\alpha^\gamma}\sum_{i\epsilon U} T_i^\gamma - \frac{(n-m)}{\alpha^\gamma}t_c^\gamma \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i\epsilon U} T_i^\gamma + (n-m)t_c^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right] \\
&= mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]
\end{aligned}
$$

Taking derivatives of $l(\gamma,\alpha)$ wrt γ and α yield,

$$
\begin{aligned}
\frac{d}{d\alpha}l(\gamma,\alpha) &= \frac{d}{d\alpha}\left(mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]\right) \\
&= 0 - \frac{m\gamma}{\alpha} + 0 + \frac{\gamma}{\alpha^{\gamma+1}}\sum_{i=1}^{n} Y_i^\gamma + 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{d\gamma}l(\gamma,\alpha) &= \frac{d}{d\gamma}\left(mlog(\gamma) - m\gamma log(\alpha) + (\gamma-1)\sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\left(\sum_{i=1}^{n} Y_i^\gamma\right) \; + \; log\left[\prod_{i=1}^{n} K(Y_i,\delta_i)\right]\right) \\
&= \frac{m}{\gamma} - mlog(\alpha) + \sum_{i\epsilon U} log(Y_i) - \frac{1}{\alpha^\gamma}\sum_{i=1}^{n} log(Y_i)Y_i^\gamma + \frac{1}{\alpha^\gamma}log(\alpha)\left(\sum_{i=1}^{n} Y_i^\gamma\right) + 0
\end{aligned}
$$

Then setting the derivatives equal to 0 yield the following equations:

$$
\hat{\alpha} = \left(\frac{1}{m}\sum_{i=1}^{n} Y_i^{\hat{\gamma}}\right)^{1/\hat{\gamma}} \qquad (1)
$$

$$
0 = \frac{1}{\hat{\gamma}} + \frac{1}{m}\sum_{i\epsilon U} log(Y_i) - \frac{\sum_{i=1}^{n} Y_i^{\hat{\gamma}} log(Y_i)}{\sum_{i=1}^{n} Y_i^{\hat{\gamma}}} \qquad (2)
$$

Notice that equation (2) involves just $\hat{\gamma}$.

A numerical solution can be easily obtained and then the value of $\hat{\alpha}$ can be obtained from equation (1).

A numerical example will be considered next to illustrate the use of SAS and R in obtaining estimators of γ and α in the Weibull model when the data has censoring.

EXAMPLE The following example (slightly modified) is from *Statistical Analysis of Reliability Data* by M.J. Crowder, A.C. Kimber, R.L. Smith, and T.J. Sweeting. In an experiment to determine the strength of a braided cord after weathering, the strengths of 48 pieces of cord that had been weathered for a specified length of time were investigated. The company wanted to estimate the probability that the cord would have strength of at least 53, that is, estimate $S(53) = P[T > 53]$. Seven cords were damaged during the study which resulted in a decrease in their strength. Therefore, the study produced right censored strength values. The strengths of the remaining 41 cords were determined as shown below:

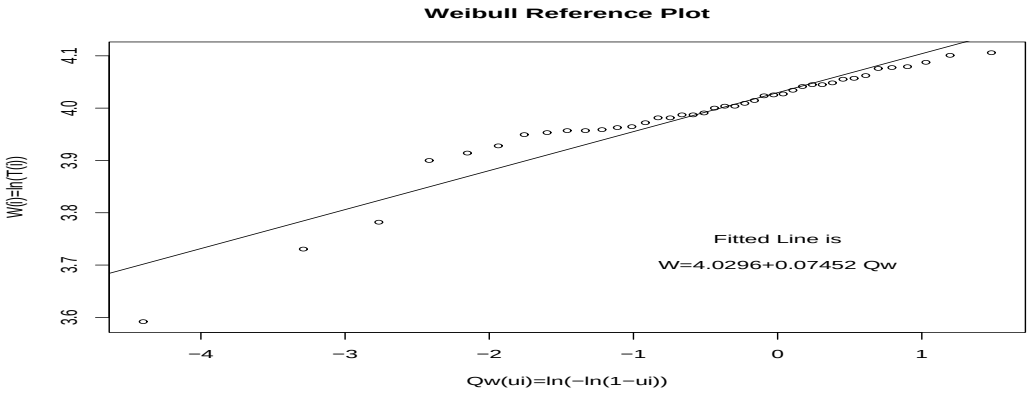
```
36.3  52.4  54.8  57.1  60.7  41.7  52.6  54.8  57.3
43.9  52.7  55.1  57.7  49.4  53.1  55.4  57.8
50.1  53.6  55.9  58.1  50.8  53.6  56.0  58.9
51.9  53.9  56.1  59.0  52.1  53.9  56.5  59.1
52.3  54.1  56.9  59.6  52.3  54.6  57.1  60.4
```

The 7 censored strength values from the damaged cords are given next:

```
26.8  29.6  33.4  35.0  40.0  41.9 42.5
```

The true strength values of the 7 cords, T_i are unobservable but we know $T_i > Y_i$, where Y_i are the observed values.

A Weibull reference distribution plot for the 41 uncensored strengths is displayed here:



From the plot, it would appear that the Weibull model is adequate.

The following SAS program will be used to obtain the MLE's of α and γ:

```
*weib_mle_censored.sas in Files/SASCode;

option ls=75 ps=55 nocenter nodate;
title 'Strength of Braided Cord';
data cords;
input S C @@;
label S  = 'Strength of Cord' C ='Censoring (1=Yes)';
cards;
36.3 0 52.4 0 54.8 0 57.1 0 60.7 0 41.7 0 52.6 0 54.8 0 57.3 0
43.9 0 52.7 0 55.1 0 57.7 0 49.4 0 53.1 0 55.4 0 57.8 0
50.1 0 53.6 0 55.9 0 58.1 0 50.8 0 53.6 0 56.0 0 58.9 0
51.9 0 53.9 0 56.1 0 59.0 0 52.1 0 53.9 0 56.5 0 59.1 0
52.3 0 54.1 0 56.9 0 59.6 0 52.3 0 54.6 0 57.1 0 60.4 0
26.8 1 29.6 1 33.4 1 35.0 1 40.0 1 41.9 1 42.5 1
run;
proc lifereg data=cords;
 model  S*C(1) = /dist=weibull covb;
run;

OUTPUT from SAS Program:

The LIFEREG Procedure in SAS

Number of Observations                48
Noncensored Values                    41
Right Censored Values                  7
Name of Distribution             Weibull

                          Standard   95% Confidence     Chi-
Parameter     DF Estimate    Error       Limits       Square Pr > ChiSq
Intercept      1   4.0257   0.0100   4.0061   4.0454  161324     <.0001
Scale          1   0.0615   0.0077   0.0481   0.0786
Weibull Scale  1  56.0223   0.5615  54.9325  57.1337
Weibull Shape  1  16.2591   2.0363  12.7202  20.7827
```

The estimators of γ and α are obtained by recalling, $F(t) = 1 - e^{-(t/\alpha)^\gamma}$.

Thus, α is the Weibull Scale parameter and γ is the Weibull Shape parameter.

$$\hat{\gamma} = \text{Weibull Shape} = 16.2591; \quad \hat{\alpha} = \text{Weibull Scale} = 56.0223$$

With $S(t) = e^{-(t/\alpha)^\gamma}$, the estimate of $S(53)$ would be given by:

$$\widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/56.0223)^{16.2591}} = .6664$$

Thus, we would estimate that approximately 67% of the cords would have strength of at least 53

From the uncensored data, we have 28 of the 41 values are greater than 53, that is, 68.3%. The two values are very close.

The estimated standard error of the estimator (see STAT 611) is obtained from the inverse of the information matrix: $\widehat{SE}(\hat{S}(53)) = 0.046$

We can use the following R code to obtain the estimates of α and γ for a Weibull model with censored data:

```
 R Code in eCampus under the name cords_censoredMLE.R
#Estimate parameters using censoring
 library(MASS)
 library(survival)
st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)
stcens = c(rep(1,41),rep(0,7))
cords = survreg(Surv(st, stcens) ~ 1, dist='weibull')
summary(cords)

#Estimate parameters ignoring censoring
fitdistr(st,"weibull",lower=c(0,0))

OUTPUT from R Code:
Call:
survreg(formula = Surv(st, stcens) ~ 1, dist = "weibull")
             Value Std. Error     z         p
(Intercept)  4.03     0.0100 401.7  0.00e+00
Log(scale)  -2.79     0.1252 -22.3 7.82e-110

Scale= 0.0615

# MLE's ignoring censoring:

     shape        scale
   9.4232284   54.5099326
 ( 1.1920947) ( 0.8643678)
```

From the above R function we must reinterpret the parameters to obtain:

$$\hat{\gamma} = 1/\text{Scale} = 1/.0615 = 16.2602, \qquad \hat{\alpha} = e^{\text{Intercept}} = e^{4.03} = 56.2609$$

and

$$\widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/56.2609)^{16.2602}} = .6847$$

Recall $\widehat{S}(53) = .6664$ using SAS code. The difference is possibly due to round-off error or precision differences in the calculations.

If the censoring of seven values had been ignored, the mle's based on 48 uncensored values are

$$\hat{\gamma} = 9.4232, \qquad \hat{\alpha} = 54.5099, \qquad \text{and} \qquad \widehat{S}(53) = e^{-(53/\hat{\alpha})^{\hat{\gamma}}} = e^{-(53/54.5099)^{9.4232}} = .4642$$

This displays a substantial change in the estimated proportion of cords that would have strength greater than 53 units, 46.4% ignoring censoring and 66.7% taking censoring into account.

The following example will further illustrate the importance of taking into account whether or not a data value is censored. Daily rainfall in millimeters was recorded over a 47 year period in Sydney Australia. The greatest amount of rain falling in a 24 hour period was recorded for each of the 47 years. There was a problem with the instruments that recorded the rainfall and any value greater than 2000 was considered to be a very inaccurate measurement of the true rainfall. Thus, we have right censored data and all values greater than 2000 will be set at 2000 prior to the data being analyzed. A Weibull model was fit to three sets of data: Data Set 1: original data values; Data Set 2: All values greater than 2000 are replaced by 2000; and Data Set 3: All values greater than 2000 are deleted. The following estimates of α and γ in the Weibull model along with the estimated quantiles for both tails of the distribution illustrate the importance of correctly taking into account censored values. The next pages contains the three data sets and a graph of the three versions of the Weibull pdf.

```
Original Data Set:  47 Maximum Daily Rainfall Values
 1468 3830  909 1781 2675  955 1565 1800  909 1397 2002 1717 1872 1849  701
  580  841  556 1331 2718 1359  719  994 1106  475  978 1227  584 1544 1737
 1188  808  846 1715 2543 1859 1372 1389  962  850  452  747 2649 1138 1334
  681 1564

Censored   Values: Rainfall larger than 2000 replaced with 2000 and designated as Censored
   1468 2000  909 1781 2000  955 1565 1800  909 1397 2000 1717 1872 1849  701
    580  841  556 1331 2000 1359  719  994 1106  475  978 1227  584 1544 1737
   1188  808  846 1715 2000 1859 1372 1389  962  850  452  747 2000 1138 1334
    681 1564

Delete Censored Rainfalls:   All maximum rainfall values greater than 2000  are Deleted
   1468  909 1781  955 1565 1800  909 1397 1717 1872 1849  701  580  841  556
   1331 1359  719  994 1106  475  978 1227  584 1544 1737 1188  808  846 1715
   1859 1372 1389  962  850  452  747 1138 1334  681 1564

MLEs for Two Parameters and Various Quantiles in a Weibull Model:

MLE for Weibull Model Parameters :

Original Data:              shape =  2.1103  ( 0.2246)       scale = 1537.2333  ( 112.1696)
Censored Data:              shape =  2.5966  ( 0.3303)       scale = 1472.6650  (  89.1911)
Censored values deleted:    shape =  3.0409  ( 0.3811)       scale = 1314.0159  (  71.2643)

Estimates of 9 Quantiles Using the 3 Methods:

                  Q(.01)   Q(.05)    Q(.10)    Q(.25)    Q(.5)     Q(.75)    Q(.9)     Q(.95)   Q(.99)
Original Data     173.80   376.25    529.20    851.80    1292.15   1794.57   2282.34   2585.46  3169.78
WithCensored      250.44   469.16    619.04    911.43    1278.81   1670.07   2030.49   2247.60  2651.75
DeleteCensored    289.48   494.77    626.91    872.30    1164.81   1463.02   1728.68   1884.94  2171.25
```

Impact of Censoring

Original Data Set: 47 Maximum Daily Rainfall Values

1468 3830 909 1781 2675 955 1565 1800 909 1397 2002 1717 1872 1849 701
580 841 556 1331 2718 1359 719 994 1106 475 978 1227 584 1544 1737
1188 808 846 1715 2543 1859 1372 1389 962 850 452 747 2649 1138 1334
681 1564

Censored Values: Rainfall larger than 2000 replaced with 2000 and designated as Censored

1468 2000 909 1781 2000 955 1565 1800 909 1397 2000 1717 1872 1849 701
580 841 556 1331 2000 1359 719 994 1106 475 978 1227 584 1544 1737
1188 808 846 1715 2000 1859 1372 1389 962 850 452 747 2000 1138 1334
681 1564

Delete Censored Rainfalls: All maximum rainfall values greater than 2000 are Deleted

1468 909 1781 955 1565 1800 909 1397 1717 1872 1849 701 580 841 556
1331 1359 719 994 1106 475 978 1227 584 1544 1737 1188 808 846 1715
1859 1372 1389 962 850 452 747 1138 1334 681 1564

MLEs for Two Parameters and Various Quantiles in a Weibull Model

MLE for Weibull Model :

Original Data: shape = 2.1103 (0.2246) scale = 1537.2333 (112.1696)

Censored Data: shape = 2.5966 (0.3303) scale = 1472.6650 (89.1911)

Censored values deleted: shape = 3.0409 (0.3811) scale = 1314.0159 (71.2643)

	Q(.01)	Q(.05)	Q(.10)	Q(.25)	Q(.5)	Q(.75)	Q(.9)	Q(.95)	Q(.99)
Original Data	173.80	376.25	529.20	851.80	1292.15	1794.57	2282.34	2585.46	3169.78
WithCensored	250.44	469.16	619.04	911.43	1278.81	1670.07	2030.49	2247.60	2651.75
DeleteCensored	289.48	494.77	626.91	872.30	1164.81	1463.02	1728.68	1884.94	2171.25

Distribution-Free Estimators

Don't know which model fits our data (in previous example weibull was good model of our distribution.

Suppose it is **unknown** to which family of distributions the cdf of T_i belongs. We want to use the data to estimate the survival function: $S(t) = Pr[T > t]$ and parameters associated with the survival function, such as, μ, $\tilde{\mu}$, and σ.

- **No Censoring**

1. Estimator of μ based on $\hat{S}(t)$

   With no censoring, $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t)$ = proportion of $T_i's$ greater than t.

   $$E[T] = \mu = \int_0^\infty t dF(t) = tF(t)|_0^\infty - \int_0^\infty F(t)dt = \int_0^\infty [1 - F(t)]dt = \int_0^\infty S(t)dt \Rightarrow$$

   $$\hat{\mu} = \int_0^\infty \hat{S}(t)dt = \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty I(T_i \geq t)dt = \frac{1}{n}\sum_{i=1}^{n}\int_0^{T_i} dt = \frac{1}{n}\sum_{i=1}^{n} T_i = \bar{T}$$

2. Estimator of median, M based on $\hat{S}(t)$

   Recall, $S(t) = 1 - F(t)$, so we have

   $$Q(u) \;=\; inf(t : F(t) \geq u) \;=\; inf(t : 1 - S(t) \geq u)\; inf(t : S(t) \leq 1 - u)$$

   Therefore, the median is defined by

   $$M = Q(.5) \;=\; inf(t : F(t) \geq .5) \;=\; inf(t : S(t) \leq .5)$$

   Thus, we can estimate the median by

   $$\hat{M} = inf(t : \hat{S}(t) \leq .5)$$

   In general the pth quantile is estimated by

   $$\widehat{Q}(p) = inf(t : \hat{S}(t) \leq 1 - p)$$

STOP 10/1/21

Kaplan-Meier Product Limit Estimator of S(t)

Suppose we have n homogeneous units placed on test.

Let $t_1 < t_2 < \cdots < t_k$ be the times at which units fail

Let d_j units fail at time t_j for $j = 1, 2, \ldots, k$

Let m_j be the number of units that are censored in the interval $[t_j,\ t_{j+1})$ for $j = 0, 1, \ldots, k$,
with $t_o = 0$ and $t_{k+1} = \infty$

Let $n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + \cdots + (m_k + d_k)$ be number of units at risk for failure just prior to time t_j → e.g. the # of units still around at time t_j

Note: $n_1 = n - m_0,\ n_2 = n - (d_1 + m_1)$

```
     mo        d1    m1              d2       m2       d3       m3          d4
|-------------|-------------------|---------------|------------------|-----
0             t1                  t2              t3                  t4
```

The Kaplan-Meier product limit estimator of $S(t) = P[T > t]$, where T is the time to failure, is given by

$$\text{For } t \in [t_i,\ t_{i+1}), \quad \widehat{S}(t) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j}$$

The estimator $\widehat{S}(t)$ is undefined for $t > t_k$

An heuristic justification of the estimator is as follows.

For $t \in [0,\ t_1)$, $S(t) = P[T > t] = P[\text{unit survives beyond time t}] \Rightarrow$

$\widehat{S}(t) = 1$ because there are no failures in $[0,\ t_1)$

For $t \in [t_1,\ t_2)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_1,\ t] \mid \text{unit survives in } [0,\ t_1)]\ P[\text{unit survives in } [0,\ t_1)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_1 - d_1}{n_1}\right) \cdot 1$$

For $t \in [t_2,\ t_3)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_2,\ t] \mid \text{survives in } [t_1,\ t_2)]\ P[\text{unit survives in } [t_1,\ t_2)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_2 - d_2}{n_2}\right)\left(\tfrac{n_1 - d_1}{n_1}\right)$$

For $t \in [t_3,\ t_4)$,

$S(t) = P[T > t] = P[\text{unit survives in } [t_3,\ t] \mid \text{survives in } [t_2,\ t_3)]\ P[\text{unit survives in } [t_2,\ t_3)] \Rightarrow$

$$\widehat{S}(t) = \left(\tfrac{n_3 - d_3}{n_3}\right)\left[\left(\tfrac{n_2 - d_2}{n_2}\right)\left(\tfrac{n_1 - d_1}{n_1}\right)\right]$$

From the Kaplan-Meier estimator we can then obtain estimators of the mean and median similarly as was done for uncensored data:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt \Rightarrow$$

$$\hat{\mu} = \quad \text{area under} \quad \hat{S}(\cdot) \quad \text{curve} \quad = \quad \sum_{i=1}^{k} [T_{(i)} - T_{(i-1)}] S(T_{(i-1)})$$

where $T_{(i)}$ are the observed times to the event for the k uncensored units, with $T_{(0)} = 0, S(0) = 1$.

The estimated standard error of $\widehat{S}(t)$ can be obtained from the Greenwood Formula:

Case 1: No Censoring

$\widehat{S}(t) = \frac{n_t}{n}$, where n_t is the number of n units still working at time t

$Var(\widehat{S}(t)) = \frac{S(t)[1-S(t)]}{n} \quad \Rightarrow \quad \widehat{SE}[\widehat{S}(t)] = \sqrt{\frac{\hat{S}(t)[1-\hat{S}(t)]}{n}}$

Case 2: Censoring with Tied Observations

$$\widehat{SE}[\widehat{S}(t)] = \widehat{S}(t) \sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}}$$

The courses STAT 645 - 646 will extensively cover the topic of censored data and survival analysis. The following books are good references for the analysis of censored data and survival analysis.

- *The Statistical Models and Methods for Lifetime Data* by Lawless
- *Survival Analysis Using S* by Tableman and Kim
- *Statistical Methods for Reliability Data* by Meeker and Escobar

The Kaplin-Meier PL Estimator will now be obtained for the data in the cord strength example using the following SAS code:

cord_censored_KM.sas

```
option ls=75 ps=55 nocenter nodate;
title 'Strength of Braided Cord'
data cords;
input S C @@;
label S  = 'Strength of Cord' C ='Censoring (1=Yes)';
cards;
36.3 0 52.4 0 54.8 0 57.1 0 60.7 0 41.7 0 52.6 0 54.8 0 57.3 0
43.9 0 52.7 0 55.1 0 57.7 0 49.4 0 53.1 0 55.4 0 57.8 0
50.1 0 53.6 0 55.9 0 58.1 0 50.8 0 53.6 0 56.0 0 58.9 0
51.9 0 53.9 0 56.1 0 59.0 0 52.1 0 53.9 0 56.5 0 59.1 0
52.3 0 54.1 0 56.9 0 59.6 0 52.3 0 54.6 0 57.1 0 60.4 0
26.8 1 29.6 1 33.4 1 35.0 1 40.0 1 41.9 1 42.5 1
run;
proc lifetest data=cords outsurv=a plots=(s);
 time S*C(1) ;
 proc print data=a;
 run;
```

OUTPUT FROM LIFETEST:

The LIFETEST Procedure

Product-Limit Survival Estimates

S	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	48
26.8000*	.	.	.	0	47
29.6000*	.	.	.	0	46
33.4000*	.	.	.	0	45
35.0000*	.	.	.	0	44
36.3000	0.9773	0.0227	0.0225	1	43
40.0000*	.	.	.	1	42
41.7000	0.9540	0.0460	0.0318	2	41
41.9000*	.	.	.	2	40
42.5000*	.	.	.	2	39
43.9000	0.9295	0.0705	0.0393	3	38
49.4000	0.9051	0.0949	0.0452	4	37
50.1000	0.8806	0.1194	0.0502	5	36
50.8000	0.8562	0.1438	0.0544	6	35
51.9000	0.8317	0.1683	0.0581	7	34
52.1000	0.8072	0.1928	0.0613	8	33
52.3000	.	.	.	9	32
52.3000	0.7583	0.2417	0.0667	10	31
52.4000	0.7338	0.2662	0.0688	11	30
52.6000	0.7094	0.2906	0.0708	12	29
52.7000	0.6849	0.3151	0.0724	13	28
53.1000	0.6605	0.3395	0.0739	14	27
53.6000	.	.	.	15	26
53.6000	0.6115	0.3885	0.0761	16	25
53.9000	.	.	.	17	24
53.9000	0.5626	0.4374	0.0774	18	23
54.1000	0.5382	0.4618	0.0778	19	22
54.6000	0.5137	0.4863	0.0781	20	21
54.8000	.	.	.	21	20
54.8000	0.4648	0.5352	0.0779	22	19
55.1000	0.4403	0.5597	0.0776	23	18
55.4000	0.4158	0.5842	0.0770	24	17
55.9000	0.3914	0.6086	0.0763	25	16
56.0000	0.3669	0.6331	0.0753	26	15
56.1000	0.3425	0.6575	0.0742	27	14
56.5000	0.3180	0.6820	0.0728	28	13
56.9000	0.2935	0.7065	0.0712	29	12
57.1000	.	.	.	30	11
57.1000	0.2446	0.7554	0.0672	31	10
57.3000	0.2202	0.7798	0.0648	32	9
57.7000	0.1957	0.8043	0.0620	33	8
57.8000	0.1712	0.8288	0.0589	34	7
58.1000	0.1468	0.8532	0.0553	35	6
58.9000	0.1223	0.8777	0.0512	36	5
59.0000	0.0978	0.9022	0.0465	37	4
59.1000	0.0734	0.9266	0.0408	38	3
59.6000	0.0489	0.9511	0.0337	39	2
60.4000	0.0245	0.9755	0.0242	40	1
60.7000	0	1.0000	0	41	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable S

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	57.1000	55.9000	58.9000
50	54.8000	53.6000	56.1000
25	52.4000	50.8000	53.9000

Mean	Standard Error
54.1824	0.7316

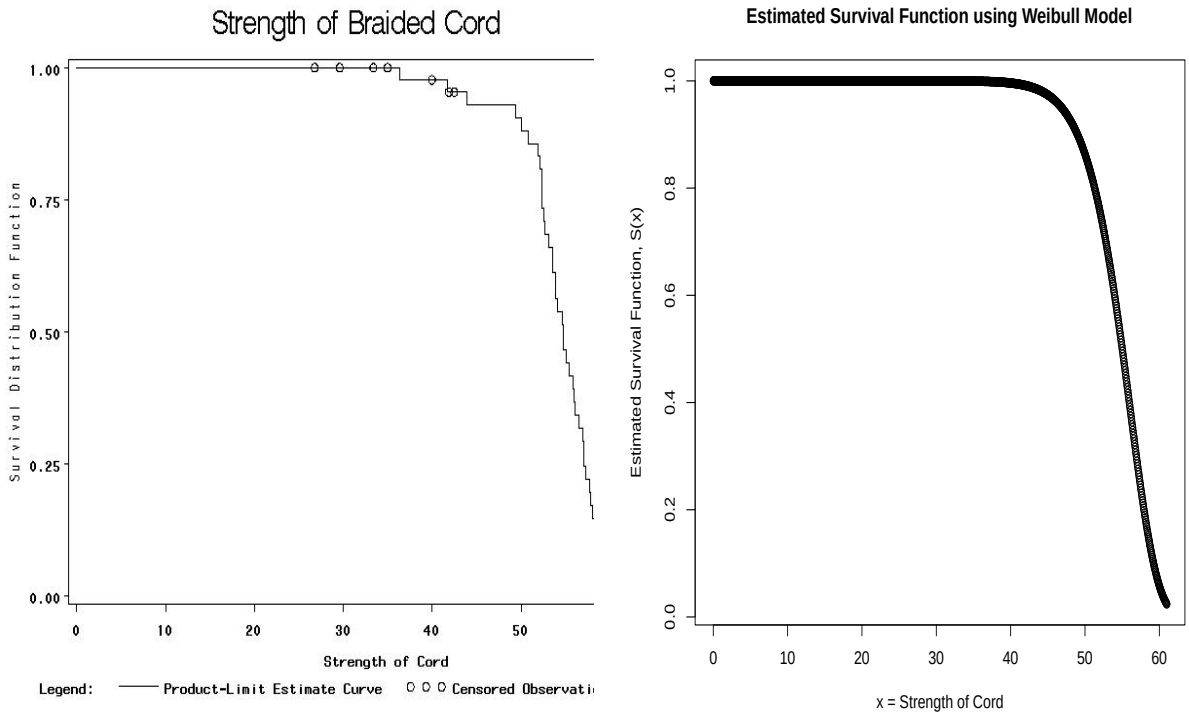
Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
48	41	7	14.58

Strength of Braided Cord

Obs	S	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
1	0.0	.	1.00000	1.00000	1.00000
2	26.8	0	1.00000	.	.
3	29.6	0	1.00000	.	.
4	33.4	0	1.00000	.	.
5	35.0	0	1.00000	.	.
6	36.3	1	0.97727	0.84941	0.99677
7	40.0	0	0.97727	.	.
8	41.7	1	0.95400	0.82832	0.98830
9	41.9	0	0.95400	.	.
10	42.5	0	0.95400	.	.
11	43.9	1	0.92954	0.79702	0.97675
12	49.4	1	0.90508	0.76630	0.96332
13	50.1	1	0.88062	0.73639	0.94855
14	50.8	1	0.85616	0.70725	0.93274
15	51.9	1	0.83170	0.67882	0.91607
16	52.1	1	0.80723	0.65102	0.89867
17	52.3	1	0.75831	0.59709	0.86206
18	52.4	1	0.73385	0.57087	0.84298
19	52.6	1	0.70939	0.54510	0.82342
20	52.7	1	0.68493	0.51975	0.80344
21	53.1	1	0.66046	0.49480	0.78305
22	53.6	1	0.61154	0.44606	0.74114
23	53.9	1	0.56262	0.39877	0.69781
24	54.1	1	0.53816	0.37565	0.67563
25	54.6	1	0.51369	0.35288	0.65313
26	54.8	1	0.46477	0.30837	0.60712
27	55.1	1	0.44031	0.28665	0.58362
28	55.4	1	0.41585	0.26528	0.55979
29	55.9	1	0.39139	0.24428	0.53562
30	56.0	1	0.36692	0.22366	0.51109
31	56.1	1	0.34246	0.20344	0.48620
32	56.5	1	0.31800	0.18363	0.46094
33	56.9	1	0.29354	0.16426	0.43528
34	57.1	1	0.24462	0.12696	0.38266
35	57.3	1	0.22015	0.10911	0.35563
36	57.7	1	0.19569	0.09188	0.32806
37	57.8	1	0.17123	0.07533	0.29988
38	58.1	1	0.14677	0.05959	0.27101
39	58.9	1	0.12231	0.04479	0.24133
40	59.0	1	0.09785	0.03115	0.21067
41	59.1	1	0.07338	0.01900	0.17881
42	59.6	1	0.04892	0.00889	0.14542
43	60.4	1	0.02446	0.00193	0.11054
44	60.7	1	0.00000	.	.

A plot of the Kaplan-Meier Estimator of S(t) and the MLE of S(t) based on a Weibull model are given below.



A comparison of the KM estimator to the values obtained from the Weibull model are displayed next:

$$\hat{S}_W(53.1) = e^{-(53.1/56.0223)^{16.2591}} = .6508 \qquad \hat{S}_{KM}(53.1) = .6605$$

$$\begin{aligned}
\hat{\mu}_{KM} &= \sum_{i=1}^{n_\nu} \left[T_{(i)} - T_{(i-1)}\right] S(T_{(i-1)} \\
&= (36.3-0)(1.0) + (41.7-36.3)(.97727) + (43.9-41.7)(.954) + \cdots (60.7-60.4)(.02446) \\
&= 54.1824
\end{aligned}$$

$$\hat{\mu}_W = \hat{\alpha}\Gamma\left(\frac{1+\hat{\gamma}}{\hat{\gamma}}\right) = \Gamma\left(\frac{17.2602}{16.2602}\right) = 54.4630$$

$$\hat{Q}_{KM}(.5) = \hat{S}_{K-M}(.5) = 54.8000$$

$$\hat{Q}_W(.5) = \hat{\alpha}(-log(.5))^{1/\hat{\gamma}} = (56.2609)(-log(.5))^{1/16.2602} = 55.0069$$

The estimates based on the Kaplan-Meier distribution free procedure and the MLE's based on a Weibull model are very close when the correct model is fit using the MLE of the parameters.

An analysis of the strength of the cords can also be analyzed using the following R code:

censoredcords_KM.R

```
 library(MASS)
 library(survival)


st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)

stcens = c(rep(1,41),rep(0,7))

Surv(st, stcens)

cords.surv <- survfit(Surv(st, stcens) ~ 1,conf.type="log-log")
summary(cords.surv)
print(cords.surv,print.rmean=TRUE)


plot(cords.surv,conf.int=FALSE,log=FALSE,
main="Kaplan-Meier Estimator of Survival Function",xlab="Strength of Cord",
ylab="Survival Function")
```

Estimator from R Code:

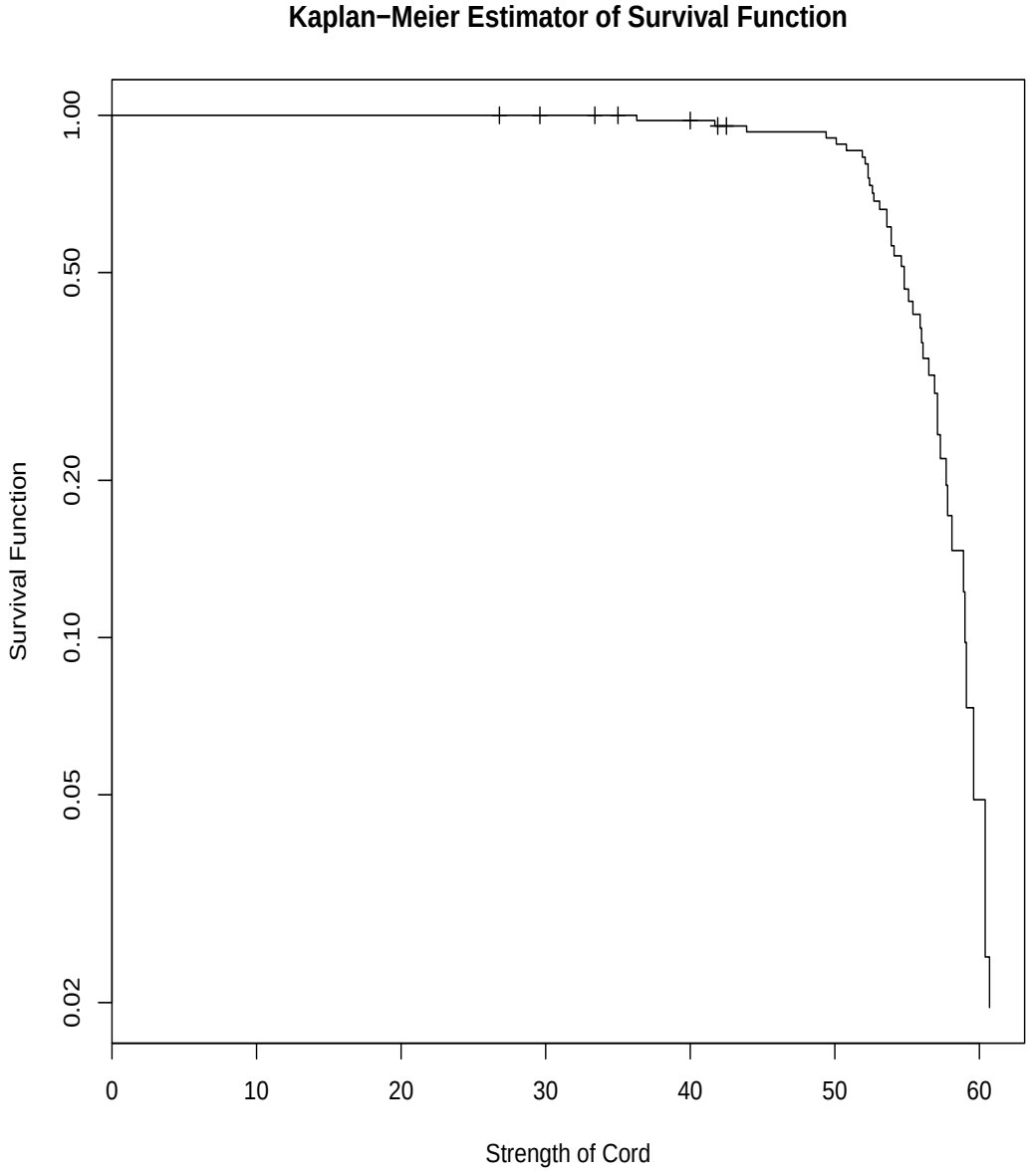
```
Call: survfit(formula = Surv(st, stcens) ~ 1, conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 36.3     44       1   0.9773  0.0225      0.84941        0.997
 41.7     42       1   0.9540  0.0318      0.82832        0.988
 43.9     39       1   0.9295  0.0393      0.79702        0.977
 49.4     38       1   0.9051  0.0452      0.76630        0.963
 50.1     37       1   0.8806  0.0502      0.73639        0.949
 50.8     36       1   0.8562  0.0544      0.70725        0.933
 51.9     35       1   0.8317  0.0581      0.67882        0.916
 52.1     34       1   0.8072  0.0613      0.65102        0.899
 52.3     33       2   0.7583  0.0667      0.59709        0.862
 52.4     31       1   0.7338  0.0688      0.57087        0.843
 52.6     30       1   0.7094  0.0708      0.54510        0.823
 52.7     29       1   0.6849  0.0724      0.51975        0.803
 53.1     28       1   0.6605  0.0739      0.49480        0.783
 53.6     27       2   0.6115  0.0761      0.44606        0.741
 53.9     25       2   0.5626  0.0774      0.39877        0.698
 54.1     23       1   0.5382  0.0778      0.37565        0.676
 54.6     22       1   0.5137  0.0781      0.35288        0.653
 54.8     21       2   0.4648  0.0779      0.30837        0.607
 55.1     19       1   0.4403  0.0776      0.28665        0.584
 55.4     18       1   0.4158  0.0770      0.26528        0.560
 55.9     17       1   0.3914  0.0763      0.24428        0.536
 56.0     16       1   0.3669  0.0753      0.22366        0.511
 56.1     15       1   0.3425  0.0742      0.20344        0.486
 56.5     14       1   0.3180  0.0728      0.18363        0.461
 56.9     13       1   0.2935  0.0712      0.16426        0.435
 57.1     12       2   0.2446  0.0672      0.12696        0.383
 57.3     10       1   0.2202  0.0648      0.10911        0.356
 57.7      9       1   0.1957  0.0620      0.09188        0.328
 57.8      8       1   0.1712  0.0589      0.07533        0.300
 58.1      7       1   0.1468  0.0553      0.05959        0.271
 58.9      6       1   0.1223  0.0512      0.04479        0.241
 59.0      5       1   0.0978  0.0465      0.03115        0.211
 59.1      4       1   0.0734  0.0408      0.01900        0.179
 59.6      3       1   0.0489  0.0337      0.00889        0.145
 60.4      2       1   0.0245  0.0242      0.00193        0.111
 60.7      1       1   0.0000     NaN           NA           NA

Call: survfit(formula = Surv(st, stcens) ~ 1, conf.type = "log-log")

records   n.max    n.start  events  *rmean  *se(rmean)  median  0.95LCL    0.95UCL
 48.000  48.000    48.000   41.000   54.182   0.723     54.800   53.100     56.100
* restricted mean with upper limit =  60.7
```

A graph of the survival is produced next:


Kaplan−Meier Estimator of Survival Function
Survival Function
1.00
0.50
0.20
0.10
0.05
0.02
0
10
20
30
40
50
60
Strength of Cord

Comparing Two Survival Curves from Censored Data

EXAMPLE from *The Statistical Analysis of Failure Time Data*

The following table gives the times from insult with the carcinogen DMBA to mortality from vaginal cancer in rats. Two groups were distinguished by a pretreatment regime. Four times to mortality were randomly censored and are denoted by an *. For these rats, we can see that their times to mortality exceed 216, 244, 204, and 344 days, respectively, but we do not know the times exactly. The random censoring occurred because the four rats died of causes unrelated to the application of the carcinogen and they were free of tumor at death, or they may simply not have developed tumor at the time of data analysis. There are two separate groups of survival times data in this example.

	Days to Vaginal Cancer Mortality in Rats									
Group 1	143	164	188	188	190	192	206	209	213	216
	220	227	230	234	246	265	304	216*	244*	
Group 2	142	156	163	198	205	232	232	233	233	233
	233	239	240	261	280	280	296	296	323	204*
	344*									

The following SAS program was used to obtain a Kaplan-Meier estimator for each of the two groups of rats.

```
 eCampus  -  kmest_rat.sas;


option ls=75 ps=55 nocenter nodate;
title 'Cancer Treatment-Estimated S(t)';
data cancer;
input T ST G @@;
LGT=log(T);
label ST = 'Censoring Indicator';
label T  = 'Time to Death';
label G  = 'Treatment Group';
cards;
143 1 1 164 1 1 188 1 1 188 1 1 190 1 1 192 1 1 206 1 1
209 1 1 213 1 1 216 1 1 220 1 1 227 1 1 230 1 1 234 1 1
246 1 1 265 1 1 304 1 1 216 0 1 244 0 1
142 1 2 156 1 2 163 1 2 198 1 2 205 1 2 232 1 2 232 1 2
233 1 2 233 1 2 233 1 2 233 1 2 239 1 2 240 1 2 261 1 2
280 1 2 280 1 2 296 1 2 296 1 2 323 1 2 204 0 2 344 0 2
run;
proc lifetest data=cancer outsurv=a plots=(s);
 time  T*ST(0);
 strata G;
run;
proc print data=a;
run;
```

OUTPUT FROM SAS:

Cancer Treatment-Estimated S(t)

The LIFETEST Procedure

Stratum 1: G = 1

Product-Limit Survival Estimates

T	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	19
143.000	0.9474	0.0526	0.0512	1	18
164.000	0.8947	0.1053	0.0704	2	17
188.000	.	.	.	3	16
188.000	0.7895	0.2105	0.0935	4	15
190.000	0.7368	0.2632	0.1010	5	14
192.000	0.6842	0.3158	0.1066	6	13
206.000	0.6316	0.3684	0.1107	7	12
209.000	0.5789	0.4211	0.1133	8	11
213.000	0.5263	0.4737	0.1145	9	10
216.000	0.4737	0.5263	0.1145	10	9
216.000*	.	.	.	10	8
220.000	0.4145	0.5855	0.1145	11	7
227.000	0.3553	0.6447	0.1124	12	6
230.000	0.2961	0.7039	0.1082	13	5
234.000	0.2368	0.7632	0.1015	14	4
244.000*	.	.	.	14	3
246.000	0.1579	0.8421	0.0934	15	2
265.000	0.0789	0.9211	0.0728	16	1
304.000	0	1.0000	0	17	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable T

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	234.000	216.000	265.000
50	216.000	192.000	234.000
25	190.000	188.000	216.000

Mean	Standard Error
218.757	9.403

Stratum 2: G = 2

Product-Limit Survival Estimates

T	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	21
142.000	0.9524	0.0476	0.0465	1	20
156.000	0.9048	0.0952	0.0641	2	19
163.000	0.8571	0.1429	0.0764	3	18
198.000	0.8095	0.1905	0.0857	4	17
204.000*	.	.	.	4	16
205.000	0.7589	0.2411	0.0941	5	15
232.000	.	.	.	6	14
232.000	0.6577	0.3423	0.1053	7	13
233.000	.	.	.	8	12
233.000	.	.	.	9	11
233.000	.	.	.	10	10
233.000	0.4554	0.5446	0.1114	11	9
239.000	0.4048	0.5952	0.1099	12	8
240.000	0.3542	0.6458	0.1072	13	7
261.000	0.3036	0.6964	0.1031	14	6
280.000	.	.	.	15	5
280.000	0.2024	0.7976	0.0902	16	4
296.000	.	.	.	17	3
296.000	0.1012	0.8988	0.0678	18	2
323.000	0.0506	0.9494	0.0493	19	1
344.000*	.	.	.	19	0

NOTE: The marked survival times are censored observations.

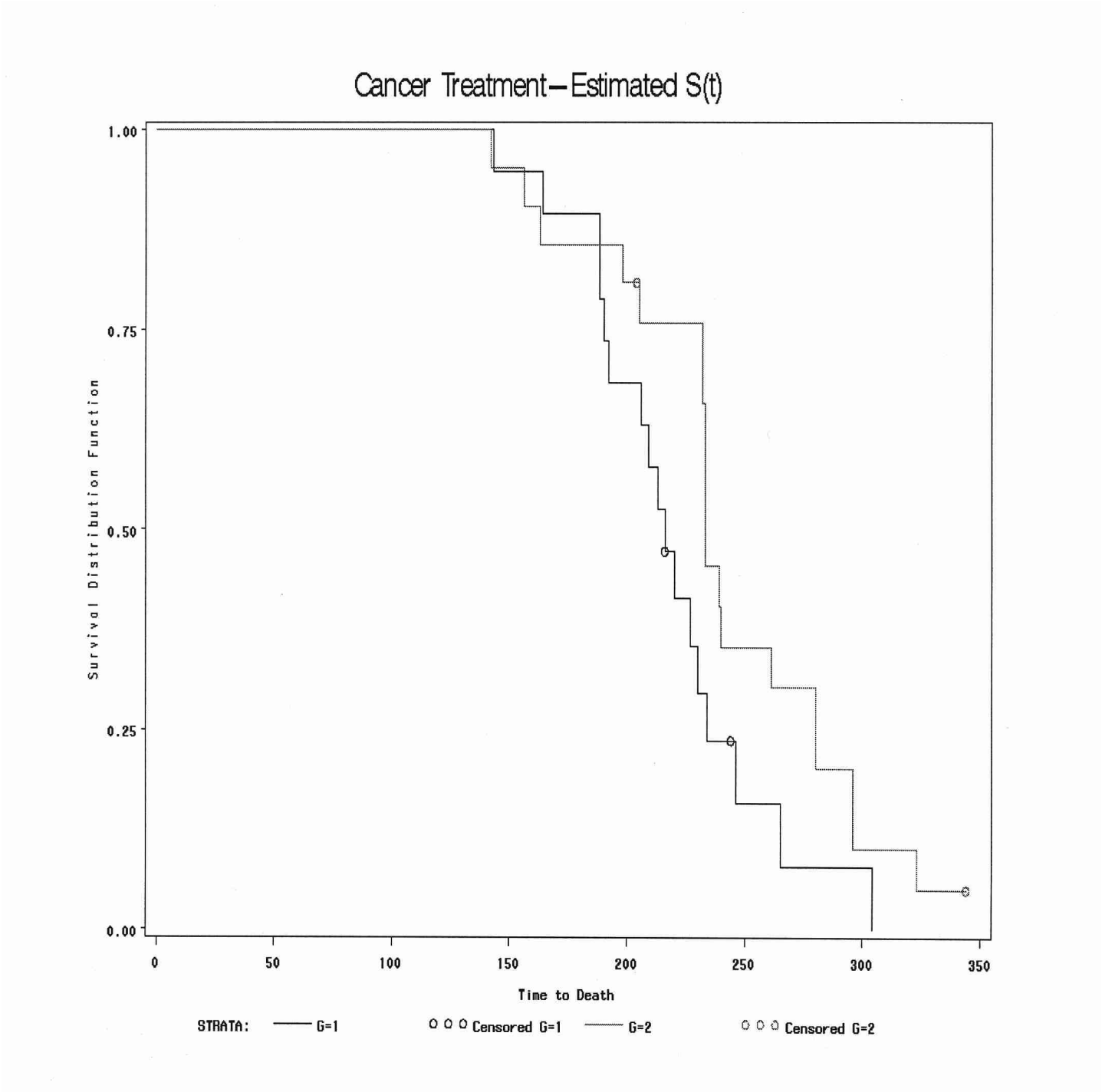
Summary Statistics for Time Variable T

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	280.000	233.000	296.000
50	233.000	232.000	280.000
25	232.000	163.000	233.000

Mean	Standard Error
240.795	11.206

A plot of the Kaplan-Meier Estimators of S(t) for the two groups is given below.



The following R code will produce similar results to what was achieved using the SAS code:

```
carcinogen_2G_KM.R

library(survival)

T = c( 143,164,188,188,190,192,206,209,213,216,
       220,227,230,234,246,265,304,216,244,
       142,156,163,198,205,232,232,233,233,233,
       233,239,240,261,280,280,296,296,323,204,
       344)

ST = c(rep(1,17),rep(0,2),rep(1,19),rep(0,2))

G = c(rep(1,19),rep(2,21))

out = cbind(T,ST,G)

Surv(T, ST)

carcin <- survfit(Surv(T, ST) ~ G)
summary(carcin)
print(carcin, print.rmean=TRUE,rmean="individual")

par(lab=c(15,20,4))
plot(carcin,ylab="Survival Function",xlab="Time to Death",mark.time=TRUE,
main="Cancer Treatment - Estimated S(t)",lty=1:2 )
legend(25,.8,c("Group 1","Group 2"),lty=1:2,lwd=2)
text(216,.4737,"+")
text(244,.2368,"+")
text(204,.8095,"+")
text(344,.0506,"+")
```

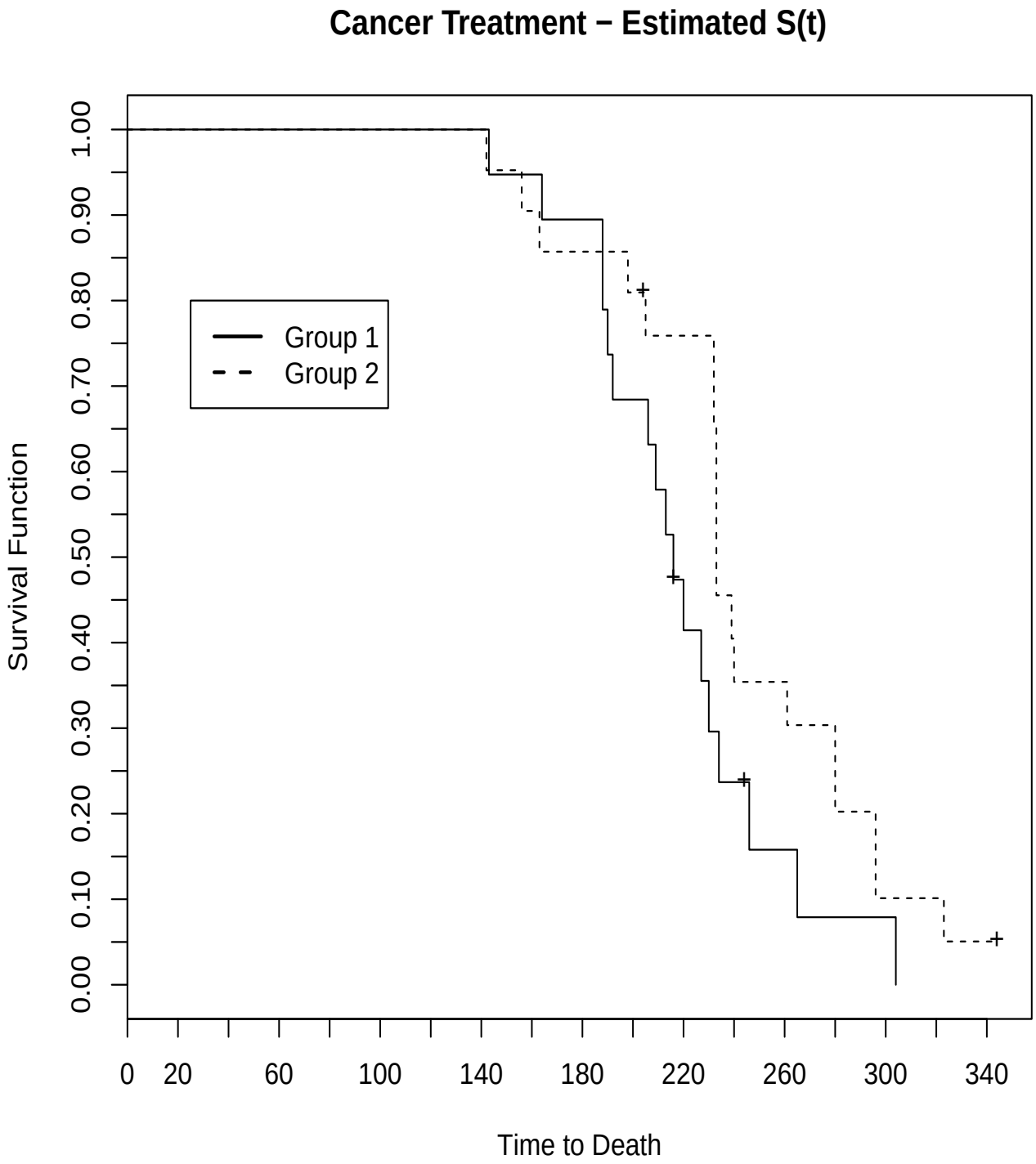
OUTPUT FROM R:

```
    records n.max n.start events *rmean *se(rmean) median 0.95LCL 0.95UCL
G=1      19    19      19     17    219       9.12    216     206     265
G=2      21    21      21     19    242      11.35    233     232     280
   * restricted mean with variable upper limit


Call: survfit(formula = Surv(T, ST) ~ G, conf.type = "log-log")

                G=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  143     19       1    0.947  0.0512      0.68119        0.992
  164     18       1    0.895  0.0704      0.64079        0.973
  188     17       2    0.789  0.0935      0.53191        0.915
  190     15       1    0.737  0.1010      0.47893        0.881
  192     14       1    0.684  0.1066      0.42794        0.844
  206     13       1    0.632  0.1107      0.37899        0.804
  209     12       1    0.579  0.1133      0.33208        0.763
  213     11       1    0.526  0.1145      0.28720        0.719
  216     10       1    0.474  0.1145      0.24438        0.673
  220      8       1    0.414  0.1145      0.19616        0.621
  227      7       1    0.355  0.1124      0.15191        0.566
  230      6       1    0.296  0.1082      0.11168        0.509
  234      5       1    0.237  0.1015      0.07578        0.447
  246      3       1    0.158  0.0934      0.03143        0.374
  265      2       1    0.079  0.0728      0.00567        0.288
  304      1       1    0.000     NaN           NA           NA

                G=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  142     21       1   0.9524  0.0465      0.70721        0.993
  156     20       1   0.9048  0.0641      0.67005        0.975
  163     19       1   0.8571  0.0764      0.61972        0.952
  198     18       1   0.8095  0.0857      0.56891        0.924
  205     16       1   0.7589  0.0941      0.51394        0.892
  232     15       2   0.6577  0.1053      0.41232        0.820
  233     13       4   0.4554  0.1114      0.23531        0.652
  239      9       1   0.4048  0.1099      0.19615        0.605
  240      8       1   0.3542  0.1072      0.15914        0.556
  261      7       1   0.3036  0.1031      0.12446        0.506
  280      6       2   0.2024  0.0902      0.06327        0.397
  296      4       2   0.1012  0.0678      0.01719        0.275
  323      2       1   0.0506  0.0493      0.00349        0.207
```


Cancer Treatment − Estimated S(t)
Survival Function
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Group 1
Group 2
0
20
60
100
140
180
220
260
300
340
Time to Death