# TABLE OF CONTENTS

- **Introduction** ........................................................................................................... 1
- **Course Structure** ................................................................................................. 2-3
- **Recommended Texts** ........................................................................................... 3-5
- **Course Evaluation/ Grading** .................................................................................. 6-7
- **Learning Objectives** ............................................................................................. 8-9
- **Course Schedule/ Outline** .................................................................................... 10
- **Course Notes/ Background Material:**
  - Lecture I .............................................................................................................. 11-24
  - Lecture II ............................................................................................................ 25-41
  - Lecture III ......................................................................................................... 42-49
  - Lecture IV .......................................................................................................... 50-76
- **Glossary** .............................................................................................................. 77-81
I. INTRODUCTION

This course will introduce you to the principles underlying the science of preventive medicine and clinical research methods, as embodied in the methodology of epidemiology and biostatistics. The goal is to give you the skills required to be critical readers of the medical literature. You will draw upon these skills as you move on to the Evidence Based Medicine Sessions in Year 2 (EPHEM II), and in your clinical years where you will learn more about the interpretation of the medical literature in the context of patient care. While these skills are broadly applicable to all of clinical medicine, we will be teaching them in the context of health promotion and preventive medicine, because the ‘population-based’ approach is directly relevant to prevention and public health; because prevention is, in general, less widely emphasized in the rest of the curriculum; and because the clinical issues in prevention should be more meaningful to first year students, since most of you have not yet taken care of patients.

You have no doubt heard elsewhere that physicians must be lifelong learners. In order to be lifelong learners, you must have the skills to read reports in medical journals and understand what they mean. As a practicing physician, you will often need to turn to published articles of original research to answer clinical questions. This requires the ability to critically appraise the methods and interpretation of studies. If this sounds easy, great! -- but recognize that it is a skill which many bright, hardworking, well-motivated physicians currently lack, as noted in a recent survey of medical residents (JAMA, 298(9):1010-1022, 2007). If you become a careful, critical, comprehending reader as a result of this course, it will have achieved its primary aim. It is expected that you will not determine the implications of a study simply by reading the last line of the abstract to see what the authors say it shows. Instead, you will critically analyze the methods and the results and form your own conclusions. This means you have to develop the belief that you are as good a judge of the import of a study -- from your own perspective, based on your experience, and within your own practice or community -- as the experts who wrote the paper. Some might call this arrogance. But the only way it’s arrogant to challenge an interpretation is if it’s unequivocally correct. And so, a crucial postulate: there are no unequivocally correct interpretations of clinical studies. This is not to say there is no truth -- simply that our attempts to uncover the truth are limited, and open to interpretation, and that what is correct in Minnesota may not be correct in the Bronx.

So as we develop our skills at critical assessment of the literature, you may find that you disagree with an interpretation or opinion of a speaker in this class. If you do, you should raise your hand and ask a question, or make your point. You are encouraged to [respectfully] challenge any statement made by your teachers or peers. Keep in mind that this course, even though it is given in a year dominated by basic science lectures, is different. While there are a series of facts you must know to succeed in this course, the learning of facts is not our primary goal. Rather, our goal is to help you to develop your critical reasoning skills, and thus, active participation in the discussions is necessary for achieving this goal.
II. COURSE STRUCTURE

Please take the time to familiarize yourself with the overall description of the course structure that follows.

PART I. BASIC PRINCIPLES: LECTURES Sessions 1-4 (March 5, 12, 19, April 16)

PART II. CASE DISCUSSION GROUPS (several sections-see group assignments on the course webpage): Sessions 5-10 (April 23, May 7, 9, 14, 21, 28).

REVIEW SESSION: May 29th at 9 am, Riklis Auditorium

FINAL EXAM: (short answers/essays, in class): Monday June 3, 9:30-12:00 pm, Robbins Auditorium.

PART I:
The lectures are designed to provide an intensive ‘mini-course’ on epidemiology and biostatistics. Lectures will cover the “Learning Objectives” which follow in this syllabus, and are designed to provide you with the basic tools required to evaluate the related case studies and to be an active participant in the case-discussion groups. Lectures are all given in Riklis Auditorium at the times noted on the schedule. The background material for these lectures is included in this syllabus. The lectures will provide some interactive problems/ and examples to build upon the material in the syllabus. For further reading, see the section below on recommended texts.

A self-study problem set will be posted on emed for each lecture. These will provide you with a review of the key points and practice answering questions based on the learning objectives.

PART II. CASE DISCUSSION GROUPS:
There will be a total of six case discussion sessions. These are the heart of the course, and provide you with the opportunity to apply the basic principles to the interpretation of actual research studies in the context of clinical scenarios. Active participation and discussion is the key to the success of these sessions. A portion of the time for each session will involve working in groups of 4-5 students to complete a hands-on problem set related to the day’s case. The remainder of the session will be a discussion of the problems and the focus readings in the context of the teaching objectives and assigned discussion questions. Group assignments for the case conferences are posted on the course webpage. Each session will be based on 1-2 focus journal articles related to an aspect of clinical research. The materials for each session will be available on the emed course page.

The expectations for each session will be clearly laid out in the weekly assignment on the course webpage. Each assignment will include:
Learning objectives for the session
Required focus article(s) for the case discussion.
Relevant sections of the course syllabus and recommended text for background reading
A brief quiz regarding the readings and background material. The quiz for each session will be available on emed and must be completed by 12pm the night prior to the scheduled session. These 6 quizzes plus the quiz after the last lecture will comprise 20% of your course grade. Responses will be used by your section leader to facilitate the discussion. These question sets will comprise 20% of your final grade.

The quizzes are designed to provide feedback once the answers have been submitted,

ADVANCE PREPARATION FOR, AND PARTICIPATION IN, THIS PORTION OF THE COURSE IS ABSOLUTELY MANDATORY. This cannot be overstated: we cannot successfully employ this teaching method if you are unprepared. We will abide by the following ground rules:

1. ATTENDANCE IS MANDATORY. As with all small group sessions at Einstein, attendance at the sessions is mandatory. Unexcused absence is grounds for failing the course, and will be reported to the office of student affairs. Absences must be excused by the course leader prior to the session, or in the case of an emergency, by the next day. A required make-up assignment will be provided for any excused session.

2. PREPARATION IS MANDATORY. Faculty will expect you to be prepared, and will elicit your contributions to discussions when you volunteer, but also sometimes when you do not.

3. COMPLETION OF THE QUIZZES PRIOR TO EACH SMALL GROUP SESSION IS MANDATORY. You will be able to use the quizzes as a self-study tool, and responses will be used by the small group leaders to facilitate the discussion, and to identify areas of confusion. Completion of these study questions will count toward 20% of your grade as noted above.

III. RECOMMENDED TEXTS.
This syllabus includes brief discussions of the major topics of the course. These are organized in relation to the topics of the 4 lectures. However, the syllabus covers this material in a cursory and concise manner, and in order to supplement the syllabus, I highly recommend the following text for the course (this is available in the Einstein bookstore):

This is a clearly written text that covers basic principles of epidemiologic research. Although some sections provide more detail than this course requires, feedback from students in later years has indicated that it is considered an excellent reference to be used throughout your medical education. Review questions in each chapter provide the opportunity to apply the principles to the interpretation of real research scenarios.

In the course schedule, I have noted the sections of this book that are relevant to the material in the lectures. In addition, the online materials for each small group session will include sections of the Gordis book that will give more in depth discussions of the learning objectives.

If you desire more detailed readings on biostatistics, I recommend the following, which will be on reserve in the library:

A concise, practical text, written by an important member of our faculty. Particularly strong are its sections on biostatistics and clinical trials, and a section on genetic epidemiology (new to the 3rd edition). Many students have reported that they have used this book, and found it very helpful.

This is a classic primer of biostatistics.

In addition, for those of you who want to go into more depth, I recommend the following texts, which will be on reserve in the library:

A highly readable text, geared toward teaching how to use knowledge of epidemiology and biostatistics in the critical evaluation of the medical literature. It will remain a useful reference during the clinical years, but lacks a public health/prevention focus, so is not entirely suitable for this course.

Statistical discussions beyond the level of this course, but it could be a useful reference for those whose curiosity is piqued to delve a bit into statistical theory.


This book provides a good overall review of Epidemiologic methods, and provides practice questions in each chapter.

A book that is particularly well suited for this course. Students may particularly like the organization of the review book, with questions and answers.

This provides a nice background to basic statistical methods used in epidemiology, in greater detail than the other suggested texts.


This is a more advanced text for those of you who really want to delve further. While this is beyond the scope of our course, it might be a good reference in the future.

IV. ON-LINE MATERIALS. The syllabus is up on our Web site, and includes some links to other useful sites. (Note: if you have any web links to suggest, please let me know.)
A particularly useful web resource for statistics information is an on-line statistics ‘textbook’ called Hyperstat: http://davidmlane.com/hyperstat/index.html

Another resource, for those who’d like to complement our lectures/readings with an on-line series of lectures on epidemiology and public health, is the “Supercourse” in Epidemiology: http://www.pitt.edu/~super1/ which compiles lectures from faculty and professionals from around the world.

Finally, there was a series of articles in the British Medical Journal regarding “How to Read a Paper” (BMJ 1997;315). These are all listed on the following website, which includes links to other relevant articles. The series provides a nice guide to applying the literature to clinical practice, and might prove useful in the future as you begin to evaluate the literature in EPHEM II and in clinical rotations.
http://infodome.sdsu.edu/research/guides/science/medlit.shtml
V. COURSE EVALUATIONS/ GRADING

This is a pass fail course. Passing the course will require the following:

- Attendance at ALL small group sessions.
- Completion of all 6 on-line quizzes BEFORE MIDNIGHT the day of each case discussion session
- Passing the Final Exam with a score of 65% or higher. (June 3rd, Robbins Auditorium 9:30-12:00), (see details below).
- An overall course grade of 65% or higher.

- Final Grade is based on:
  - 20%: Quizzes as noted above.
  - 80% final exam.

The real issue is: “how can I learn what’s important?” That should be straightforward, if you follow the steps below, you should be well prepared for the final:

- Be prepared for class by doing the assigned (required) reading
- Come to class (and come on time)
- Participate in class discussions
- Follow up on any material you didn’t understand by reviewing the suggested readings for that session, or contacting (preferably by email) your group leader or the course leader if you’re still confused.

SMALL GROUP SESSION QUIZZES (20% of final grade):

As noted above, the assignments for each small group session include completing a quiz available on the EPHEM I emed page. Quizzes will be posted prior to each session and must be completed by MIDNIGHT the day of the session. The purpose of the quizzes is to assess your preparedness for the small group sessions. Questions will be based on the assigned articles, and on the material from the syllabus/lectures that is relevant to the learning objectives for the case. In addition, each quiz will be set up to provide feedback for you to use in preparing, and to help you clarify concepts prior to the discussion. Finally, the section leaders will use the information regarding distribution of answers to identify areas that might require extra time in the discussion.

FINAL EXAM (80% of final grade/ MUST PASS THE FINAL with ≥ 65% to PASS the COURSE)

The final exam will be a short answer/short essay exam (i.e. not multiple choice), and will be administered in class. It will be based on interpretation of a recent/current
journal article (one that probably hasn’t been published yet as this syllabus goes to press). The article for the final exam will be posted on emed for you to review and study one week prior to the final exam, on Monday, May 27th. The date of the final exam is June 3, 9:30-12:00, Robbins auditorium. (Date for exam review session: May 29, 9:00-10:15 AM, Riklis Auditorium.)

VI. SUPPORT/ FEEDBACK/ QUESTIONS

Your section leaders will be available after each class, through email, and by appointment. If you ask an important question by email, your question and an answer may be sent to the entire class – but your name will be deleted from the correspondence. Contact for your section leader is listed on the course website along with the group assignments.

Also, through the duration of the course, Dr. Derby will be available to all students. Simple questions are best addressed by email; if you wish to schedule an appointment, please call Dr. Derby at 430-3882. If you have administrative questions regarding the course, please call Ms. Margie Salamone at 430-3465. ALL QUESTIONS REGARDING EXCUSED ABSENCES MUST BE directed to your section leader, with Dr. Derby copied on the email.
LEARNING OBJECTIVES

The overall goal is to acquire the tools required to critically evaluate the medical literature and to understand how studies conducted in populations relate to clinical practice and care of the individual patient.

At the end of this course, you should be able to define and interpret each of the following concepts: (bold print = major objectives):

TYPES OF STUDY DESIGNS and the pros and cons of each
- cross-sectional
- ecological
- case-control
- cohort
- clinical trials

MEASURES OF DISEASE/ RATES
- prevalence and incidence and the difference between them

MEASURES OF ASSOCIATION (RISK ESTIMATES)
- absolute risk (incidence)
- relative risk
- odds ratio
- attributable risk
- number needed to treat (NNT)/ Number needed to harm

TYPES OF BIAS
- selection
- detection
- recall
- confounding

SCIENTIFIC REASONING
- the null hypothesis/ hypothesis testing
- causal inference

STATISTICAL INFERENCE/
VARIABILITY AND ESTIMATION
- sampling distributions
- standard deviation, standard error
- the null hypothesis
- p-values
- confidence intervals
chi-square, Fisher’s tests
t-tests, ANOVA, Wilcoxon
power and type II error
significance and type I error
1- vs. 2-tailed tests
Ways to control for confounding
  multivariate analysis
  logistic regression
  stratification
life tables/Cox models

CLINICAL TRIALS
  randomization
  stratification
  blinding
  intention-to-treat

SCREENING AND CLINICAL DECISION MAKING
  pros and cons of screening tests
  lead-time and length bias
  sensitivity
  specificity
  positive and negative predictive values
<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture / Case Conference</th>
<th>Relevant Chapters in Gordis</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 5: 1:30-3:30</td>
<td>LECTURE I: Introduction: What kind of study is this, and why does it matter?</td>
<td>Chp 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chp 7:131-137</td>
</tr>
<tr>
<td></td>
<td></td>
<td>138-140</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chp 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chp 10:177-180</td>
</tr>
<tr>
<td>March 12: 1:45-3:45</td>
<td>LECTURE II: How strong are these results and should I believe them?</td>
<td>Chp 3:37-40,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43-47, 50-52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chp 10:180-191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chp 11:215-216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>220-221</td>
</tr>
<tr>
<td>March 29: 1:45-2:45</td>
<td>LECTURE III: From screening to diagnosis.</td>
<td>Chp 5:85-91,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96-101</td>
</tr>
<tr>
<td>April 16: 1:30-3:30</td>
<td>LECTURE IV: Are these results “Significant”??: Sampling and Principles of statistical inference.</td>
<td></td>
</tr>
<tr>
<td>April 23: 1:30-3:30</td>
<td>CASE CONFERENCE 1: Prospective population based studies: The Framingham Study</td>
<td></td>
</tr>
<tr>
<td>May 7: 1:45-3:45</td>
<td>CASE CONFERENCE 2: Case-Control Studies. Doll and Hill on Lung Cancer</td>
<td></td>
</tr>
<tr>
<td>May 9: 1:45-3:45</td>
<td>CASE CONFERENCE3: What puts the P in P-values? Practice with basic inferential statistics.</td>
<td></td>
</tr>
<tr>
<td>May 14: 1:45-3:450</td>
<td>CASE CONFERENCE 4: Randomized Clinical Trials: The Lipid Research Clinics Study</td>
<td></td>
</tr>
<tr>
<td>May 21: 1:30-3:30</td>
<td>CASE CONFERENCE 5: PSA Screening for Prostate Cancer</td>
<td></td>
</tr>
<tr>
<td>May 28: 1:30-3:30</td>
<td>CASE CONFERENCE 6: Critiquing and Article.</td>
<td></td>
</tr>
<tr>
<td>May 29: 9:00-10:15</td>
<td>Review Session : Riklis Auditorium</td>
<td></td>
</tr>
<tr>
<td>June 3: 9:30 am -12pm</td>
<td>FINAL EXAM: ROBBINS AUDITORIUM</td>
<td></td>
</tr>
</tbody>
</table>
EPHEM I
Lecture I: Introduction to the Epidemiologic Approach

This lecture will provide an introduction to epidemiology and its applications to clinical practice, will discuss basic concepts of study design, and an introduction to basic measures of disease occurrence.

Learning Objectives:

At the end of this lecture, you should be able to:

1. Discuss the difference between observational and experimental studies.
2. Define ecological and cross-sectional studies.
3. Describe the design of cohort (prospective) studies, and of case-control (retrospective studies, and discuss the advantages and disadvantages of each.
4. Define the advantages of randomized controlled trials.
BACKGROUND LECTURE I: PRINCIPLES OF STUDY DESIGN AND ANALYSIS

I. Introduction: What is Epidemiology/ Why do we need to know?

Epidemiology is the study of the distribution and determinants of disease in populations. The roots are Greek: “epi” meaning upon; and “demos” meaning the people. So what we’re concerned about is that which is upon the people (such as plague, pestilence, and the like). The goals of epidemiologic studies are to identify the extent and causes of disease, study the natural history and prognosis of disease, test preventive strategies and treatments and provide data for making public health policy decisions.

How does epidemiology relate to clinical practice? First, the ability to make a diagnosis depends upon knowledge obtained from studies relating clinical findings to pathology/ disease in large groups of patients. Second, the ability to give a prognosis is based upon population data regarding the clinical course of large groups of persons with a specific disease. Finally, choices regarding treatments are based on results of clinical trials that evaluate therapies in large groups of patients. Thus, while clinical decisions relate to individuals, the knowledge/ data that form these decisions are population based.

II. Overview of Study Designs used in Clinical Research:

To be able to critically interpret studies in the literature (or to design your own), you must have an understanding of the issues involved in study design, to learn the impact of design on what we observe and what inferences we can make.

In general, the goal of any epidemiologic study is to establish whether there is an association between an exposure (or characteristic) and an outcome.

Epidemiologic studies may be classified into two groups:

Experimental studies are those in which the investigator manipulates or controls the exposure of interest and examines the association of the exposure to a particular outcome. Randomized Controlled Trials are considered the “gold standard” in terms of proving causation or evaluating a treatment.

However, there are situations in which it is not ethical to randomize subjects to exposure. For example, if one suspects that a toxic pesticide is a carcinogen, it would not be ethical to randomize study participants to exposed and unexposed groups. Instead, you must design a study in which you observe the disease rates in those with and without the exposure of interest, or an observational study. There are many types of observational studies, and we will study these and clinical trials in depth in the remainder of the course, using the case conferences to explore their design and interpretation.
A. Observational Studies

1. Case Series: Clinical observations in series of patients or case-series often provide clues to possible etiologic factors. For example, one of the first suggestions that smoking might be related to lung cancer came from the observations of a surgeon who observed that the vast majority of his lung cancer patients reported that they smoked cigarettes. However, case-series are limited in that they do not include a control or comparison group. Without comparing the rates of smoking in lung cancer patients to those without lung cancer, there is no way to determine whether smoking might be related to the disease.

2. Ecological Studies: An ecological study is based on comparisons of group data, and can provide clues regarding exposure/disease associations. However, ecological studies are distinguished from other observational studies by the fact that they do not have data on individuals, and thus the inferences to be drawn from the study are limited. For example, if you were comparing mortality in Alaska and Florida, you would be conducting an ecological study. Based on the observation that mortality rates are lower in Alaska, we might make the inference that exposure to cold air increases longevity. We might make this inference -- but I expect we wouldn’t. There are two important reasons to avoid this inference:
   a. Cold air exposure is not the only difference between these populations. Even after adjusting for age and sex, there are important differences between the populations other than the climate. For example, many people move to Florida when they retire, and many people move to Alaska to find employment. This does, to be sure, contributes to the difference in age distribution -- but even after taking that into account, the fact remains that people who are working, at a given age, are healthier than those who don’t (the so-called ‘healthy worker effect’), and that healthier people have a lower mortality rate;
   b. Since we’re looking at summary statistics for an entire population, we’re not able to specifically associate any particular risk factor with the people who actually have the outcome. Sure, it’s warmer in Florida -- but that doesn’t mean that the particular people who died were exposed to excessive warmth. To take it to an absurd extreme, what if all those who died in Florida worked in a refrigerated meat packing plant, and all those who died in Alaska worked in a hot iron-smelting factory? That is, in ecological comparisons, even if a putative risk factor seems to be associated with an overall rate, we can’t determine whether anyone with the outcome was actually exposed to the specific risk factor. This is called the ‘ecological fallacy.’

3. Cross-Sectional Studies: A cross-sectional study is one in which the presence of a suspected exposure and the presence of the outcome (or disease) are ascertained
simultaneously. In other words, we take a “snap-shot” or survey of the study population at a particular time. These studies can be valuable in establishing associations between suspected factors and disease, but are limited in that we do not know which came first. In other words, we do not know the temporal sequence of the exposure and the outcome. The problem is that we sometimes want to make inferences from cross-sectional studies that are not appropriately made by observations at a single point in time.

For instance, we might be interested in finding out whether doctors read fewer medical journals as they get older. We do a cross-sectional study of practicing physicians, and ask them how many journals they read each month. The results are summarized in the chart to the right.

It’s certainly tempting here to conclude that as physicians age, they read fewer journals. But think about it: we’re not looking at any physician as s/he ages, but rather at different age groups of physicians at a single time. Sure, they MAY change their habits over time: but it’s just as plausible that physicians in their sixties ALWAYS read fewer journals (in fact, many fewer journals were published when they started in practice, and perhaps physician habits re: reading were different). So what we may be seeing is not a change over time in individuals, but a change in each individual cohort: those who became physicians in the 1960s, 70s, 80, or 90s. This is what’s known as the cohort effect, and underscores the danger in making longitudinal inferences from cross-sectional data.

In summary each of the study designs above suffers from design limitations that limit the inferences that can be made from the results. Case series lack a control or comparison group, ecological studies lack data on individuals, AND neither ecological nor cross-sectional studies provide information regarding the temporal relationship between exposure and disease.

The two the remaining types of observational study designs improve upon these designs cohort studies and case-control studies.

4. Cohort/ Prospective Studies: One way to determine whether an exposure increases the risk of developing a disease is to select a group of people who have the exposure and a group that does not, and to follow these groups to see the rates at which each develops disease. This study design is called a cohort study. If, for instance, we believe that hypertension may lead to subsequent heart attacks and
Design of Cohort (Prospective) Studies

Design of Cohort (Prospective) Studies

Begin Study

Follow-up

Exposed

Disease

No Disease

Not Exposed

Disease

No Disease

Defined Population Without Disease

or

strokes, we can study people with hypertension with those without hypertension, and compare the incidence rates of these outcomes. If we enroll those patients today, and look for MIs and strokes 10 years from now, that would be a cohort study. (The Framingham Heart Study is the classic example of such a study.) Cohort studies are often called prospective studies, because we are prospectively following exposed and unexposed groups through time to determine occurrence of the outcome.

4a. Identification of subjects: Choosing a cohort and determining exposure status

When we decide to perform a cohort study, we first try to identify a group with some similarities (for instance, born between 1930 and 1939), and without the disease of interest at entry into the study (we are interested in incidence or development of new disease). Such a group is called an inception cohort. There are several approaches. One may begin by selecting a cohort who lives in a particular area (say residents of Framingham MA). Another approach is to select persons in particular groups, such as physician, nurses, veterans, or medical plan subscribers.

Once a group is selected, the investigator must determine who in that group was 'exposed'. This term is itself an oversimplification, since exposure is usually not an all or none issue, but is graded. For instance, 'hypertension' is an arbitrary term, and could be defined in any of a number of ways (e.g. diastolic blood pressure greater than 90 mm Hg; systolic BP over 160 mm Hg).

In the scenarios above, the unexposed, or control group neatly follows from the definition of the exposed group, and the inception cohort is divided into those with and
without the exposure. This is not always the case. If one is interested, say, in an occupational exposure, one might begin by selecting a specific exposed group (such as ship workers, or atomic bomb survivors, or seventh day Adventists). In this case, controls are not available from the same inception cohort, and it is necessary to identify a different group to function as controls. For instance, a study was done to evaluate the effects of exposure as a rubber worker on mortality. It turned out that the mortality rates of rubber workers was only 82% of that for US males overall. Does this mean being a rubber worker is good for you? The healthy worker effect means that the exposed group is in general healthier than the general population. To control for this phenomenon, it would be helpful to find an unexposed group in whom the same effect is operating, say another working population. This can be tricky given that you always have the concern of comparability on factors other than the exposure.

4b. Assessing exposure

It is important that data be reliable in cohort studies. Accurately determining the exposure to the risk factor is critical, and such exposure must be carefully measured. Even when the investigator is doing the measurements himself, there must be a standardized, reliable protocol [e.g. since blood pressure measurements have intrinsic variability, there’s a standard way to measure it: in a patient who’s been seated at least 10 minutes, in the left arm, using an average of three such measurements on different days over 2 weeks]. One of the biggest issues in prospective research is that exposures may change over time. While exposure to Agent Orange or an atomic bomb blast occurred in a finite period of time, exposure to coffee drinking or dietary fat tends to change over time. This must be taken into account in the analysis and interpretation of prospective studies. In studies where exposures are measured repeatedly over time, it is crucial that measurements be obtained in the same way at every point in time.

4c. Assessing Outcomes

Also important is the methodology for the assessment of outcomes. The best studies will have appropriate, standardized assessment of possible outcomes for ALL study subjects, and this assessment should be done in a blinded way (that is, those who determine whether or not the outcome has occurred should be unaware of the exposure status of the individual). Suppose we measure blood pressure now, and look for strokes later. Which patients will get CT scans for headaches? (Clinicians might be more likely to order this expensive test in people they perceive at high risk for intracranial hemorrhage -- such as those with hypertension.) How will mild abnormalities be read? (Radiologists might be more inclined to read subtle findings as abnormal in higher risk patients -- such as those with hypertension.) How will causes of death be classified? (You see the point.) All of these problems are issues of BIAS. Such biases are eliminated with a standardized, blinded assessment of outcomes.

5. Case-Control Studies: Often the conduct of a cohort study is not practical, for example when the disease or outcome is rare, and thus would be necessary to follow a
huge population for a very long time in order to reach a conclusion. For example, Parkinson’s disease has an incidence rate of approximately 1/1,000. Coenzyme Q\textsubscript{10}, a vitamin like substance found primarily in mitochondria, has been suggested to reduce the risk of PD. However, given the low incidence rate, even assuming that CoQ10 cuts risk in half, a cohort study would require following 31,443 persons for a year, while a case-control study would require only 177 subjects, and no follow-up. At other times, although a longitudinal study might be feasible, not enough is known regarding the potential exposure-disease association to warrant the time and expense required for a prospective study.

The case-control study design allows us an effective and convenient way to look back at past exposure to potential risk factors for a disease. We begin the study by defining cases (people who have the disease) and appropriate controls (people without the disease), and then ascertain whether the past (i.e., prior to the onset of disease) exposure to a potential risk factor differed between these two groups. Because you are looking back in time to ascertain exposure, the term retrospective is often used as a synonym of case-control -- however, while all case-control studies are retrospective, not all retrospective studies are case-control. Therefore, most epidemiologists prefer the term case-control. What distinguishes a case-control study from a cohort study is that in a case-control study we begin by selecting those who already have the outcome and those who do not.

5a. Identification of Cases

When we decide to perform a case-control study, we first try to identify a group with the disease. Perhaps the most important consideration is that the disease itself must be strictly defined. The 'pornography' approach (I can't tell you what it is, but I know it when I see it) does NOT work in epidemiology -- we need strict criteria. Thus, not every person who carries a clinical diagnosis of a disease will necessarily meet the strict criteria required for entry into a study. For example, if an exposure is suspected to be related specifically to a particular cell-type breast cancer, then selecting a heterogeneous group of all breast cancer cases might dilute the association you are looking for.
The source of the cases must also be considered. Usually a convenient source of cases is hospitalized patients, but they are not always representative of all cases of the disease. Especially when cases from only one hospital are selected, it is possible that referral patterns or other factors distort the estimates of exposure in your case group. In other words, you want to try to select cases that are representative of all cases with respect to their exposure status. However, combining this goal with the issue of strict case definitions described above, points to a common dynamic tension in study design: the conflict between the internal validity of the study and its generalizability (or external validity). Many have said that validity should not be compromised in an effort to achieve generalizability, and this makes sense. If the cases are chosen, therefore, from a group believed to be more 'reliable' historians than the general population (such as patients of high socioeconomic status [SES]), we ought to select a similar control group. If we were enrolling high SES cases with MI, and high SES controls, we might validly conclude, say, that physical inactivity is associated with MI in our study population. However, we must accept that the findings of such a study may NOT be generalizable beyond the specific SES stratum evaluated; for instance, 'physical activity' may mean something very different in a working class population.

Another issue in case selection has to do with the decision to study prevalent cases or incident cases. The study of incident cases requires waiting for the cases to develop and be diagnosed, while prevalent cases are in general more readily available. However, the study of prevalent cases is tricky, because it is always possible that risk factors we identify might be related to survival with the disease rather than development of the disease. Exposures related to higher case fatality will be underrepresented in a prevalent case series, while those that increase survival will be overrepresented. (Refer back to the relation of prevalence to incidence in Lecture I).

5b. Identification of Controls

This is often the crux of the case-control study. As noted in the example above, for purposes of internal validity, the control group must be comparable to the cases in basic demographic terms. They should also be representative of the general, non-diseased population with respect to their exposure status.

Thus, in a case-control study exploring whether coffee drinking causes pancreatic cancer, a group of highly expert epidemiologists chose a group of hospitalized cases in several hospitals in Massachusetts and Rhode Island. In order to make their control group as similar as possible to the cases, they selected controls who were hospitalized patients of the same doctors in the same hospitals. They found a history of coffee drinking was significantly associated with the presence of pancreatic Ca: more pancreatic Ca cases had a history of coffee drinking than did non-cancer controls.
It was concluded that coffee may cause pancreatic cancer. But it turned out that this inference was an error. The problem was that cases (those with pancreatic cancer) had doctors who were, by and large, gastroenterologists. Since the investigators chose hospitalized controls with the same doctors, we have to think about what kinds of disease lead to hospitalization of patients of gastroenterologists. Two big diagnostic categories include peptic ulcer disease and inflammatory bowel disease -- two groups of patients who are often advised not to drink coffee. Thus, in this case, the apparent association wasn’t seen because cases were MORE likely to drink coffee, but because controls were LESS likely to drink coffee than the general population. Clearly, not an easy task, this business of selecting controls.

6. Comparing Case-Control and Cohort Studies

**Advantages of case-control studies**
1. Particularly useful in the study of rare diseases
2. Inexpensive, quick
3. No loss to follow-up
4. Multiple exposures can be assessed for a single disease outcome

**Advantages of cohort studies**
1. Temporal relation appropriate
2. Incidence rates can be determined (therefore true relative risk can be calculated; see lecture II)
3. Less subject to bias (e.g. recall bias)
4. Useful with rare exposures
5. Multiple disease outcomes can be assessed for a single exposure

Just to muck this up a bit, it is possible to have a cohort design that starts in the past (the so-called retrospective cohort study). The basic idea is the same as any cohort study: a group of individuals at risk is classified as exposed or not exposed, and then followed forward to see how many develop the disease. The difference is, their exposure status had been defined sometime in the past (at a time when they were at risk for but had not developed the disease), so that the subsequent development of the disease has also been in the past. For instance, if I have good data on Vietnam vets with respect to their exposure to Agent Orange, I can classify them as exposed or not exposed from 1960-1970, and then follow them for the development of cancer between 1980-1990. This would give me incidence rates.

Alternatively, it’s possible that a new idea is generated during a prospective study, and that cases and controls are assembled within the cohort study, to look back at the entry data (or perform new assays on pre-collected specimens, such as blood or DNA) to determine exposure. This is called a “nested case-control study.” A nested case-control study is still a case-control study (individuals are entered into the study based on whether they have the condition of interest [cases] or not [controls]), but it is “nested” within a cohort study. A major advantage of this approach is that data on
exposure has already been collected, and such data will be “blinded” to disease status (by definition, since disease will not have occurred yet). In Case Conference V we will explore the nested case-control study in more detail.

The most important aspect of the distinction between a ‘cohort’ and a case-control design is the ability to calculate incidence rates. In a cohort, characterized by exposure status and then prospectively followed over time, incidence rates can be calculated. That is, if I follow 500 women exposed to DES, and after 20 years I find that 25 of them have developed breast cancer, I know that the incidence of breast cancer in this group is 2.5 per thousand per year. In a case-control study, where I collect a group of women with breast cancer and select controls that do not and then determine who was exposed to DES, I can’t calculate an incidence rate. The temporal sequence will not allow it.

**B. Experimental Studies/ Clinical trials:** When we want to know whether a therapy is effective, there is a paradigm: the PROSPECTIVE, RANDOMIZED, DOUBLE-BLIND, PLACEBO CONTROLLED TRIAL (RCT).

While there may be some redundancy here, but each word does have a specific meaning.

PROSPECTIVE -- unless the treatment precedes the outcome, it can’t cause it
RANDOMIZED -- this allows us to avoid bias in assignment of treatment. (Please note: randomization is NOT done *in order to* achieve uniform distribution of potential confounding factors. Although that is the usual consequence of randomization, particularly if the study is large enough, it is not its intent. Thus, while it’s common to look at the traditional Table 1 in a published randomized trial, wherein the baseline characteristics of the two randomly assigned groups are compared, as checking to see if randomization worked, that is incorrect. Randomization always works, because it always allocates treatment in an unbiased fashion. Table 1 tells you whether baseline variables are evenly distributed in the two groups: that *usually* works, but if it doesn’t, we may revert to analytic techniques to control for potential confounding factors (see above). So even if there are, say, more smokers in group A than group B, randomization still met its primary objective -- but the investigators may still need to control for smoking in the analysis.)
DOUBLE-BLIND -- neither the subject nor the investigator knows who is receiving the active treatment and who is not. This avoids bias in categorizing outcomes.
PLACEBO-CONTROLLED -- this is done to achieve blinding. In some cases, the goal of the study is not to determine the efficacy of a particular treatment compared to NO treatment, but compared with a STANDARD treatment. In such a case, there would be no placebo -- though it would still be important to do this in a double blind fashion.
It could be (and has been) argued that in the absence of random allocation of treatment, there is no way to reliably evaluate the efficacy of therapy. This is a rather harsh statement: it imposes a severe limitation on what can be learned from observational data. But it suggests that absent randomization, there are always potential differences (biases) between patients who take a particular treatment and those who don’t, or those who practice a particular health behavior and those who do not; biases that may invalidate the inference that differences observed between the groups must be caused by the treatment itself.
The following true example makes this point: A large, pooled study was done of cholesterol lowering drugs, where the investigators compared compliant men (defined as those who took $\geq 80\%$ of medication) with non-compliant men (took $< 80\%$).

<table>
<thead>
<tr>
<th>Compliant</th>
<th>N</th>
<th>Adjusted mortality *</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>882</td>
<td>26%</td>
</tr>
<tr>
<td>Yes</td>
<td>1813</td>
<td>16%</td>
</tr>
</tbody>
</table>

* adjusted for 40 baseline risk factors
Relative Risk Reduction = $(0.26 - 0.16)/0.26 = 38\%$
$P= 0.0000000073$

The only apparent interpretation of these data is that the drug must be effective. After all, the only difference between the two groups is whether or not they took the drug. Perhaps one may argue there’s a bias here: compliant patients tend to be healthier, more health conscious, etc. -- wouldn’t that bias the results in favor of what we observed? Well, yes -- but we’ve already controlled for every risk factor that we could measure. Is it plausible to suggest that such nebulous factors as health behaviors could account for this substantial, and highly significant, reduction in mortality? It doesn’t seem plausible, at least not at first -- but it turns out that it must be the explanation, because this analysis was done entirely within the placebo group of a large RCT! That is, we’re comparing men who were compliant with their placebo with men who were non-compliant with their placebo. Since it’s a placebo, our observed effect can’t be from the action of a drug, but must be due to something else about people who are or are not compliant.

If an RCT is required to truly evaluate the effect of therapy, a corollary would be that you can’t learn from your clinical experience. This is true in the sense of learning about the effect of drug "per se" -- and nowhere is it more true than in preventive medicine. There is absolutely no way to know if, as you treat hypertensive patients, you’re preventing strokes. That’s not because you aren’t seeing enough patients, or haven’t organized your observations well enough, but because you just can’t know. If you treat a patient and he doesn’t have a stroke, that doesn’t mean you’ve prevented it -- lots of hypertensives never have strokes, even if they’re not treated. Likewise, if you fail to treat a patient and he does have a stroke, that doesn’t mean your inaction caused it -- even well treated hypertensives can get strokes. The fact is, we do know hypertension treatment prevents strokes, but we know it because RCTs were done. That’s the only way we know it.

But please don’t interpret this as a knock on the value of clinical experience. Experienced clinicians are better able to determine how to approach hypertension treatment in their clinical populations -- what classes of agents are likely to lead to specific problems with certain patients, how to improve compliance, how to arrange follow-up. The point here is that there are limits to what you can learn by experience,
not that experience isn’t very important. A great example is given in a current raging clinical controversy: the treatment of localized prostate cancer in older men. Many observational studies have suggested that the mortality may be unchanged by surgery, when compared with “watchful waiting.” Urologists and patients are fond of sharing anecdotes: “I’m alive today because I had my PSA test, and got surgery in time!” But think about it: how can he possibly know that? There’s no way to know if you wouldn’t be just as alive, and feeling just as well, even if you hadn’t had the surgery. We really need a randomized trial to know that. In fact, in 2005, a randomized trial reported that prostate cancer mortality was reduced as a result of radical prostatectomy, but the benefit was limited to the group of patients younger than 65 years of age.

However, there are some things clinical experience can tell you, more reliably than published data. If the literature suggests that 30% of men will become impotent after surgery, and you know, based on referring 100 patients to a specific urologist, that your surgeon has an impotence rate of only 5%, then you shouldn’t expect a 30% rate in your center. In fact, well-designed studies may use treatments less (or more) effectively than is available to your institution, in which case the results may not be applicable to your patients.

1. Ethical considerations

This is an issue of prime importance. Patients must consent to participate in trials, and such trials must be approved by IRBs (Institutional Review Boards) before they are undertaken. We have come a long way since earlier research excesses, but this is not ancient history. Willowbrook and Tuskegee were just a few decades ago. (If you haven’t heard about these events, you may want to read about them.)

Some have argued against the RCT, saying it’s unethical to experiment on humans. They say that the usual justification, that of ‘equipoise’ (it’s equally likely that the treatment will be beneficial or harmful) is a canard, because investigators wouldn’t do a study unless they had reason to believe the treatment would work. A response: sure, the investigator always thinks the treatment will work; but he doesn’t know. An important object lesson was the CAST (Cardiac Arrhythmia Suppression trial) study, published in the New England Journal of Medicine in 1989; 321: 406-12. In this study, several drugs that were known to suppress arrhythmia’s, along with placebo, were administered in random fashion to a group of patients with arrhythmia’s, in a double-blind study. The question was: even though we know that arrhythmias predict mortality, and even though we know these drugs suppress arrhythmias, will the drugs reduce mortality? The surprising answer: the drugs increase mortality! A wonderful lesson about the dangers of confusing what we believe with what we know.

2. Study Populations

This is the next major issue in RCT design. As noted earlier, there is a constant dynamic tension between precision and generalizability; or internal versus external
validity. Sometimes, the most reliable research subjects are not drawn from the general population: this is exemplified by the Physician’s Health Study, which used male physicians to study beta carotene and aspirin. An improvement in reliability within the study (physicians share many characteristics, which will be comparable in exposed and unexposed groups) is tempered by the difficulty in extrapolating the findings from a mostly white, all male, highly educated population to the rest of the world.

No matter how broad the inclusion criteria for your study population are, you’re always using that relatively small proportion that is willing to give you informed consent and volunteer. Bottom line -- volunteers are unusual people.

3. Endpoint assessment

Investigators must carefully plan out their strategy for assessing endpoints. First, criteria for determining that an endpoint has occurred must be carefully laid out prior to the study, along with a clinical approach to making the diagnosis. Second, to avoid bias, those determining if the endpoint has occurred must be blinded to study treatment. Even if, for some reason, a double-blind study could not be performed, it may still be possible to blind the outcome assessment, and that should be done.

4. Randomization and Confounding

Remember, randomization is not done to create similar groups, but to minimize bias in the assignment of treatment. This usually results in similar groups, but not always, there is no guarantee that randomization will yield similar distributions of characteristics across the treatment arms of a study. If not, the discrepancies (which would be noted in the traditional Table 1) are corrected for in the analysis, either by stratified, restricted, or multivariate analysis.

5. Intention to Treat

A common problem faced by investigators: how do you analyze patients who are not compliant, or who refuse the treatment that they were initially randomized, or who drop out of the study before it’s complete? Do you ‘cross them over’ in the analysis if they take the other treatment? Do you drop them out of the study? Actually, what’s usually done is called the intention to treat method. This analysis involves maintaining subjects in the group to which they were initially randomized, whether or not they ever receive the treatment (or even if they cross over to the other group). While it’s somewhat counterintuitive to analyze subjects as if they’ve received a treatment they didn’t get, it’s clearly the best option, for three reasons:

A. it is methodologically appropriate, since it maintains the random allocation (thus, is the only way to avoid bias);
B. it is statistically conservative (thus, if you find a statistically significant
difference, it’s more likely to be real);
C. it is clinically reasonable, since it measures the impact of a decision at a given
point in time (that is, just because you prescribe a treatment doesn’t mean the
patient will comply -- so we see in the intention to treat method the outcome of
your prescription, rather than the treatment *per se*).

6. **Subgroup analysis**
   In a word: BEWARE! Investigators love to try to make a silk purse out of a sow’s
ear by massaging the data until something pops out, and this often plays out as a
subgroup analysis. However, it’s often important to look at specific subgroups. In
general, think of this as hypothesis generating, rather than hypothesis testing, unless
these subgroups, and the analysis of them, was planned *a priori* in designing the study
(i.e. not part of a data-dredging exercise).
EPHEM I
Lecture II

How strong are the study results? Should I believe them? (Bias and Confounding)

This lecture describes the measures used in epidemiologic studies to quantify the associations between exposures and disease. We will discuss how bias and confounding can distort these measures.

Learning Objectives:

At the end of this lecture, you should be able to:

1. Understand the difference between incidence and prevalence, and how they are related.

2. Understand adjusted rates, the difference between crude and adjusted rates.

3. Calculate and interpret Relative Risk.

4. Interpret absolute risk (incidence), relative risk and attributable risk, and contrast the measures.

5. Calculate and interpret the Number Needed to Treat

6. Define person-years and why person years are used to summarize event rates from cohort studies.

7. Calculate and interpret the Odds Ratio, and describe the difference between the Odds Ratio and the Relative Risk.

8. Understand the role of Bias and confounding
BACKGROUND MATERIAL: LECTURE II.

I. MEASURES OF ASSOCIATION: HOW STRONG ARE THE STUDY RESULTS?

A. Review of Measures of Disease Occurrence

Central to the methodology of epidemiology is the process of measuring and quantifying as precisely as possible. Only with precise measurements can we identify patterns: trends over time, differences between populations, and the impact of risk factors. Clues regarding etiology and avenues for prevention are obtained by describing disease occurrence in relation to person, place and time, or by describing the who?-what?-where? of disease occurrence. Therefore, any discussion of this methodology must begin with a nosology of the measurements, so we all agree that we’re talking about the same things.

1. Rates

In understanding disease occurrence, we are dependent on clear data. The usual representations are rates and proportions of disease, or outcome. Each of these measures includes both a numerator and a denominator. Numerator and denominator must be clearly defined, and everyone in the numerator must also be included in the denominator. A proportion tells us the percent of the population that is affected, while a rate includes a unit of time, or tells us how quickly the outcome is occurring. For example, the annual rate of Hepatitis C infections in the state of NY. There are all kinds of ways to present such data: depending on your goals, you can manipulate denominators to make your point. How commonly are data presented in a manner like: someone is involved in a car accident every 9.3 seconds? (The poor guy!) Epidemiologists do use different methods of presenting rates, but the standard methods are the best.

2. Morbidity Rates

When describing the frequency of occurrence of disease in a population (morbidity), there are two important rates which must be distinguished. The prevalence rate tells us how common the condition is in the population. That is, in a given population, what proportion has the condition at a given time? Let’s say I’m working for a major pharmaceutical company, interested in marketing my new beta-blocker to patients who have previously suffered a myocardial infarction (MI, or heart attack). If I’m going to project sales estimates, I need to know the prevalence of MI in the population (let’s say the Bronx). So I may randomly sample 1,000 Bronx residents this week, ask them “have you ever had a heart attack?”, and perhaps 50 say “yes.” Therefore, I say that the prevalence of the condition is 50/1,000 (and since there are 1.2 million residents in the Bronx, we might infer that there are 60,000 Bronx residents who have had an MI). Could be an important market to go after!
In contrast, the *incidence* rate tells us how many new cases of the condition occur over a given time period. Perhaps I’m working for a different pharmaceutical company, which is trying to market a thrombolytic (clot-dissolving) drug in the Bronx. Since this drug is only useful in new (i.e., incident) cases, the prevalence rate doesn’t concern me: I need the incidence. So I follow 1,000 Bronx residents *who have not had a prior MI* (these “prevalent cases” are excluded so that we can look at the occurrence of first MI) for 5 years. At the end of that time, 40 new MIs occur. Therefore, the incidence rate is 40/1,000/5 years, or 8/1,000/yr. Similar projections to the 1.2 million Bronx residents (assuming, of course, that our sample is representative of the risk of the entire Bronx population) would estimate about 9,600 new or incident MIs per year.

There is an interesting relation between incidence and prevalence:

\[
\text{Prevalence} = \text{Incidence} \times \text{Duration} \text{ (or Median Survival)}
\]

That is, the number of existing cases (old and new) at any given time, depends on the rate at which new cases develop (incidence) and the length of time that a person remains a case (duration). So in this case, since prevalence is 50/1,000, and incidence is 8/1,000/yr., the median survival is 6.25 years/case (check the calculations, if you wish: from above, survival is approximately P/I).

There is an important corollary to this relation: that a declining prevalence is not necessarily a good thing! Of course, one way to diminish prevalence is to reduce the incidence -- that would be good. But another way is to reduce the survival in patients who develop the disease -- and that’s not good. Imagine that the incidence of MI remained at 8/1,000/yr., but that the average survival reduced from 6.25 years to 4 years. Over time, the prevalence would reduce from 50/1,000 down to 32/1,000. Would we want to proclaim to the residents of the Bronx: “Eureka! We’ve achieved a major public health breakthrough! The prevalence of MI has been reduced by 36%!”? This is hardly something to brag about, when you see where it comes from.

### 3. Mortality Rates:

While incidence and prevalence tell us about the rates of disease occurrence in the population, mortality rates provide important information regarding the severity of disease, whether particular groups are at increased (or decreased risk) of dying from a disease, whether there have been trends in prevention or treatment of disease over time, and in cases where the disease in question has a very high death rate, mortality can be a marker for incidence rates. Different types of mortality rates are used depending upon the information required. Annual death rates are simply the total number of deaths in a given period divided by the total population at risk of dying from the disease during that period. These may be overall, age-sex- or cause specific.

A *case-fatality* rate is a particular type of rate that should be distinguished from mortality rates. Case fatality tells us the proportion of persons diagnosed with a particular
disease who die of that disease in a given interval. It differs from a mortality rate in an important way: In a mortality rate, the denominator is the total population at risk; in a case-fatality rate, the denominator includes only persons diagnosed with the disease in question.

4. Crude and specific rates

In the preceding discussion, we assumed that the rates in our study sample would apply to the entire Bronx population. That would mean that we sampled randomly from the entire population of the Bronx -- which would not be an efficient way to conduct a study. Why should we go to the effort of including, say, 8 year old children, who are very, very, very unlikely to have an MI over the next 5 years? Since the incidence of MI is so different in different age groups, we may not get very useful information from an overall (or crude) incidence rate. What we want, instead, is age-specific incidence. We may find, for instance, that the rates of MI break down by age as follows (note: these are all fictitious data):

<table>
<thead>
<tr>
<th>Age category</th>
<th>MI rate (per 1,000/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;18</td>
<td>0.02</td>
</tr>
<tr>
<td>18-34</td>
<td>0.1</td>
</tr>
<tr>
<td>35-54</td>
<td>6.2</td>
</tr>
<tr>
<td>55-64</td>
<td>9.8</td>
</tr>
<tr>
<td>65-74</td>
<td>12.3</td>
</tr>
<tr>
<td>75+</td>
<td>23.6</td>
</tr>
</tbody>
</table>

These data would then allow us to make more reasonable projections to the Bronx as a whole (assuming we know the age distribution in the Bronx population). It would also allow us to make more reasonable comparisons to other populations, if that is our goal: we might compare, for instance, age specific rates for those 65-74. Then we know we’re comparing apples and apples. Imagine if we compared crude rates in two populations, one where the mean age was 25, and another where it was 65. Certainly this would be apples and oranges.

We might be interested in all kinds of specific rates: age-specific, sex-specific, race-specific. We could even look at specific outcomes, rather than specific populations. For instance, an overall (or crude) mortality rate in a population tells you something, but a cause-specific mortality rate tells you something else. We may want to know not how many men in their 60s are dying each year, but how many are dying from MI each year. So we would calculate an age- and sex-specific, cause-specific mortality rate from MI. (If this sounds complicated, just keep in mind that different ways of looking at data can tell you different things.)

5. Comparing Rates Across Different Populations: Adjusted rates
Let’s say that the outcome of interest is associated with age, and that the age distributions of the populations being compared are very different. Differences in the crude (unadjusted) rates might simply reflect differences in the age distributions, rather than differences in the occurrence of the outcome. As shown above, one approach is to compare age specific rates. However, the problem of looking at specific rates is the basic problem of lumping versus splitting: after splitting the data up into all the different age groups, for example, it’s hard to see an overall general pattern. To accomplish this, epidemiologists use adjusted rates, to give an overall summary estimate after controlling for the variable in question (e.g., age, education or sex).

**EXAMPLE** (Data from Klein RJ, Schoenborn CA. Age adjustment using the 2000 projected U.S. population. Health People Statistical Notes, no. 20. Hyattsville, Maryland: National Center for Health Statistics. January 2001.): Let’s say you wish to compare smoking rates among persons with high education (high school graduate) with rates in those with no high school diploma. First, you go to the National Center for Health Statistics data base and determine that overall, the percent of smokers in those with no high school diploma is 30.5% compared with 28.7% among those with a high school diploma. However, you also learn that smoking prevalence varies by age:

You then determine that the age distributions of the two groups you wish to compare (i.e., those with less than HS education and those with a high school diploma), also differ:

So, the outcome of interest (smoking) is related to age, and the age distributions of the two groups you are comparing are not the same. You need to age adjust the smoking rates in each education group to make a fair comparison.

The goal is to compare the rates across groups, assuming they had the same underlying age distribution. There are 4 steps to this procedure, and the table below illustrates the calculations:
1. Select some standard age structure to which you will adjust the rates of each of the groups you are comparing (In this case, the total US population 2000 COLUMN D).

2. For each group (education level in this example), multiply that group’s age specific smoking rates (COLUMN C) by the number of persons in the standard population who are in that age stratum (COLUMN D). This will yield the "expected" number of smokers in that education group in that age stratum, assuming the age distribution of the standard population (COLUMN E).

3. Sum the expected numbers of smokers across all age groups to get a total (COLUMN E Total).

4. Divide the total number of smokers by the total size of the standard population to get a summary adjusted rate (COLUMN F).

Example of Age Standardization: Smoking Rates by Education Level

<table>
<thead>
<tr>
<th>A Education:</th>
<th>B Age Group</th>
<th>C Observed Smoking Rate (%)</th>
<th>D 2000 US Standard Population Age Distribution</th>
<th>E Adjusted (expected) Number of Smokers</th>
<th>F Age Adjusted Percent Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; HS</td>
<td>18-24</td>
<td>36.8</td>
<td>26,258,000</td>
<td>9,662,944</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25-44</td>
<td>41.1</td>
<td>81,892,000</td>
<td>33,657,612</td>
<td></td>
</tr>
<tr>
<td></td>
<td>45-64</td>
<td>34.9</td>
<td>60,991,000</td>
<td>21,285,859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>13.2</td>
<td>34,710,000</td>
<td>4,581,720</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30.5</td>
<td>203,851,000</td>
<td>69,188,135</td>
<td>33.94</td>
</tr>
<tr>
<td>HS Diploma</td>
<td>18-24</td>
<td>31.4</td>
<td>26,258,000</td>
<td>8,245,012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25-44</td>
<td>35.5</td>
<td>81,892,000</td>
<td>29,071,660</td>
<td></td>
</tr>
<tr>
<td></td>
<td>45-64</td>
<td>28.7</td>
<td>60,991,000</td>
<td>17,504,417</td>
<td></td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>12.2</td>
<td>34,710,000</td>
<td>4,234,620</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>28.7</td>
<td>203,851,000</td>
<td>59,055,709</td>
<td>28.97</td>
</tr>
</tbody>
</table>

In summary, you applied the age specific rates for each education group to a standard population, to estimate the numbers of smokers that would be in each group, if each group had the age distribution of the standard population. The resulting age adjusted rates (COLUMN F) show that after age adjustment, the differences in smoking prevalence are somewhat greater than those observed using the unadjusted rates (COLUMN C total or overall rate).

B. Measures of association in Epidemiologic Studies: How much is X associated with disease?

1. Cohort Studies
1a. Relative Risk: The goal of a cohort study is to determine whether there is an association between the exposure of interest and the outcome studied. Is there excess
risk among those exposed? The most commonly used index of association is the **relative risk (RR)**. Since, in epidemiological jargon, the word “risk” means “rate,” what we are really talking about is the relative incidence rate. Remember, the incidence is the probability, or risk of developing disease. The relative risk is defined as:

\[
\text{(Incidence rate in the exposed)} \div \text{(Incidence rate in the unexposed)}
\]

In the table below, the incidence in the exposed would be \(a/\text{alb}\) and incidence in the unexposed would be \(c/\text{cod}\). Therefore, the Relative Risk would be defined as: \(a/(a + b) \div c/(c + d)\).

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>(a)</td>
<td>(b)</td>
<td>(a + b)</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>(c)</td>
<td>(d)</td>
<td>(c + d)</td>
</tr>
</tbody>
</table>

A relative risk of 1.0 would therefore indicate no difference in risk between exposed and unexposed persons, or no association. If we do a prospective study on the association between exposure to Agent Orange and the development of cancer, and the RR is 1.4, we would interpret that to mean that men exposed to Agent Orange have 1.4 times the rate of cancer of those not exposed; or, alternatively, a 40% increased risk of cancer. If we conduct a study of calcium supplementation and hip fractures and find a relative risk of 0.65 among women using supplements versus those who do not, we would say that women using supplements have 0.65 times the risk of fractures than do non-supplement users, or, alternatively have 35% reduced risk of fractures.

**1b. Attributable Risk**: Another important measure of association is the **attributable risk** (or ‘absolute risk increase’ or ‘absolute risk reduction’ or simply ‘risk difference’). This is a very useful clinical measure. Similar to relative risk, it seeks to describe the relation between two incidence rates: the rate in the exposed population versus the rate in the unexposed population. However, instead of determining the ratio between the two, it simply describes the difference, as:

\[
\text{Attributable Risk} = \text{(Incidence rate in the exposed)} - \text{(Incidence rate in the unexposed)}
\]

Referring back to the 2 x 2 table above, this would be: \(a/(a+b) - c/(c+d)\).

The attributable risk is expressed as a rate: that is, it has units (cases/person/unit time). The RR is unitless: it tells you how much the risk increases (or decreases) *relative* to a baseline risk, but can’t tell you how many people actually will get sick as a result of their exposure. **Attributable risk** provides a tool for weighing the number of excess cases in exposed group that are attributable to the exposure. Sometimes this is expressed as a proportion, or the proportion of the incidence in exposed group that is attributable to exposure.
Referring to the table above, this would be:

$$\frac{(Incidence \ rate \ in \ the \ exposed)}{Incidence \ rate \ in \ the \ exposed} - \frac{(Incidence \ rate \ in \ the \ unexposed)}{Incidence \ rate \ in \ the \ exposed}$$

To illustrate the difference in interpreting relative risk and attributable risk, let’s talk about postmenopausal estrogen use (estrogen therapy, or ET). Please note: the data that underlie the following analysis are based on a long history of observational research in this area. It’s an analysis that would have made great sense – in fact, the sort of analysis that strongly influenced clinical behavior – prior to 2002, when the results of the Women’s Health Initiative (WHI) study were published. We will be looking at that study in depth during one of your case conferences.

Prior to the WHI, while there was some equivocation about the association between ET and breast cancer, the data on endometrial cancer were consistent with a relative risk of about 6.0 (if UNOPPOSED estrogen [that is, without progestin] was used). Observational data on the risk of coronary heart disease (CHD) suggested a relative risk of 0.5. Thus, women on ET were shown to be about 6 times as likely to develop endometrial Ca, and about half as likely to develop CHD, compared to those not on ET. (This is how you interpret relative risk.) Increasing cancer 6-fold should outweigh the benefits of reducing CHD by only half. However, while it may seem that the negative impact is obviously greater, the attributable risk of cancer is actually less than the attributable risk benefit for CHD. This is because baseline incidence rates are so different. For endometrial cancer, a 55 year old white woman’s risk is 62.9/100,000/year (this is the baseline incidence rate for an average 55 year old white woman). If her risk increases 6 fold with estrogen therapy, then it becomes 377.4/100,000 women/year (that is, 62.9 * 6). For CHD, the average incidence rate for a 55 year old white woman is 990/100,000/year. On estrogen therapy, her risk becomes 495/100,000/year (calculated as 990 * 0.5).

Therefore, attributable risk for endometrial Ca: \(\frac{(377.4 - 62.9)/100,000}{year} = \frac{14.5}{100,000/year}\).  
Attributable risk for CHD: \(\frac{(495 - 990)/100,000}{year} = \frac{-495}{100,000/year}\).

For these two diseases, the net effect in 100,000 women each year is 315 new cases of endometrial Ca, compared with 495 cases of CHD prevented. So, in the entire population of 100,000 women, there are 180 more who benefit than who are harmed. Now the hard part -- how do you balance the actual risk versus benefit? Which outcome is more important -- and how can you make this judgment? And, of course, how do you factor in all the other effects of ERT: possible increased risk of breast Ca? Reduced risk of osteoporosis and fractures? Effects on symptoms and quality of life? These are the ultimate clinical decisions that must be made: epidemiologic research provides the data to guide such decisions.
1c. Person-Years

Often in a cohort study, not everyone has equal follow-up time. All subjects are not enrolled on the same day (or even week or year). People drop out at varying times during follow-up, and the outcomes develop at different rates. The problem is: what number do you use for a denominator in calculating the incidence rate? How do you decide the number at risk over the study period? One way to deal with this in the analysis is to assume that all subjects were followed to the end of the study (not optimal, as you might surmise, as you are inflating the denominator). Another choice would be to base the analysis only on those who were followed for the entire study (what is the cost of excluding a portion of the subjects, how are they different that those who stayed in the study?). In order to use information for each person for as long as they contributed to the study, and to account for the precise amount of follow-up included in the incidence rates, person years are used. Person years are the total number of years of follow-up for each person under study, summed across all persons. For example, 10 people followed for 1 year each would be 10 person-years. A study that followed 50 people for 1 year and another 100 people for 6 months would have 100 person years of follow-up. Once person-years are calculated, incidence rates can then calculated as the number of cases that developed during the study divided by the total person-years of follow-up, or the rate per unit person-years.

The figure below shows an example:

Seven persons were enrolled in a study of treatments for prevention of diabetes. 3 incident cases developed during a four year follow-up (2 at year 2 and one at year 3).
person was lost to follow-up (LFU) at year 3 and 3 completed the study without diabetes. Total person years =21. Incidence rate; 3/21 person years, or 14/100 person years.

There is an inherent assumption made when using person-years: all person-years are considered equal. That is, one person followed for 20 years is equated with 5 persons followed for 4 years, or 20 people followed for one year. Depending upon the question being studied, this may or may not be valid.

2. Measures of Association in Case-Control Studies:

In our discussion of cohort studies, we expressed the relation of exposure to disease in terms of the relative risk. In order to calculate relative risk, it was necessary to compute the incidence of disease in the exposed and unexposed groups. However, in a case-control study, we start by selecting those with and without disease. As a result, the proportion of persons with disease is not a function of the risk of developing disease in the general population, but rather is the result of the selection procedure used by the investigator. For example, one might choose 1 control for every case, 2, 3 or even more, with each choice yielding a different proportion with disease. In other words, because of the design, it is not possible to calculate incidence or compute the relative risk in a case-control study.

The measure of association used in case-control studies is the odds ratio (OR), which is an estimate of the relative risk that can be readily calculated from the 2x2 table in a case control study:

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
</tbody>
</table>

Remember that the relative risk is based on a ratio of proportions, or probabilities of disease (among those in the exposed and unexposed groups). To calculate these proportions, all those in the numerator are also included in the denominator. For example, in the table above, the probability of disease in those exposed group would be a / a+b. In contrast, the odds ratio is based on a ratio of the odds of disease in those with exposure to the odds of disease in those without the exposure. Briefly, odds are the number of ways an event can occur divided by the number of ways that it can not. In the table above, the odds of disease in the exposed group would be a/ b. Similarly, the odds of disease in those without the exposure would be c/d. The odds ratio is then defined as a ratio of these two, or a/b ÷ c/d, or ad/bc.

The OR is a mathematically sound estimate of the RR based on data from case-control studies -- so much so that it’s interpreted as if it were the relative risk. So if a case control study were done to evaluate the association between exercise and
coronary heart disease, and the OR was 0.7, we would say that those who exercise have 70% of the rate of coronary disease of those who don’t exercise; or, alternatively, a 30% reduction in risk.

The following hypothetical example illustrates that the RR is not a valid measure of association in a case-control study, while the OR is not affected by the proportions of cases and controls selected. Assume you have a population in which 50% of those with disease X and 10% of those without disease X are exposed to Agent Q. You decide to conduct a case control study and select 100 cases and 100 controls. The results are summarized below:

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Unexposed</td>
<td>50</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

“RR” = \( \frac{50}{60} \div \frac{50}{140} = 2.3 \)

OR = \( \frac{50}{50} \div \frac{10}{90} = 9.0 \)

Now, let’s say that you decide that your budget is larger than expected and you want to increase the size of your study. You select the same 100 cases, but this time, you choose 400 controls and calculate the summary measures.

You can see that the RR is not valid, as it changes with the ratio of cases to controls, which was determined by the investigator. The OR gives the same estimate regardless of the ratio of cases to controls selected.

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Unexposed</td>
<td>50</td>
<td>360</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>400</td>
</tr>
</tbody>
</table>

“RR” = \( \frac{50}{90} \div \frac{50}{410} = 4.6 \)

OR = \( \frac{50}{50} \div \frac{40}{360} = 9 \)

3. Clinical Trials: **Number Needed to Treat**: An important measure, particularly when considering the clinical and/or public health implications of intervening on a risk factor or of a therapeutic intervention, is the **Number Needed to Treat (or NNT)**. This is calculated as the reciprocal of the risk difference or attributable risk (usually, in
therapeutic context, this would be called the “absolute risk reduction”). This will come up in our case conferences this year, and will certainly come up in the setting of your “Evidence-Based Medicine” (or “EBM”) sessions next year. The NNT is typically calculated in the context of a clinical trial.

For example: The Diabetes Prevention Program (NEJM 346(6):393-403, 2002) tested whether treatment with metformin (an antihyperglycemic agent) prevents or delay the onset of diabetes. The risk of diabetes in the placebo group was 28.9/100 over 3 years, while the risk in the metformin group was 21.7/100 over 3 years. So, the attributable risk or risk difference was: 28.9/100/3 years – 21.7/100/3 years = 7.2/100 fewer cases of diabetes in metformin users compared with placebo over 3 years. Therefore, the number needed to treat was: 1/AR or, 1/.072 or 14. This means that in order to prevent one case of diabetes, 14 people need to be treated with metformin for 3 years.

C. Are my estimates of association distorted? (Bias and Confounding)

1. Bias

Bias is defined as any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and risk of disease. It’s important to note here that there is nothing in this definition that refers to intent -- nor should there be. In contrast to the way you and I use this word in everyday English, saying that a study is affected by bias doesn’t imply maliciousness or ill-will on the part of the investigator. While bias is something we try to avoid as much as possible as we design studies, we consider its potential effects without casting aspersions on the investigator, recognizing that the complete elimination of research bias is often unachievable. This issue affects all clinical research. In any study (regardless of the design) potential biases needs to be considered and all efforts should be made to avoid them in the design, and to consider the implications when drawing conclusions.

There are complex classification schemes of bias, and sometimes this nosology may be of help to an investigator planning a study. For our purposes, we should focus our attention to two basic types of bias: selection and information.

1a. Selection bias: Selection Bias refers, in general, to the problem that subjects included in a study may not be representative of the population. “The population” may refer to the universe of cases from which the sample is drawn, or the population to which you’re trying to apply the results of the study.

For example, in a prospective study of cardiovascular risk factors in relation to cognitive decline, elderly participants who volunteer to report for annual clinic visits may be healthier in general than elderly persons in the same community who are unable or unwilling to participate due to health concerns. Thus, the estimated frequency of cardiovascular disease may be reduced in the study population, or the group who participate may all have some protective factor (say a gene?) which has enabled them to survive with intact cognition despite the presence of vascular disease.
Another form of selection bias in observational studies has to do with selection factors which determine who is and is not exposed. For example, earlier population based studies of postmenopausal hormone therapy have shown that women who choose to take hormones long term are in general more healthy than are women who do not. How would this affect the results of a cohort study addressing whether hormone therapy reduces risk of subsequent cardiovascular disease?

Loss to follow-up is a major type of selection bias that is particularly problematic for cohort studies. Bias is introduced if “selection” for loss to follow up is associated with either the exposure or the outcome. For instance, consider a study on Agent Orange. If we have no follow-up data available on a large proportion of Vietnam vets, could this introduce bias? Sure -- if there’s some association between exposure and loss to follow-up (e.g. those exposed were more likely to be ground troops, who tend to come from more transient backgrounds and are thus harder to find) or to outcome (e.g. cancer patients are more likely to be lost to follow-up than healthy people). [exercise: what would be the effect of each of these potential biases?]

Loss to follow-up is not entirely avoidable, but with aggressive and creative approaches to data collection, loss to follow-up can be kept at a minimum. As a rule of thumb, it has been suggested that if fewer than 80% of subjects are available at the end of the study, it should be viewed cautiously. Better than this rule of thumb, an approach can be to recalculate observed rates assuming 'best' and 'worst' case scenarios. For instance, say that there are 1000 hypertensives and 1000 normotensives; at the end of 10 years of follow-up, data are available on 900 hypertensives (who have 180 strokes) and on 950 normotensives (who have 40 strokes). The Relative Risk is 180/900 over 40/950, or 0.2/0.04, or 4.75. We could assume that all the 'lost' hypertensives did not have strokes, while all the 'lost' normotensives did. The new RR = 180/1000 over 90/1000 (2.0). So while the effect is less, even in this worst case scenario, hypertension is still positively associated with the risk for stroke, so the major inference of the study is still valid.

A number of selection biases have been defined as they relate to case-control studies. For instance, in a case-control study, biases may occur if exposed cases are more likely to be selected for study, or conversely if controls are be less likely to be selected if exposed. [An example of the latter was given above, in the discussion of the study of coffee and the risk of pancreatic Ca]. An example of the former is the association between estrogen use and endometrial Ca. Since estrogen use leads to post-menopausal bleeding, and since this symptom leads to testing which can diagnose endometrial Ca, those women taking estrogen are more likely to be diagnosed, not necessarily more likely to get the disease. (This assumes that some undiagnosed endometrial cancer is present in the community from which controls are drawn.) A way to avoid this bias would be through a prospective study in which all subjects underwent the same scrutiny in the search for a diagnosis (e.g. all had yearly endometrial biopsies).]
Another example comes from a case-control (Selby JV, Friedman GD, Quesenberry CP Jr., Weiss NS. A case-control study of screening sigmoidoscopy and mortality from colorectal cancer. *New Engl J Med* 1992; 326:653-7), comparing rates of sigmoidoscopy in those who had died of colon cancer (cases) with those who had not (controls). What do you think the impact of selection bias would be among those who were screened; that is, how are those patients who are referred for or who choose to come for screening likely to be systematically different than those who don’t get screened?

More family history -- hence outcome (cancer) more likely.
More clinical symptoms -- hence outcome more likely.
More informed, healthier behaviors -- hence outcome less likely.

Thus, in this case, it’s difficult to assess overall impact. However, this particular example demonstrates how creative investigators can come up with study designs to help assess and/or minimize the impact of bias. In Selby’s study, the investigators performed two separate analyses: one used the outcome of mortality due to distal colon cancer (those cancers which arise in the distal end of the colon, within the range of the sigmoidoscope, for which outcome could be conceivably influenced by screening); and another analysis used the outcome of death from proximal Ca (beyond the reach of the sigmoidoscope, for which outcome could not be influenced by screening). The biases discussed above would have the same effect in both analyses, since the issue of selection bias shouldn’t be effected by the ultimate outcome. However, in the analyses, screening was associated with lower mortality ONLY in distal (not proximal) Ca, lending substantial strength to the inference that this benefit was due to the effect of screening, rather than to selection bias.

1b. Information bias: Information bias includes a broad range of systematic errors possible in the collection of information. The issue is that information on exposure must be collected evenly or equally well in those with and without disease. Cohort studies are less prone than other designs to information bias, since the information on exposure is ascertained prior to the development of disease. There is no recall required, and the investigator is not aware of who will ultimately become a case. Information biases will be discussed in more detail when we talk about case-control studies.

There is a special case, related to bias, called misclassification. The distinction between bias and misclassification is that bias requires a *systematic* process, and misclassification does not. In other words, misclassification occurs when there is a systematic over or under estimation of exposure or disease status. We distinguish between “non-differential” misclassification (say where we underestimate the disease rate equally in both exposed and unexposed study subjects), and “differential misclassification” (for example where the outcome is ascertained more completely in exposed and unexposed groups). When non-differential misclassification occurs, it
always biases toward the null (that is, it always will make our observed RR closer to 1 than it otherwise would be- think of it as adding random “noise” to your data). Thus, when investigators have a positive result, non-differential classification will actually strengthen (or at least reinforce) the findings -- but it’s hard to be sure our misclassification is non-differential. If differential misclassification occurs, then study results may be biased either toward or away from the null and this is difficult to determine.

One type of information bias particularly relevant to case-control studies is recall bias. For instance, imagine a case-control study seeking to evaluate the possible association between physical trauma and breast cancer. The investigator asks a group of cases and controls the same question: “Have you ever had significant trauma to your breast, enough to leave a bruise or have noticeable soreness the next day?” Note that the investigator is being careful to define what s/he means by ‘significant,’ attempting as much as possible to standardize the results. But imagine the subjects: cases all have received a diagnosis of breast cancer; controls have not. Imagine a case/control pair, both of whom had been hit by a baseball 10 years ago, enough to meet the criterion above. Does it seem likely that the breast cancer patient will be more likely to recall that past trauma than the control (or, equivalently, that the control will be more likely to forget it)?

There are other potential sources of information bias that may arise in a case-control study. Some may be introduced by the investigator (in a manner perhaps more reminiscent of our general use of the word bias). Imagine, in the study above, that the investigator performs all the interviews him/herself, and is really convinced, based on pilot data, that there’s a real association here. It’s conceivable that s/he will probe a little harder for possible exposure in the diseased group. ("You mean you NEVER had ANY trauma to your breast? Not even a LITTLE bit?") This could clearly be a problem -- but the good news is (here’s preventive medicine) that this bias is entirely preventable through BLINDING -- just make sure that whoever’s doing the interview (or record review) doesn’t know who’s a case and who’s a control.

**1c. Estimating the Impact of Bias**

This is where we often get more ‘art’ than science. Are we over- (or under-) representing exposure among cases (or controls)? Are we over (or under-) diagnosing illness among exposed (or unexposed)? We must conceive of differential effects, and think of what kinds of impact these are likely to have on our observations. Sometimes these are quite predictable, and we may conclude, for instance, that our observed RR is likely to be an overestimate of the true RR (if we predict that our biases tend to increase the RR). When biases predictably tend to **reduce** our estimate of effect, that often increases our confidence in the results: “while our observed RR was 2.3, since the biases in this study are likely to have reduced our observed effect, the true RR is likely to be even higher than 2.3.”
Sometimes, there are many potential biases, some of which may tend to overestimate, some underestimate, the effect. This can get complicated. For example, we’d like to determine whether screening sigmoidoscopy reduces risk of colon cancer mortality. [Question for thought: how might you do such a study?]  

2. Confounding  

In all observational (i.e., non-experimental) assessments of association, one must be concerned about the possibility of confounders. A confounding variable is one which interferes with our observed association because of two factors: it is associated with the putative causal factor; and it is associated with the outcome of interest. Confounding can be controlled for in design or analysis.  

Note that by definition, confounders must be associated with both the potential risk factor and the outcome. One is not enough.  

The following graphic representation may help clarify this:  

\[ \text{A} \xleftarrow{C} \xrightarrow{} \text{B} \]  

In this example, we’re wondering if exposure A is truly associated with outcome B. There’s a potential confounder, C -- and since it is associated with A (that’s the squiggly line), and it independently leads to B (that’s the arrow), then it is a confounding variable.  

For example, let’s say we’re interested in the association between oral contraceptive (OC) use and venous thrombosis (blood clots). We might consider that Catholicism might be a confounder, since Catholics may be less likely to use OCs than non-Catholics. However, since there is no relation between Catholicism and the risk for blood clots, this can not be a confounder. Similarly, we may be concerned that aspirin use, which might prevent blood clots, could be a confounding variable. But if there’s no association between aspirin and OC use, then it’s not a confounder. On the other hand if we are interested in the association of smoking and breast cancer, then alcohol use is a confounder, because smokers in general have higher alcohol consumption, and alcohol has been associated (in some studies) with increased risk of breast cancer.  

There are several ways to deal with confounding:  

CONTROL for CONFOUNDING in DESIGN  

Randomization: very effective, but not always practical – for example, it’s hard to randomize people to exercise or not exercise. Also note that randomization does not ensure that there is no confounding. There can still be important group differences by chance which could cause confounding.
**Restriction**: limit the study to, say, only non-smokers. This controls for smoking as a confounding factor, but has a major limitation: we won't know if exercise protects *smokers* from MI (only *non-smokers*).

**Matching**: for every 'exerciser' who smokes, we'd need to enroll a 'non-exerciser' who smokes; likewise for non-smokers. This is cumbersome, and allows opportunities for the introduction of bias, but is very useful in certain cases.

**CONTROL for CONFOUNDING in ANALYSIS**

**Stratification**: we can categorize subjects into strata, based on the confounding variable. Special analytic techniques are available -- and often, the pattern can be clear even without fancy statistics.

For example, here’s the kind of data we might find:

<table>
<thead>
<tr>
<th>Exercise</th>
<th>MI</th>
<th>No MI</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>75</td>
<td>925</td>
<td>1000</td>
</tr>
<tr>
<td>No</td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
<tr>
<td>Total</td>
<td>175</td>
<td>1825</td>
<td>2000</td>
</tr>
</tbody>
</table>

In this mythical study, 1000 exercisers are compared with 1000 non-exercisers. The rate of MI in the former is 75/1000, and in the latter is 100/1000, so the RR is 0.75.

However, what if we stratify by smoking status, and come up with the following data:

**SMOKERS**

<table>
<thead>
<tr>
<th>Exercise</th>
<th>MI</th>
<th>No MI</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>60</td>
<td>340</td>
<td>400</td>
</tr>
<tr>
<td>No</td>
<td>90</td>
<td>510</td>
<td>600</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>850</td>
<td>1000</td>
</tr>
</tbody>
</table>

**NON-SMOKERS**

<table>
<thead>
<tr>
<th>Exercise</th>
<th>MI</th>
<th>No MI</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>15</td>
<td>585</td>
<td>600</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>390</td>
<td>400</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>975</td>
<td>1000</td>
</tr>
</tbody>
</table>

RR=1 (check the math)

In this example, smokers are, overall, 6 times more likely to have an MI than non-smokers (nb: this is an exaggeration of reality), and are 50% less likely to exercise. When the data are analyzed by strata, there is no association between exercise and MI, so our presumed association (based on the overall data) was accounted for entirely by the confounding factor of cigarette smoking.

**Multivariable analysis**: this is a more complex, but somewhat analogous, approach. It is based on certain mathematical assumptions, and the computations required are substantial, and usually done by computer -- thus, there’s the loss of intuitive understanding of the approach that we can have for simple stratification, as just demonstrated. However, these methods (such as logistic regression) are generally considered valid (here again, we defer to our statistician colleagues), and often are the only efficient way to simultaneously control for several confounding variables.
This lecture will provide an overview of basic concepts of screening and the application of screening to clinical practice.

At the end of this lecture you should be able to:

1. Define sensitivity and specificity.

2. Use sensitivity and specificity to calculate positive and negative predictive value.

3. Explain the relationship between prevalence and pre-test probability.

4. Explain the relationship between pre-test probability and post-test probability.

5. Distinguish screening from diagnosis.
BACKGROUND MATERIAL LECTURE III: FROM SCREENING TO DIAGNOSIS

I. SCREENING AND PREVENTION

The act of screening for disease usually falls under the heading of ‘clinical preventive medicine.’ While that’s where it does belong, it’s important to remember that screening does not actually prevent the occurrence of disease. Since the goal is to detect disease early, it doesn’t seek to reduce disease incidence, but rather to reduce disease mortality. There’s an important corollary of this: unless early intervention affects outcome, there’s no reason to screen. That’s why we screen for some things (hypertension, breast cancer), don’t screen for some things (lung cancer), and don’t know what to do with some other things (prostate cancer).

This is one of the criteria that helps define a useful screening test: that early intervention makes a difference. It’s actually the most important criterion. Others include other characteristics of the disease (it should have substantial prevalence, and important [negative] clinical consequences) and of the test used for screening (it should be accurate, acceptable to patients, and cost-effective).

In preventive medicine, it’s traditional to categorize activities as primary, secondary, or tertiary prevention. The categorization goes like this:

PRIMARY PREVENTION: prevents a pathologic process from occurring (e.g. vaccination, universal salt restriction);
SECONDARY PREVENTION: after pathologic process has begun, but before symptoms occur (screening fits in here); and
TERTIARY PREVENTION: rehabilitation or arresting progression in patients with established disease.

While this scheme is traditional, and may even find its way onto board exams, it’s not terribly useful. Particularly problematic is the fact that some actions fit into this pattern with substantial fluidity. Take the treatment of hypertension. Telling the general population to reduce its sodium intake is an attempt at primary prevention of hypertension -- but (as we’ll discuss further during the second part of this course) hypertension isn’t really a disease, but a risk factor for CAD, stroke, and death. Thus, treating hypertension is often referred to as primary prevention of stroke or CAD. However, treating a hypertensive with CAD is referred to as secondary prevention of MI, and reducing blood pressure after an MI is, probably, tertiary prevention of MI (or, since it reduces risk of a second MI, could be secondary prevention). You see the point here.

A more useful categorization is between clinical prevention and public health interventions. As physicians, you will be spending most of your time involved with clinical preventive medicine: you’ll be screening for CAD risk factors and for cancers, you’ll be administering immunizations, you’ll be trying to promote smoking cessation.
But you’ll also have a role to play in public health, or population based approaches. Perhaps you’ll choose a career in public health. But even if you don’t (and most of you won’t), your role as a physician will make you a powerful ally in the population-based approach. Public health campaigns against smoking, promoting vaccination, encouraging seatbelt and helmet use, discouraging drug use -- all of these are more effective when physicians counsel their patients and reinforce the message. So even here, the overlap in ‘classifications’ is substantial.

A. How do we know that a screening test is effective?

The problems we have identified in the context of study design are relevant here. While we’ve said that a screening test is only useful if its use leads to improved outcomes, the question is, how can we know if that’s the case? It turns out that there’s only one way to Know (with a capital K), and that’s if we do (you guessed it) a randomized clinical trial. However, there are only a few screening tests for which we have good RCT evidence that screening leads to reduced mortality (mammography in women over 50 reduces breast cancer mortality; stool occult blood tests in men and women over 50 reduces colon cancer mortality), and that is about it.

1. Bias in the Evaluation of Screening: We’ve talked about bias before, and the art of predicting the impact (and, most importantly), the direction of the impact of biases. There are three important biases to consider which limit our ability to make reliable inferences from observational studies of screening tests. Importantly, all these biases tend to work in the same direction: to make screening look good, even if it’s not.

These biases are:

1a. VOLUNTEER BIAS. People who choose to come for screening tests tend to be more health conscious, and have better health outcomes. (Note: for this one, we have to equivocate on the direction of effect, because sometimes people choose to be screened because they’re at higher risk of disease.)

The usual impact of volunteer bias can be seen in these data from the HIP mammography trial (an RCT of mammography):

<table>
<thead>
<tr>
<th></th>
<th>Breast Ca</th>
<th>All other causes</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40-49</td>
<td>50-59</td>
<td>60-69</td>
</tr>
<tr>
<td>Control</td>
<td>2.4</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Exp</td>
<td>2.5</td>
<td>2.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Subgroups -- volunteered</td>
<td>42</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>[about 1/3] refused</td>
<td>77</td>
<td>38</td>
<td></td>
</tr>
</tbody>
</table>

The first two lines of the table show the efficacy of mammography (this is from the RCT). The second two lines compare two groups within the experimental group: those who came for the mammograms (e.g. were compliant with the recommendations), and those who did not. The numbers speak for themselves.
1b. LEAD TIME BIAS. This refers to the fact that even if you don’t alter the natural history of a condition, diagnosing it earlier means you’ll live longer with it. Of course, what we want to do to increase your survival after diagnosis is make you live longer -- but the same effect will follow if we just make the diagnosis earlier.

1c. LENGTH BIAS. It turns out that not all people with disease have the same severity. In cancer, for example, some tumors are aggressive, and grow rapidly, and some are more indolent. Since your likelihood of being diagnosed through screening depends on how long your tumor is present in ‘preclinical’ stage, the slower growing tumors are more likely to be diagnosed through screening.

2. Test performance and interpretation

A major clinical skill is the ability to accurately interpret the results of screening, and diagnostic, tests. This seems so basic that one wouldn’t have to even say this -- but the sad fact is, most docs don’t do this too well. Many do, however, and you can too -- there’s a rather simple conceptual (and numerical) basis to this that will help you.

In general, we tend to refer to our test results as “positive” or “negative.” This is an oversimplification: there really aren’t any clinical tests that are truly dichotomous. Everything depends on measurement and subsequent categorization. But operationally, we do need to dichotomize our results, because our goal is to use the results to guide decision making.

Once we’ve categorized our results as positive or negative, we need to determine how well these results predict ‘the truth.’ This approach is somewhat similar to our discussion of alpha and beta error, but with an important difference: here, when we refer to the ‘truth,’ it’s not an unknowable metaphysical concept, but a reasonable diagnosis based on our predetermined standard. This is often referred to as a “reference standard” or a “gold standard” test. When this is done, we end up with a 2x2 table:
### THE TRUTH (based on the gold standard)

<table>
<thead>
<tr>
<th>Your diagnosis (based on the test in question):</th>
<th>DISEASE PRESENT</th>
<th>DISEASE ABSENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST POSITIVE</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>True positive</td>
<td></td>
<td>False positive</td>
</tr>
<tr>
<td>TEST NEGATIVE</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>False negative</td>
<td></td>
<td>True negative</td>
</tr>
</tbody>
</table>

The accuracy of a test is measured by both its sensitivity and specificity.

**SENSITIVITY**: proportion of individuals with disease who have a positive test;

**SPECIFICITY**: proportion of individuals without disease who have a negative test.

Thus, referring to the table above, sensitivity is $a/(a+c)$, or $[true\ positives]/[true\ positives + false\ negatives]$; and specificity is $d/(d+b)$, or $[true\ negatives]/[true\ negatives + false\ positives]$. Perhaps you're more comfortable trying to memorize these formulae, but years of experience with students suggests that that will ultimately be confusing. No matter what field in medicine you pursue, you'll be using this knowledge (as well as that relating to predictive values, below) -- so an appropriate learning strategy is to understand what these terms mean, rather than memorize the formulae.

Sensitivity and specificity are properties of the test itself. The important feature of these terms is that since they are proportions, they focus your attention on the denominator. Since sensitivity is defined as the proportion of individuals with disease who have a positive test, the denominator is: individuals with disease. Similarly, the denominator for specificity is persons without the disease. Therefore, interpreting sensitivity and specificity requires some knowledge of who has the disease. Usually, if we knew that, we wouldn’t do the test!

What we want clinically is a number that will help us evaluate the likelihood of disease for a given test result. If the test is positive, how likely is it that the person has the disease? If the test is negative, how likely is it that the person does not have the disease? These kinds of questions are answered by predictive values, defined as:

**PREDICTIVE VALUE OF A POSITIVE TEST** (or PV+): what proportion of positive tests are correct?

**PREDICTIVE VALUE OF A NEGATIVE TEST** (or PV-): what proportion of negative tests are correct?

The concept is simple: while we consider sensitivity and specificity to be properties of the test itself, the predictive values depend not only on the test but on the population tested. More specifically, the predictive value of a test depends not only on sensitivity and specificity, but also on disease prevalence in the population being tested.
More specifically yet, we know that, as disease prevalence decreases in a population, PV+ goes down, and PV- goes up. Conversely, as disease prevalence increases, PV+ goes up, and PV- goes down. This is because prevalence is related to the probability that an individual in the population has the disease: the lower the prevalence, the less likely that any given person has the disease. It has to do with the “pre-test probability” of disease. You can see this yourself, without needing to use Bayes’s theorem, by making 2x2 tables.

As an example, let’s look at the Papanicolaou (Pap) test to screen for cervical cancer. This is a widely accepted screening strategy that is often credited for the dramatic decline in cervical cancer mortality in the US during the latter half of the 20th century. The focus of this discussion is not to question its utility (though some authors have raised this question, in the absence of randomized controlled trials to support its efficacy), but rather to assess its accuracy in clinical application.

For any test, determining the sensitivity and specificity requires research that meets certain methodological standards. Most importantly, all subjects should have both the test in question, and the “gold standard” test performed. These tests should be independent, and the assessment of one should be blinded to the assessment of the other. For Pap testing, the gold standard can be provided by colposcopic biopsy, though there may be some debate about what pathologic finding should define a gold standard positive result. Without dwelling on this (I’m not an expert, and you’ll learn more about this next year), there are estimates of test performance given in the Guide to Clinical Preventive Services, 2nd Ed., written by the US Preventive Services Task Force. NOTE: this is a wonderful reference resource, and as a government funded document, it’s available to all of you on the web, at http://odphp.osphs.dhhs.gov/pubs/guidecps/default.htm

From the chapter on screening for cervical cancer, we can find the following estimates for the accuracy of the Pap test, and the prevalence of cervical cancer:

SENSITIVITY = 0.70
SPECIFICITY = 0.90

PREVALENCE is actually much harder to estimate. The Guide only tells us that there are approximately 16,000 cases diagnosed each year. But what’s the denominator? If this includes all adult US women, the denominator could be about 105,000,000 (current US census for women 18+ years), for a prevalence rate of 0.00015 (or 15/100,000). Of course, many women are NOT screened, so a more meaningful prevalence estimate is probably twice that. For a first look, let’s imagine the prevalence is as high as 1 per thousand (0.1%). Here’s how we can calculate the predictive values:
<table>
<thead>
<tr>
<th></th>
<th>Cervical Ca</th>
<th>No Ca</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pap +</td>
<td>70</td>
<td>9,990</td>
<td>10060</td>
</tr>
<tr>
<td>Pap -</td>
<td>30</td>
<td>89,910</td>
<td>89,940</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>99,900</td>
<td>100,000</td>
</tr>
</tbody>
</table>

Positive Predictive Value = 70 / 10,060 = 0.007
Negative Predictive Value = 89,910 / 89,940 = 0.999

Despite the fact that specificity is relatively high, more than 99% of all positive tests are false positives. This makes sense, since with such low prevalence, the PV+ will be low – since so few of the women screened actually have the disease, and most positive tests are false positives. But it can sometimes be surprising – even upsetting – to see just how dramatic this effect can be, as in this example.

But what if we were to consider the accuracy of the Pap test not as a screening test in an asymptomatic population, where the prevalence is only 0.1%, but in a population of women at extremely high risk. Imagine doing Pap tests in a group of women who’ve never been screened, smokers (yes, smoking is implicated in this cancer, too!) of low SES who have had multiple sexual partners and infection with human papilloma virus (HPV). Let’s imagine that this targeted strategy identifies a population with a very high prevalence of 10% (note: this is just for illustration purposes, not real).

The new table would look like this:

<table>
<thead>
<tr>
<th></th>
<th>Cervical Ca</th>
<th>No Ca</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pap +</td>
<td>70</td>
<td>90</td>
<td>160</td>
</tr>
<tr>
<td>Pap -</td>
<td>30</td>
<td>810</td>
<td>840</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>900</td>
<td>1,000</td>
</tr>
</tbody>
</table>

PV+ = 70 / 160 = 0.44
PV- = 810 / 840 = 0.96

Note: we’ve kept sensitivity and specificity constant, but only changed prevalence. (The way to do that with this 2x2 table approach is to fill in the bottom line first [I like to always put “100” as the total for the left-hand column] to reflect prevalence, and then use sensitivity and specificity calculations to fill in the boxes in the table.)

Predictably, with fixed sensitivity and specificity, as prevalence goes up, positive predictive value goes up, and negative predictive value goes down. In this case, if the test is negative, the “pre-test” probability of cancer (10%) is reduced to 4% – perhaps enough to take no further steps. A positive test still indicates a less than 50% chance of cancer – but given a positive test, the change from a 10% “pre-test probability” to a 44% “post-test” probability is probably enough to warrant further steps.

Now, of course, further work is required. The screening test leads to further diagnostic or therapeutic actions, and the value of screening depends on the outcomes.
of these actions. Thus, screening is not necessarily good -- there may be downsides. (Can you think of possible downsides in this scenario?) But the problem is translating this into clinical practice: how does this play out for an individual woman? How do you assess risks and benefits, and how do you balance these? When we talk about “quality adjusted life,” what does that mean? Whose quality? Once again, we can see that the job of the physician involves quantitative and factual knowledge, but also depends on judgment. Here’s the ‘art’ of medicine, and part of the frustration, joy, and challenge of medical practice.
Lecture IV: Principles of Scientific and Statistical Inference

This lecture will provide an overview of basic concepts of scientific reasoning, and basic statistical concepts required to critically analyze the medical literature.

At the end of this lecture you should be able to:

1. Outline criteria for making causal inferences.
2. Define sampling variability
3. Define measures of variability: standard deviation and standard error and describe the difference between the two.
4. Explain hypothesis testing.
5. Define type I (alpha) and type II (beta) error and their relation to study power.
6. Explain the difference between 1-tailed and 2-tailed tests of significance.
7. Outline the factors to be considered when computing sample size for a study.
BACKGROUND MATERIAL LECTURE IV: BASIC PRINCIPLES OF SCIENTIFIC AND STATISTICAL INFERENCE

I. The Philosophical and Conceptual Basis of Research

The goal of any clinical or epidemiologic research project is to make inferences from the population under study that may apply to the broader population. Imagine if all that the investigator could learn about were the subjects of the study themselves: that would hardly be useful, since anyone reading the report is unlikely to care for those particular patients! We must be able to extrapolate the findings on a limited sample to a broader population if the results are to have any utility at all. Thus, we must understand how such extrapolations can be reasonably and fairly made.

II. Scientific Inference, Deduction, and Induction

You've probably learned elsewhere about “deductive” versus “inductive” reasoning. A simplified (and admittedly simplistic) approach is to consider as “deductive” truths of logic and mathematics. If we know that all men are mortal, and that Socrates is a man, we can DEDUCE that Socrates is mortal. This must be the case -- logic does not permit it to be otherwise.

However, while this works well for truths of reason, our knowledge of the physical world comes through experience. In worldly matters, we induce general truths based on a careful interpretation of empirical evidence. We move from specific observations to general principles. This approach seems to work for us, some of the time -- but we all know of cases where it doesn't. Let's say you come for your first visit to New York City. You stop someone on the street to ask directions, and he's absurdly rude. The next person you stop won't even answer you. So you conclude, “All New Yorkers are rude.” But this isn't true, is it? After all, you're a New Yorker, and you're not rude! Clearly, inducing general truths from specific observations is shakier than deducing specifics from general truths.

In fact, some out-of-towners seem to dearly hold the ‘truth’ that all New Yorkers are rude. Once this is accepted as a true postulate, then it is a deductive inference that if you're a New Yorker, then it must be true that you are rude. Therefore, everything you say sounds rude. This specific example is broadly reflective of a common problem: we make generalizations based on empirical evidence (this is induction -- and an important, reasonable way to make inferences about the world); we then internalize the generalization, and accept it as a real ‘truth’ (this is psychology -- and a mistake); and now we use it to predict what will happen in the future (this would be deduction, if the principle were actually true).

In clinical and epidemiologic research, we make inductions based on evidence -- and how to do that properly is the stuff of this course. We then apply those general principles to specific situations, but with a sophisticated psychology: we learn to
recognize that these “truths” are simply our best theory at the moment, and we’re willing to change as new evidence becomes available. This, of course, is true of all science: while the Bohr model of the atom adequately predicts much of what we need to know in chemistry, as we learn about subatomic particles and quantum mechanics, we reject it in favor of a more complex model. But most of us (I suspect) still have this image in our minds of protons and neutrons being orbited by electrons -- and to the extent that gets us what we need, that’s ok!

III. Falsification versus Affirmation

When using empirical observation to make inductive inferences, we have a much greater ability to falsify a principle than to affirm it. This was laid out in the late 1950s by the philosopher Karl Popper, who illustrated the point with this now-classic example: if we observe swan after swan, and each is white, we may infer that all swans are white. We may observe 10,000 swans, all white, and feel confident in our inference. All of this evidence may make us think that we can prove that all swans are white. However, all it takes is a single observation of a non-white swan to disprove the assertion that all swans are white.

Based on this theoretical underpinning, our contemporary approach to statistical inference takes the following form: we have a scientific theory; we create a null hypothesis, which essentially says that our theory isn’t correct; and then we attempt to disprove it. This process often seems contrived and confusing -- as if we’re building a straw man and then knocking him down -- but is in fact logically mandated by the Popperian view described above. All of our current statistical theory, used in hypothesis testing, is based on this approach. The ‘P-value,’ which you must understand, and about which more will follow, really just gives us a mathematical probability (hence “P”) that we’d find the difference observed or a more extreme difference if the null hypothesis is true. The lower that probability, the greater the confidence we have in saying that we have falsified the null hypothesis.

IV. Causal Inference

By means of the type of reasoning just described (empirical observation, inductive inference), we may conclude that there is a real association between two variables; say, A is associated with B. Often, what we really want to know is, does A cause B? Here’s an example of this problem: in February 1997, a paper published in the New England Journal of Medicine seemed to demonstrate that people are more likely to be involved in car crashes if they’re using their cellular phones. So a big question is: does the use of cell phones cause car accidents? [If so, this could have significant policy implications.]

In thinking through how you make scientific inferences, we refer to ‘principles,’ rather than to rules. To advance our understanding beyond “A is associated with B” to make the assertion that “A causes B,” we have several principles to guide us:
1. TEMPORAL ASSOCIATION. The cause must precede the effect. (Of all the principles of causal inference, this is the only one that’s logically REQUIRED.)

   [In the cell-phone example, the phone calls can only cause an accident if they happen BEFORE the accident. This is particularly relevant here since, as the investigators point out, the number of phone calls immediately increases AFTER an accident (e.g. calling for help).]

2. STRENGTH OF THE ASSOCIATION. The stronger the association, the more likely it is to be causal.

   [In this case, the relative risk of an accident is 4.3, indicating that motorists are more than 4 times as likely to have an accident when they’re using their cell phones than when they are not. This is more likely to be a causal connection than if the relative risk were, say, only 2 -- but less likely than if it were 8.]

3. DOSE-RESPONSE RELATIONSHIP. The greater the exposure to the potential cause, the greater should be the risk of the outcome.

   [In this case, being on the phone at the time of the accident is an all or none variable -- so the issue of dose-response is not relevant. But if our putative risk factor were drinking and driving, we might see that the more you drink, the greater the likelihood of an accident.]

4. CONCEPTUAL PLAUSIBILITY. If A really causes B, it should make sense that A could cause B.

   [It is reasonable to speculate that talking on a cell phone could reduce reaction time, decrease concentration, limit dexterity, etc. (in fact, earlier research in driving simulators demonstrated that such effects occur), so a causal connection is plausible. If, for instance, it were found that people in red cars are more likely to be in accidents, it’s only remotely plausible that the fact that the cars are red causes the accident -- rather, that there’s something about people who drive red cars that makes them more prone to crash.]

5. CONSISTENCY. If the association is causal, it is more likely to show up in numerous studies (not just one), even in different populations, and with different study designs.

   [In the cell-phone example, there are no other studies that demonstrate this association -- and two (industry-sponsored) studies that show no such association -- so this criterion is not met.]
The traditional list goes on from here to include the concept of “specificity”: that is, that every time the cause is present, the effect follows; and every time the effect is observed, it had been preceded by the cause. Since this almost never occurs in biologic systems (as it does, for instance, in physics), most contemporary thinkers have dropped this ‘criterion’ from the list. Even if talking on car phones does cause accidents, it would be neither a necessary (e.g. you could still have another car run into you head-on when you’re not on the phone) nor a sufficient (e.g. if you’re on the phone you still may not have an accident unless it’s foggy and the roads are slick and some other driver swerves) condition.

It is an interesting aside to note that several “anti-car phone” ordinances have been passed since the publication of this paper. These laws have often referred to this paper as justification, along with the “face validity” concept that says, “of course it’s dangerous to drive while you’re talking on the phone.” One of many examples of the conflict between personal liberty and public health.

II. Basic Statistics

OK, here it comes -- we’re going to work with numbers! For many of you, this won’t be a stumbling block -- the pre-med process requires enough mathematics that you’ve all had a reasonable exposure. But experience suggests that a substantial proportion of you may retreat to the stance: “oh, I’m just not good with numbers.” Sorry, that’s no excuse. Imagine instead a doctor saying, “Well, perhaps my patient notes are meaningless, but I’m just not good with words.” (Actually, having read many student and house staff workups, perhaps some think this is a reasonable stance.) While we understand that illiteracy is common, it cannot be accepted in a doctor. Likewise, innumeracy is unacceptable. Think about what doctors do. Essentially, every action can be thought of in terms of either determining etiology (with its implications to prevention); making diagnoses; estimating prognosis; or prescribing therapy. All of these actions are quantitative, since they involve an understanding of rates and a balancing of probabilities. The modern practice of medicine requires physicians who are both literate and numerate.

A. Sampling

In statistical analysis, our goal is to use a limited number of observations and make reasonable inferences to a larger population. As indicated above, if we could only apply the results of medical research to the research subjects themselves, there would be precious little we could do with the information. The utility of a study’s findings is directly proportional to our ability to extrapolate its results to another clinical population.

In this type of approach, we recognize that it is rarely possible to include in a study the entire population of interest. For instance, we may want to determine the health impact of radon exposure -- but we can’t study everyone who’s ever been
exposed to radon. What we must do then is study a reasonable sample of people who have been exposed – and not exposed – to radon.

1. Probability sampling is a way of drawing a sample from a population in such a way that the probability of any unit in the population entering the sample is known. A starting point for this type of sampling is a list enumerating the population; this list is known as the sampling frame. Individuals are then selected singly from this frame by a random mechanism such as a table of random numbers. If after a unit is selected for the sample it is returned to the population, this is known as random sampling with replacement. The advantage of such a scheme is that if there are N units in the sampling frame, the probability of selection for each unit remains 1/N every time a unit is selected. The disadvantage of this method is its potential wastefulness in that the same unit can be included in the sample on different draws. To overcome this disadvantage, a modification is to not replace a unit after it has been selected from the population. This type of scheme is known as random sampling without replacement. The complication here is that at the first draw each unit has probability of 1/N of inclusion in the sample; at the second draw the remaining units have probability of 1/(N-1) of being included in the sample, etc.

2. In both schemes discussed above, the user is depending on the random mechanism to provide a representative sample on other characteristics (such as age, sex, race) of individuals in the population. If it is important that a particular subgroup or subgroups be included in the sample, or that they be included in proportions different from how they appear in the general population, then the investigator can insure this by performing stratified random sampling. This is accomplished by stratifying the sampling frame into subgroups and then doing random sampling, with or without replacement, within each stratum.

3. A systematic sample is illustrated by the following example. Suppose we want to obtain a 5% sample from a given population. We would first enumerate the population to define the sampling frame. We would use a random mechanism to generate a number between 1 and 20, say 6. Then the 6th individual in the sampling frame would be the first individual included in the sample. Then we would select every 20th individual from the sixth for sample inclusion, i.e. 26th, 46th, 66th, etc. Systematic sampling has two advantages over simple random sampling. First, it is easier to perform, since only one random number is required; and second, it distributes the sample more evenly over the listed population. It may have a built-in stratification, if we assume the population is enumerated by strata. Systematic sampling often gives substantially more accurate estimates than simple random sampling, and it has become a popular sampling technique. However, it has the disadvantage of having no reliable method of estimating the standard error of the sample mean (a concept that will be explained further below). Also it has the potential disadvantage of being badly biased if the population list contains periodic variation that occurs on the same wavelength as the interval between sampled units.
4. A method to be avoided is the haphazard sample or the sample of convenience. As an example, suppose an investigator were interested in studying the population of homeless persons housed in shelters in New York City and wanted to draw blood samples for laboratory analysis. If only some homeless were asked if they would give blood, and only half of those approached consented to participate, then this group would be a sample of convenience. There might be any number of selection factors operating so that the sample would be unrepresentative of the larger population.

An unhappy truth: in most clinical and epidemiologic studies, we have convenience samples, not random samples. If we’re looking at the health impact of radon exposure, we might pick a geographic area, send a questionnaire to homeowners, and study those who respond to the questionnaire and are willing to have their homes tested and are tested by the study team during the enrollment period and have adequate testing samples. Rarely would the investigators have a list of every homeowner, randomly select people from the list, and then aggressively enroll everyone randomly selected. This is particularly problematic when we realize that most statistical tests we use are predicated on an assumed random sampling approach. The distinction between the ideal and the practical is quite stark here.

B. Types of Variables

CONTINUOUS VARIABLES are those that theoretically can take an infinite number of values along a continuum within a particular range. For instance, blood pressure is often recorded as, say, 120/80, the 120 indicating that the systolic pressure is about 120 mm Hg. There is a 0 digit preference because clinicians often don’t feel compelled to be too precise; and an even digit preference because that’s how the scale is marked on the sphygmomanometer. But it could be 121, or, for that matter, 121.3579 mmHg, if we had an instrument which could measure the pressure that precisely. For continuous variables, it’s appropriate to calculate an average or mean value for a group.

DISCRETE VARIABLES can not take on all these intermediate values. These may be numerical (such as number of students in a med school class: there may be 176 or 177, but can’t be 176.32). Some variables are naturally broken into categories (hence are categorical); these may be dichotomous (e.g. sex) or multichotomous (e.g. race); they may be naturally ordered (ordinal, such as grades [A, B, C]) or non-ordered (nominal, such as blood type [A, B, AB, O]). For these variables, groups are usually described by proportions with a particular characteristic.

There is a valid distinction to be made between continuous variables and numerical discrete variables (e.g. serum cholesterol, which is continuous, and numbers of live births, with is discrete [you can have 1 or 2 live births, nothing in between]). However, here again the theoretical and the applied are somewhat different. Discrete variables should be analyzed differently than continuous variables, but in practice, numerical discrete variables tend to be handled as continuous variables in the medical
literature. This is the kind of problem that probably does [and should] upset statisticians--but as physicians, you can’t get too worked up about this as you critically read the medical literature. As you’ll see, there are bigger things to worry about.
DATA REDUCTION AND PRESENTATION. It is common to take one type of variable and transform it into another. This is reasonable, because we humans need to "lump" data in order to understand them. One convenient way to deal with a morass of numbers is to "categorize" them (that is, break them up into categories). This would, in effect, transform a continuous variable into a categorical variable.

I may be interested in knowing, for instance, whether pulse varies with age. In Table 1 (see next page), data are presented from a real study of 178 people with acute myocardial infarction, and age has already been broken down into categories (since age is truly a continuous variable, and in this dataset was calculated from birthdates, we had 178 separate age groupings before they were categorized). It’s difficult to make sense out of this table, except maybe to note that certain pulse rates seem to come up more frequently than others. [QUESTIONS FOR THOUGHT: which are the most common pulse rates? Why are these most common? Does this represent physiology, or something else? (Think about preference for recording specific numbers)]
### TABLE 1. Pulse rate by age

<table>
<thead>
<tr>
<th>PULSE</th>
<th>AGE &lt;55</th>
<th>AGE 55-64</th>
<th>AGE 65-74</th>
<th>AGE 75+</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>53</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>55</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>57</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>61</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>62</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>63</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>66</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>68</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>71</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>72</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>74</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>75</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>76</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>78</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>79</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>82</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>83</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>84</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>86</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>88</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>89</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>91</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>92</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>93</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>94</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>96</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>98</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>104</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>106</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>107</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>108</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>112</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>120</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>18</td>
<td>28</td>
<td>51</td>
<td>81</td>
<td>178</td>
</tr>
</tbody>
</table>
Note that the above inferences on the distribution of the pulse rates completely ignore any potential relation between age and pulse. But it’s difficult to look at this complex table and make much sense of how these two variables relate. To do that, we’ll have to simplify the data, by categorizing the observations. But before we do, let’s follow along the question about the distribution of pulse rates. In Table 1, we looked at the last column to see patterns in pulse distribution. Perhaps we can get more out of this by presenting the data in graphic form: a histogram.

**Distribution of Pulse Rates (1)**

Squeezing each individual observation into this graphic does allow us to clearly see the MODAL values: it’s striking how a pulse of ‘80’ is recorded far more than any other. But there are also several missing values -- and while the curve sort of looks like a ‘normal distribution,’ or bell-shaped curve, it’s hard to be sure. We may see this more clearly if we categorize the pulse rates: 45-49, 50-54, 55-59, 60-64,....,120-124. Here’s the histogram that follows (next page):
By grouping the observations by sets of 5 bpm, we are smoothing out the curve a bit, and it’s easier to see the pattern. Note how the distribution in this histogram is somewhat less symmetric than in the earlier one -- but since every value has a bar associated with it, it’s a more complete representation of the data. So data reduction can serve an important function.

Now, let’s get back to the association between pulse and age, as presented in Table 1. As we noted earlier, there’s too much data to make real sense of here. In Table 2, the same data have been grouped into 4 categories of pulse rates. Clearly, it’s easier to begin to see some patterns here. Starting with pulse itself, we see (consistent with the histograms above) that the extremes (bradycardia, tachycardia) are less common than the middle categories. [In contrast to the earlier question re: Table 1, this is likely to represent real physiology.] Also, it begins to look like older people may have, on average, higher pulse rates.

**TABLE 2. Pulse rate by age**

<table>
<thead>
<tr>
<th>PULSE</th>
<th>AGE &lt; 55</th>
<th>AGE 55-64</th>
<th>AGE 65-74</th>
<th>AGE 75+</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 60</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>60-79</td>
<td>9</td>
<td>13</td>
<td>25</td>
<td>30</td>
<td>77</td>
</tr>
<tr>
<td>80-99</td>
<td>5</td>
<td>7</td>
<td>13</td>
<td>38</td>
<td>63</td>
</tr>
<tr>
<td>100+</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>TOTAL</td>
<td>18</td>
<td>28</td>
<td>51</td>
<td>81</td>
<td>178</td>
</tr>
</tbody>
</table>
But it’s still difficult to clearly see the patterns here. So more simplification may be in order. Why not dichotomize both sets of data: age greater than or less than 75 (which looks like it’s about the median) and pulse greater than or less than 80 (likewise). This gives us a new table, using the same data we started with:

**TABLE 3. Pulse rate by age**

<table>
<thead>
<tr>
<th>PULSE</th>
<th>AGE &lt; 75</th>
<th>AGE 75+</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 80</td>
<td>58</td>
<td>33</td>
<td>91</td>
</tr>
<tr>
<td>80+</td>
<td>39</td>
<td>48</td>
<td>87</td>
</tr>
<tr>
<td>TOTAL</td>
<td>97</td>
<td>81</td>
<td>178</td>
</tr>
</tbody>
</table>

Here the association between age and pulse rate is rather clear, especially if you take the important next step of calculating proportions. For instance, 48 out of 81 “elderly” subjects have “higher” pulse rates, compared with 39 of 97 “younger” subjects: 59% versus 40%. Is this a “statistically significant” association? We’ll learn about this later. Is it meaningful? That’s a different question. Even if a difference is statistically significant, it may not be meaningful, either because the difference is too small to have an impact on patients (that’s a clinical decision), or the study design has substantial flaws (that’s an issue in clinical epidemiology). Both may apply in this case. First, does it really matter if older people have a pulse rate a few beats higher than younger folks? Are there any health implications of this? Second, did this study include an appropriate sample to ask this question? Actually, while these are real data, they come from a study looking at treatment in patients with acute myocardial infarction, a study not designed at all to look at age and pulse. (I just used these data for teaching purposes, looking for continuous variables that could be readily categorized.)

Back to data transformation. It is important to recognize that such categorization is arbitrary, and hence leaves a lot of room for "fudging". Therefore, the categories should be selected for SOME logical reason (e.g. convention, clinical intuition, previous studies); and they should NOT be selected based on the appearance of the data themselves. The categorization scheme used above is arguably clinically reasonable (pulse categories could be considered as bradycardia, low normal, high normal, or tachycardia; age as very old or younger than that), but also may have been influenced by the appearance of the data. The best way to ensure against such “fudging”: make your categorization decisions *a priori* in designing your research projects.

An advantage of categorization is that it enables you to visually present your data in useful ways, such as tables (frequency distributions), bar charts (histograms), and line charts (frequency polygons). However, while categorization can ease the interpretation of continuous data, there is some judgment that is allowed to enter the picture here. In addition, there are always some data that are "lost in the translation"; the original dataset contains detail which tends to be obscured. We would like a way to describe such data without arbitrary categorization -- and this is where we look for C. Measures of Central Tendency.
MEAN
MEDIAN
MODE

You’ve probably all heard these terms before (see glossary).
The mean, remember, is the average. If we add all the pulse observations, and
divide the total by 178 (the number of observations) we get the mean (in this case,
79.3). The median is the number for which half the observations (that is, 89
observations or half of 178) are higher, and half are lower. In this case, the median is
78 (check it out for yourself and see if that’s right). The modal value, that value which
appears most frequently in the data, is clearly demonstrated in the first figure: 80.

In a situation like this, when the mean, median, and mode are so close, we have
a symmetrical distribution, such as the so-called ‘normal,’ or ‘Gaussian,’ curve.
Sometimes they’re not too close to each other -- a so-called ‘skewed’ distribution.

In descriptive (and, ultimately, analytic) statistics, there is often a question of
which measure of central tendency is the most informative. The mean is most widely
used, since it is most widely understood, includes all the measures, and is most
amenable to statistical testing. It is, however, very sensitive to “outliers” (extreme
measurements on one side or the other, as are often found in skewed distributions).
When outliers are a problem, the median is preferred. Imagine that I have two groups
of 5 people, and compare their incomes. In group I, the incomes are
$25,000, $27,000, $29,000, $31,000, $33,000

so the mean and median are both $29,000.

In group II, the incomes are
$25,000, $27,000, $29,000, $31,000, $175,000

In this case, the median is still $29,000, but the mean is now $57,400. Deciding which
measure is more informative is a matter of some judgment -- which number best
describes the “average” income in this group? In general, we tend to prefer the median,
because it tells us the most about the greatest number of individual observations. For
instance, in group II, 4 out of 5 people have incomes very close to the median. If we
described group II by the mean income, we’d have the funny situation that 4 out of 5
people earn substantially less than the mean. But the choice of the right measure also
depends on the goals of our description. If we’re trying to pool our resources, obviously
the mean tells us more: there’s a lot more money in group II than in group I, even
though the medians are the same.
D. Measures of Variability

One way to help deal with the fact that very different data sets can have the same mean value is to consider how much variability exists around the mean. What we’d like is a measure that will tell us how much each individual observation varies from the mean. We could sum up each individual difference:

\[ \text{obs1} - \text{mean} + \text{obs2} - \text{mean} + \text{obs3} - \text{mean} + \ldots + \text{obsN} - \text{mean} \]

but this sum will always be equal to 0 (you may confirm this yourself if you like).

To get around this problem, statisticians have developed the measure you all know, the standard deviation (SD). The SD is calculated by squaring the difference between each observation and the mean, summing, and dividing this sum by \((n-1)\) [Why \((n-1)\) and not \(n\)? Well, it isn’t that important, but the idea is that if you know the mean, any \((n-1)\) of the observations can then determine the value of the remaining observation. You have therefore lost a “degree of freedom” – a somewhat confusing term you will sometimes encounter in tables in medical journal articles or outputs from statistical software. Obviously, with a lot of observations, the difference is minor.]

An important feature of the SD, in normally distributed data, is that about 95% of the sample will be contained within the interval

\[ \text{mean} \pm 2\text{SD}. \]

About 2/3 of the observations are contained within 1 SD of the mean, and almost all (>99.7%) are within 3 SD of the mean. Again, this is a characteristic of the normal distribution. The trouble is, often data don’t follow a normal distribution. Real data tend to have more extreme observations, and sometimes the extreme observations are more in one direction than the other. For that reason it is a good idea to verify whether your data are normally distributed before selecting a statistical test.

There are other measures of variability that are often used when non-normal distributions are described (particularly when the median is the measure of central tendency). The most useful is the interquartile range: the range within which the middle half of the observations occur. Since the median value is the 50th %ile, the interquartile range is given by [25th %ile - 75th %ile].

E. Analytic Statistics: Hypothesis Testing

1. Statistical significance

Statistical analysis usually focuses on answering the following question: are the differences that we observe between two (or among several) groups greater than what we would expect by chance alone, given the intrinsic variability of our observations? Going back to our earlier discourse on the philosophy of science, we will invoke the null hypothesis.
Please bear in mind, before you read the next few paragraphs, that this whole thing may be confusing at first. Such phrases as “null hypothesis,” “disproving the null hypothesis,” “failing to disprove the null hypothesis,” may seem unnecessarily cluttered by double- and triple-negatives. However, these usages were not coined just to confuse the hapless medical student. They are quite logical, and with time will enhance your real understanding of the phrase “statistically significant” (a phrase you will hear, and use, throughout your medical career, so you might as well understand it).

As we had previously outlined, the process of falsification is more robust than affirmation in empirical research. Statistical inference is based on the assumption of the null hypothesis (that two groups are not different), and, using statistical theory and the data in hand, estimates the probability of finding a difference as large or larger than that observed in the study. We are seeking to DISPROVE the null (since we can only falsify, not affirm); thus, a POSITIVE study (which finds a statistically significant difference) is one that DISPROVES the null hypothesis. Further, a NEGATIVE study (one which does not find a statistically significant difference) is one that FAILS TO DISPROVE the null hypothesis (note, it never proves the null hypothesis). Under the assumption that the null hypothesis is true, the probability of finding a difference as large as that in your data, or greater, in the direction of the alternative hypothesis, is given by the \( P \) value. For most data analysis problems in which the sample size is fixed at the beginning of the study, a lower \( P \)-value represents stronger evidence against the null hypothesis and we speak of it as showing greater statistical significance.

What \( P \)-value is “statistically significant”? The conventional level of 0.05 is simply that -- convention, not an absolute. It has nothing to do, directly, with the fact that the mean \(+2\)SD includes 95% of the observations in a sample. However, both relate to a similar convention -- that we tend to view 95% certainty as pretty good. (See confidence intervals, below.) And the 0.05 level has become so well established that to use anything else makes it difficult to compare results across studies.

Now, the null hypothesis doesn't have to be the hypothesis we attempt to falsify, but it is a convenient one to use, since there is only one null, but an infinite number of other choices. Thus, if we are comparing Group A with Group B, our null is that A=B. We might choose to disprove, instead, that A=2B, that A=B/2, that A=B+4. You see the point. Not only is there only one null, but it’s also intrinsically of interest to determine whether two groups are “the same” versus “different.” So there’s an inherent advantage to the null hypothesis as our “straw man.”

In hypothesis testing, an important statistical decision is the choice of performing a \textbf{one-tailed test} or a \textbf{two-tailed test}. One way to think of this is to decide whether the research question is: are the 2 groups \textit{different}? Or: is the outcome measure of interest in one group \textit{more} (or \textit{less}) than the other? We can make this distinction by stating our \textbf{alternative hypothesis}. Once we decide to attempt to disprove the null hypothesis, there are only 3 choices of alternative hypotheses: 1. A is not Equal to B.
2. \( A > B \).
3. \( A < B \).

Using alternative hypothesis (1) is called a 2-tailed test. Either (2) or (3) is directional, and is called a 1-tailed test. In medical research, a 2-tailed hypothesis test is almost always preferred, because (a) it’s harder to find a statistically significant difference [so if we find a statistically significant difference, it’s more likely to be real], (b) it is rarely the case that only one direction of effect is possible or important (e.g. any drug which can have a beneficial effect can always conceivably have a detrimental effect), and (c) the 2-sided 0.05 significance level has become the accepted level for significance, which makes comparisons across studies easier.

The unavoidable fact in medical research is that, no matter how carefully we design our studies, our conclusions may be wrong. Remember, we’re only looking at a sample of the whole population; and we’re lumping individual observations, with their inherent intra- and inter-individual variability, into groups. The beauty of statistical analysis is that it allows us to come up with reasonable estimates of how likely it is that we’re wrong, given the data at hand and assuming a well-designed study.

We’re usually concerned about 2 possible errors, called (not surprisingly) Type I (or \( \alpha \)) error and Type II (or \( \beta \)) error. A Type I error is the error of saying that the groups are different when in fact they’re not. Can you figure out what number tells you the probability of making this error? Since a Type I error can be rephrased as saying that the null hypothesis is false when it’s in fact true, you can see that this probability is given by the P-value.

A Type II error is the error of saying that the groups are NOT different when in fact they are. This is, essentially, the problem of failing to reject the null hypothesis when it’s in fact false. Remember, it’s harder to affirm than to falsify, so the probability of a Type II error is usually greater than that of a Type I. The following 2x2 table should help clarify this:

<table>
<thead>
<tr>
<th>YOUR CONCLUSION FROM THE STUDY:</th>
<th>THE TRUTH!</th>
</tr>
</thead>
<tbody>
<tr>
<td>there IS a difference (+ study)</td>
<td>there IS a difference</td>
</tr>
<tr>
<td>ok (the study is correct)</td>
<td>Type I (( \alpha )) error</td>
</tr>
<tr>
<td>Type II (( \beta )) error</td>
<td>ok (the study is correct)</td>
</tr>
<tr>
<td>there is NO difference (- study)</td>
<td></td>
</tr>
</tbody>
</table>

The probability of a Type II error is given by the value of \( \beta \). This is also related to another statistical measure, called power. Power refers to the probability that your study will find a real difference (that is, if a difference actually exists). You may see that this concept represents the complementary probability of a Type II error: that is,
Power = 1 - β

Another important concept: power is directly proportional to the sample size of the study. Thus, the smaller the study (other things being equal), the greater the chance of a Type II error. This relationship is of major importance in critically evaluating a study that purports to show a negative result. Since we can never prove the null hypothesis, our failing to disprove it raises the possibility of a Type II error, so our first question is: was the study large enough to detect an important difference that might exist?

In fact, all these features (alpha, beta, effect size, and sample size) are interrelated:

In fact, all these features (alpha, beta, effect size, and sample size) are interrelated:

\[ \alpha \rightarrow N \rightarrow \beta \rightarrow \Delta \rightarrow \alpha \]

where Δ is the magnitude of the difference you’re looking for. Knowing any three of these values, you can calculate the fourth. These interrelationships are extremely important to consider in the planning stages of a research project, particularly in determining the sample size required for a study. Usually, the investigator specifies α and β (often relying on convention to provide levels of 0.05 and 0.20, respectively), then specifies Δ, based on clinical considerations: how large a difference would be clinically important enough to plan the study to be able to detect?

2. Statistical tests

There are a whole host of statistical tests used, and, as practicing physicians, you’re unlikely to need to actually do the calculations involved in any of them. (If you become a researcher, you will want to take a statistics course in which you learn these and many additional methods!) What you will do is utilize the fruits of such analysis (e.g. interpret P-values and logistic odds ratios). While we understand that most of you will not need to perform statistical analyses, we would like to briefly introduce you to the basic concepts that underlie the two most widely used statistical tests: the t-test and the chi-square test. Both are used when two groups are being compared; the t-test is used for continuous variables, and the chi-square test for categorical variables. Another commonly used test is the correlation coefficient, used to test the association between two continuous variables, which we will briefly discuss as well. There are many other statistical tests you will come across in your reading, and much thought goes into the choice of the appropriate test to use, but for the most part, you will be at the mercy of the authors and journal editors to determine that the appropriate tests were done.

(a) The chi-square (χ²) test
The chi-square test is used when an investigator wants to test the null hypothesis that the proportion with a particular characteristic is the same in each of two groups. (Actually, the chi-square test can be used with more than two groups, but we will discuss here only the special case of two groups.) This testing takes advantage of a well-known principle of conditional probability: that the probability of two independent events both occurring is the product of the individual probabilities. If we put this mathematically, we can say that

\[ P(A \& B) = P(A) \times P(B). \]

For example, if I pick a card, the probability that it will be a spade is 1 in 4 (there are equal numbers of cards in each of 4 suits), and the probability that it will be an ace is 1 in 13 (there are equal numbers of cards in each of 13 denominations). Therefore, the probability that it will be the ace of spades is \(\frac{1}{4} \times \frac{1}{13} = \frac{1}{52}.\) (Of course, you can immediately see that this must be correct: since there are 52 cards, and only one is the ace of spades, then the probability of picking it must be \(\frac{1}{52}.\))

How does picking cards relate to statistics? Because if we are going to assume that the null hypothesis is true (as we do in all statistical testing), then we can infer that the probability of two variables coinciding is the product of the individual probabilities. Let’s go back to the 2x2 table we made earlier (remember age and pulse rate?):

<table>
<thead>
<tr>
<th>TABLE 3. Pulse rate by age</th>
</tr>
</thead>
<tbody>
<tr>
<td>PULSE</td>
</tr>
<tr>
<td>80+</td>
</tr>
<tr>
<td>&lt; 80</td>
</tr>
<tr>
<td>TOTAL</td>
</tr>
</tbody>
</table>

These numbers contain information we need to calculate probabilities. For example, what’s the probability (in this sample) of being younger than 75 years old? (answer: 97/178, or 0.54) What’s the probability of having a pulse less than 80 beats per minute? (answer: 91/178, or 0.51) And the real crux of this comparison: what proportion of younger (age < 75) patients have lower pulse rates (<80 bpm), compared to the proportion of older patients with lower pulse rates? (answer: 0.60 versus 0.41) It’s this last difference for which we’re asking the question: is this observed difference more than we’d expect by chance alone?

If we assume that age and pulse rate are independent, then we can say that the probability of being BOTH younger than 75 AND having a pulse less than 80 should be 0.51x0.54, or 0.28. That is, this is the proportion we would EXPECT under the assumption of independence. How many people, therefore, would we expect in the first cell of the 2x2 table (that’s the one that currently contains the number 58)? (answer: 0.28x178, or 50) We OBSERVE 58, not 50, people. What proportion of the sample does this represent? (answer: 58/178, or 0.33)
Therefore, we expect, under the assumption of independence (that is, assuming the null hypothesis), that 28% of our sample would be younger than 75 and have a pulse under 80. We observe that 33% of the sample fits this bill. OK, so this is more than we expect -- but is this measly 5% difference within the range of variability that we might expect by chance alone? Answering this question, my friends, is the heart of classical statistical inference! How much do our observations differ from our expectations, and does the magnitude of this difference exceed that which we would expect by chance alone? Here’s where we are at the mercy of the statisticians to develop a mathematical approach to allow us to determine this, and I for one am comfortable with that. Here’s where they’ve given us the chi-square test. Essentially, the chi-square statistic is calculated with a rather simple formula:

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

where O is each observed frequency in the table, and E is each corresponding expected frequency.

So, continuing our above example, we can see that the table we’d EXPECT (keeping the row and column totals, or the ‘marginal frequencies,’ fixed) is:

<table>
<thead>
<tr>
<th>PULSE</th>
<th>AGE &lt; 75</th>
<th>AGE 75+</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 80</td>
<td>49.59</td>
<td>41.41</td>
<td>91</td>
</tr>
<tr>
<td>80+</td>
<td>47.41</td>
<td>39.59</td>
<td>87</td>
</tr>
<tr>
<td>TOTAL</td>
<td>97</td>
<td>81</td>
<td>178</td>
</tr>
</tbody>
</table>

You may note that once you’ve entered the first number into the first cell, all the rest are determined (given fixed marginals). This is what is meant, in this case, by “1 degree of freedom.”

Our chi-square statistic would be calculated as:

$$(58 - 49.59)^2/49.59 + (33 - 41.41)^2/41.41 + (39 - 47.41)^2/47.41 + (48 - 39.59)^2/39.59$$

or

$$1.426 + 1.708 + 1.492 + 1.787 = 6.413$$

[By the way, there’s a simpler computational formula for chi-square with 2x2 tables. Labeling the cells a, b, c, and d as in Table 7 (see notes for Lecture II, page 42), the formula is:

$$\text{Chi-square} = \frac{(ad-bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

where N is the total sample size, or a+b+c+d. So, in this example, taking the numbers from Table 3, we get]
Do the arithmetic, and you should get the same answer as before (minus rounding error).]

Now comes the part where we have to defer to the statistical theorists. They tell us that for any particular P-value (and with reference to a particular number of degrees of freedom), there is a critical value of chi-square. For our 2x2 table, the number of degrees of freedom is 1; and for our conventional level of statistical significance (P=0.05) the critical value of chi-square is 3.84. So, to determine if our observations represent a statistically significant difference in proportions between the two groups (at P < .05), we ask: “does our calculated value of chi-square exceed the critical value?” Since it does (6.39 > 3.84), we can say that the difference is statistically significant (P < 0.05, or the probability that a difference this large would have occurred if the null hypothesis were true is less than 5%).

Now, after all of this, we can say that 60% of patients under age 75 have a pulse less than 80 bpm, compared with 41% of patients age 75 or more, and that this difference is statistically significant. Please remember -- this doesn’t mean it’s important! Statistically significant and clinically important are two completely different concepts. Perhaps it is, perhaps it isn’t -- and since this discourse is primarily concerned with the statistics, I'll leave the clinical significance for you to decide.

ANOTHER EXAMPLE: BASED ON A QUESTIONNAIRE FILLED OUT DURING MCFM (CLASS OF 2005) ON DIET AND EXERCISE

Please note -- we will use this example several times, and this will be recognizable by the use of italics.

Questionnaires were administered during MCFM, and 126 members of the Class of 2005 filled them out.

Questions for thought: given your class size of about 175 students, is this a good response rate? Is this sample representative of 20-26 year old Americans? of all medical students? of all AECOM medical students?

We were interested in seeing if there was an association between students’ sex and the number of times students report that they’d previously been on a weight loss diet. While some of you may have a hypothesis about such an association, our null hypothesis is that there is NO association.

The data were a provided in the next table (note: only 106 out of 126 students answered both these questions meaningfully (in fact, 3 answered the “sex” question, but put “other” instead of male or female – which either demonstrates a
common problem in clinical research [accuracy and completeness of data], OR that this class is more “diverse” than most):
TABLE 6. Sex and past attempts to lose weight: Class of 2005

<table>
<thead>
<tr>
<th>Sex</th>
<th>0 or 1</th>
<th>2 or more</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>39</td>
<td>18</td>
<td>57</td>
</tr>
<tr>
<td>Women</td>
<td>15</td>
<td>34</td>
<td>49</td>
</tr>
<tr>
<td>TOTAL</td>
<td>54</td>
<td>52</td>
<td>106</td>
</tr>
</tbody>
</table>

What do these data show? First, to look at them intelligently, you need to think about the proportions being compared. Don’t just run to do a chi-square -- you won’t know what it means! One way of looking at it: among men, 18/57, or 32%, have tried to lose weight 2 or more times, compared with 34/49, or 69%, of women in the class.

Now we can do the chi-square to see if this difference (32% vs. 69%) is statistically significant. The chi-square is 15.07, for a P value of 0.0001.

How do you interpret this finding? Also, please note that while one of these 2 variables being assessed is by its nature dichotomous (sex), the other is not, and was ‘reclassified’ in order to turn into a categorical variable. This would lead to a desire to compare means:

(b) the t-test

We started with an approach that involved the comparing of proportions -- in this case, proportions with a particular pulse rate which had been somewhat arbitrarily categorized. It’s not necessarily true that a pulse greater or less than 80 is an important cut-off. We may prefer to compare the average pulse rate in each group, to determine if the average pulse rate is significantly higher in younger than in older patients.

As previously discussed, the average can be described in several ways, most notably the mean, median, and mode. The modal values are not usually amenable to statistical testing, and, besides, they’re generally the least informative. Median values can be, and are, compared using statistical methods (as you read the clinical literature, for instance, you’ll see reference made to the Wilcoxon rank-sum test, which is one way statistician compare medians). Most commonly, still, statistical comparison is made between group means -- and the most common test used for such comparisons (again, when two groups are being compared) is the t-test. It is also known as r Student’s t-test, named for the pseudonym of W.S. Gosset, who published this technique in 1908. The reason he had to publish under a pseudonym is that he was working for the
Guinness Brewery in Ireland, which didn’t want its employees to publish anything because they were afraid of leaking trade secrets to their competitors!

In the t-test, we again compare the difference we observe with some measure of what we expect. In this case, our expectation is a function of the variance associated with our measures -- or more specifically, the **standard error of the mean (SEM)**. The standard error is similar to the standard deviation, but there’s an important difference. In *descriptive* statistics, we’re concerned with how much variance exists in our sample, and the standard deviation tells us that. As indicated above, if we have a mean and standard deviation from a normal distribution, we can infer that about 95% of the sample will fall in the interval described by the mean ± 2 sd.

In *analytic* statistics, our goal is not simply to describe the sample we’ve studied; rather, it’s to make inferences to the larger population from which this sample was drawn. Assuming random sampling (see discussion above), we are able to make inferences about the true population *mean* based on our particular sample *mean*. (An easy way to remember that standard error relates to inferences about the *mean* is to remember the phrase “standard error of the mean.”) Thus, if you drew an infinite number of samples and plotted the resulting sample means, the result would approximate a normal distribution centered on the true (unknown) population mean, and the mean ± 2 se would include 95% of the observed sample means. In statistical theory, this is related to the Central Limit Theorem.

The important take home concept: standard deviation is useful for describing the distribution within a sample; standard error is useful for making inferences about the true population mean based on a sample mean.

It’s also instructive to know that the standard error is given by the formula:

\[ se = \frac{sd}{\sqrt{n}}, \]

from which you can see the importance of the sample size in determining the standard error. Essentially, n comes into the formula twice: once in calculating sd, and again to calculate se. The larger the n, the smaller the se (for a given amount of variance). That makes sense, of course: if we have a larger sample, then our sample estimates are more precise estimates of the true population values of that sample statistic (say the mean or a proportion). In other words, the se gives us an estimate of the variability of a sample statistic, because recall, we are always using data from our sample to estimate the true (and un-measurable) population values. One of the main uses of the standard error is to construct confidence intervals.

Confidence intervals provide information regarding how confident we are that our sample estimates are good estimates the true population values of a given parameter. For example, a 95% confidence interval is estimated as the mean ± 2 se. The strict definition of a 95% confidence interval is that if you repeated the study an infinite
number of times, this interval will contain the true value of the population mean 95% of the time.

Let’s see how this plays out in the case of our data about pulse rate and age. Our question can be framed as follows: is the average pulse rate different among patients younger than 75 compared with those 75 years or older? Our null hypothesis:

\[ H_0: \text{the average pulse rate is the same for younger and older patients} \]

We then statistically compare the mean pulse rates in both groups, measuring the difference between the mean values compared with a pooled estimate of the standard error, and compare the resulting value (the t statistic) with a critical value for t (for a specific P-value, and with a specific number of degrees of freedom, in this case n-2, or 176 degrees of freedom). Just as we did earlier with the chi-square test, if our calculated t statistic exceeds the critical value, we say that the difference is statistically significant at the pre-specified level.

In this case, the mean pulse rate in patient under age 75 was 77.3 ± 1.62 (se), and in those over age 75 it was 81.6 ± 1.57. The t-statistic calculated from these data is -1.8743, which, with 176 degrees of freedom, corresponds to a P-value of 0.0625. This does not meet our prespecified level for statistical significance -- so the difference is NOT statistically significant. That is, we have failed to reject the null hypothesis in this analysis.

\[ \text{STUDENT QUESTIONNAIRE DATA: for this analysis, we used information provided by students in the class about their ‘actual’ and ‘ideal’ weight. We defined “Weight Difference” as current weight (in pounds) minus ideal weight, as stated by the students. We compared mean Weight Difference among students who had ever tried to lose weight with those who never had tried to lose weight. Data follow:} \]

<table>
<thead>
<tr>
<th></th>
<th>MEAN WEIGHT DIFFERENCE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever tried to lose weight</td>
<td>4.16</td>
<td>2.86</td>
</tr>
<tr>
<td>Never tried to lose weight</td>
<td>7.96</td>
<td>1.97</td>
</tr>
</tbody>
</table>

The t calculated for this is \(-1.09, P=0.28\). WHAT DOES THIS MEAN?

\( (c) \) correlation

In the preceding analyses, we were comparing two groups with respect to either proportions (the chi-square test) or means (the t-test). In these examples, we were interested in the relation between adolescent and young adult body weight. Classifying students as either ‘underweight’ or ‘not underweight’ adolescents, we compared both proportions currently overweight (the chi-square test) and mean current BMI (the t-test).
Sometimes, however, you may be interested in knowing whether two continuous variables are related to each other -- and this is usually referred to as ‘correlated,’ since we're not really saying that one is ‘dependent’ and the other ‘independent,’ but just that they each relate to the other. In this case, we are simply assessing association -- no causal link is implied, or should be inferred. The most common measure of correlation is the correlation coefficient (often called the Pearson correlation coefficient). This measure, $r$, varies from +1 (perfect positive correlation) to -1 (perfect negative correlation); a value of 0 means no association between the two variables. This is graphically displayed in the figures below:

$\begin{align*}
\text{r=+1} & \\
\text{r=-1} & \\
\text{r=0} & 
\end{align*}$

STUDENT QUESTIONNAIRE DATA:
We looked for a correlation between 2 continuous variables: students’ BMI (based on self-reported height and weight), and the number of times each student had been on a weight-loss diet in the past. The correlation coefficient (and associated P-value) is: 0.294 (P=0.0021).
FOOD FOR THOUGHT: Is this a strong or weak correlation? What does the P value mean? Does this suggest that going on a weight-loss diet makes you gain weight?
F. Statistical Considerations in Clinical Trials

We’ve already discussed the issue of Type I and Type II errors, and both must be considered in the design and interpretation of an RCT. The important point here is that the probability of a beta error is dependent on 3 things: the alpha level; the effect size; and the sample size. All of this must be considered in performing your "power analysis" before beginning your RCT.

Let’s say you have developed a new drug for dyspepsia (indigestion, heartburn) under the name regurganot, which you wish to test. Your experience leads you to believe that the "natural history" of dyspepsia will lead to a 20% response rate, while the drug will yield a 60% response rate. If you were to observe these rates with 10 subjects randomized to each group, what would you find?

RESPONSE
Yes  No
Drug  6   4   10
Plcbo  2   8   10
Chisquare= 1.875 (p=0.17)

If the same rates applied, but there were 20 patients per group, we’d see:

RESPONSE
Yes  No
Drug  12  8   20
Plcbo  4  16  20
Chisquare= 5.104 (p=0.02)

Thus, a larger sample size gives a more statistically significant finding with the same proportionate findings.

What if we were interested in a smaller real difference: e.g. 40% vs 20%?

RESPONSE
Yes  No
Drug  8  12  20
Plcbo  4  16  20
Chisquare= 1.7 (p=0.30)

As previously discussed, it’s harder to find statistical significance with smaller differences (with a given sample size). With huge sample sizes, however, even tiny difference that would never be clinically significant may turn out to be statistically significant and it is important to bear that in mind when interpreting any data analysis.

Please note: here we are only talking about the endpoint of drug efficacy. We must also be concerned about possible side effects. That’s because any drug which
has sufficient physiologic power to have a potential beneficial effect must also have potential detrimental effects. Medicines can hurt you; when medical treatment leads to harm, it’s called *iatrogenesis*.

Remember that side effects are often rare: for instance, chloramphenicol, a potent antibiotic, was only found to cause fatal aplastic anemia after it was in widespread use. This side effect occurs in about 1 of 20,000 patients. Some have stated that for a study to be large enough to detect a side effect, the sample size should be 3 times the reciprocal of the rate -- so you see that that may be difficult to achieve in clinical trials. That’s why continuing to collect data after a drug is in widespread use (post-market surveillance) is important.

Another statistical consideration is that of early termination of trials. For ethical reasons, it’s important to allow the possibility of stopping a trial early, in case the drug turns out to be clearly beneficial (or harmful). To do this, you have to analyze the data periodically to see what’s happening, and this raises a statistical problem. Similar to the problems of performing multiple comparisons in studies, you must be wary of the increased possibility of chance effects when you look at the same data many times. If your alpha level is set at 0.05, and you then proceed to do 50 separate comparisons, the chance that one of them will end up with a statistically significant finding is actually substantially greater than 5%. So if you take multiple looks at the same data, you should pre-specify when those will be done and how the analysis will be adjusted to control for the multiple looks. The methods to do this are complicated and you should definitely consult a statistician early on in your design if you ever decide to lead such a study.
GLOSSARY OF TERMS

ADJUSTED RATES: Rates which have been statistically manipulated to allow comparison of groups that may differ in their background distribution of important co-variables. An example of this: age-adjustment to compare mortality rates in two populations where the age distribution differs. Note that adjusted rates are useful for comparison only: the rates themselves are fictitious, and should not be used as if they truly described the rate in the particular population.

ALPHA ERROR: The probability of erroneously concluding that an observed difference, or association, is real. (see P-value; Type I error)

ATTRIBUTABLE RISK: The risk (or incidence rate) of disease among subjects exposed to a risk factor, minus the risk (or incidence rate) in subject not exposed. Used in cohort studies; often used to determine the clinical or public health impact of an exposure.

BETA ERROR: The probability of a study failing to detect a difference (or association) when a real difference (or association) exists. (see type II error)

BIAS: a systematic error in which data collected in a study deviate from or do not represent those of the underlying population, such that they may lead to erroneous conclusions.

BLINDING: A strategy to avoid bias, in which subjects, investigators, or both are unaware of important clinical information.

CASE-CONTROL STUDY: an observational retrospective study where cases (those with the disease) and controls (a comparable population without disease) are compared with respect to past exposure to a potential risk factor.

CAUSAL INFERENCE: Inferring that an observed association truly represents a causal connection: not only is A associated with B, but A causes B. Such inference is strengthened if the time order is correct; if the association is strong; if it’s consistently seen; if there’s a dose-response relation; and if it’s biologically plausible.

CLINICAL PREDICTION RULES: mathematically (statistically) derived functions, often expressed as algorithms, that allow a reasonable estimate of various outcomes based on specific clinical factors.

CLINICAL TRIAL: an experimental study where the investigator assigns (usually randomly) an exposure (usually a treatment) to a study group, and no exposure (usually placebo) to another group.

COHORT STUDY: an observational study where individuals at risk for a disease outcome are evaluated for exposure to a specific risk factor, and followed over time to
determine the incidence of disease with and without the risk factor—also called a prospective study.

CONFIDENCE INTERVALS: If you repeated a study an infinite number of times, 95% of the estimated 95% confidence intervals will contain the “TRUE” population value of the measure you are estimating.

CONFOUNDING VARIABLE: a variable that may either attenuate or exaggerate an association between a risk factor and an outcome. Such a variable must be directly associated with the risk factor (exposure) in the study, and independently associated with the outcome.

CROSS-SECTIONAL STUDY: an observational study in which the presence of several factors is assessed at a single point in time. Difficult to infer a temporal sequence or causal connections, but prevalence can be determined.

CRUDE RATE: The overall rate in the entire population- not adjusted for any other factors (e.g., age, sex, education level).

DEDUCTION: The process of inference that moves logically from general principles to specific observations. For instance, if all patients with tuberculosis are infected with *Mycobacterium tuberculosis*, and if Mr. Jones had TB, then he must be infected with *Mycobacterium tuberculosis*.

ECOLOGICAL STUDY: comparison among different populations (e.g. different countries or states). Problem is the *ecological fallacy*: although a risk factor and an outcome may appear to be associated in an ecological comparison, there’s no guarantee that there is any association at all within individuals at risk for the outcome.

INCIDENCE: The number of new cases of a disease (during a specific time period), divided by the population at risk (during that time period). Usually referred to as ‘incidence rate’; expressed as a proportion per unit time. You CAN NOT estimate incidence in a case-control study.

INDUCTION: The process of inference that moves logically from multiple observations to a general principle (empirical). For instance, if every patient you’ve ever seen with lung cancer was a cigarette smoker, you might induce that all cases of lung cancer occur in smokers.

INTENTION-TO-TREAT: The strategy of data analysis in clinical trials which keeps research subjects in the group to which they were initially [randomly] assigned, even if they cross over to other treatments.

LEAD-TIME BIAS: The appearance of increased survival following early diagnosis, even if treatment is ineffective. If a disease is diagnosed earlier, the time from diagnosis to death will increase even if the natural history is not altered.
LENGTH BIAS: The appearance of benefit from screening programs due to the fact that indolent cases of disease are more likely to be picked up through periodic screening, while aggressive cases of disease would become clinically apparent between screenings.

MEAN: Arithmetic average; sum of observations, divided by the number of observations.

MEASURE OF ASSOCIATION: The method chosen by the investigator to measure the strength of an effect, or association. Examples: relative risk; attributable risk, odds ratio. The measure used depends on the design of the study. (For example, The relative risk cannot be estimated in a case-control study, as incidence is not available)

MEDIAN: 50th percentile. The observation for which half the population has a higher value and half has a lower value.

META-ANALYSIS: a systematic review of similar studies, often involving a pooled analysis of data.

MODE: The most frequently observed value in a population.

NEGATIVE PREDICTIVE VALUE: The probability of not having a disease, given a negative test result.

NESTED CASE-CONTROL STUDY: A case-control study in which cases and controls are drawn from within a prospective study. All cases who developed the outcome of interest during follow-up are selected and compared with a subgroup of the non-cases. Exposure is defined prior to disease development based on data collected at baseline or on assays conducted in biological samples collected at baseline.

NULL HYPOTHESIS: The statement that assumes that the groups you will be comparing (and in which you’ll be looking for differences) are basically the same (that is, do not have the differences you’re looking for).

NUMBER NEEDED TO TREAT (NNT): a useful clinical measure that estimates the number of patients who must be treated to achieve a single desired effect. Calculated as 1/[attributable risk].

ODDS RATIO: the odds of exposure to a potential risk factor among subjects with a specific disease outcome, divided by the odds of exposure among those without the disease. Used in case-control studies; interpreted as [and closely approximates] the relative risk.
P-VALUE: The numerical probability that a difference as large as that observed in your study could occur by chance alone, if the null hypothesis is true. The smaller the P-value, the less likelihood of observing your finding under the null hypothesis, and thus the greater the ‘statistical significance’ of the finding. (In other words, the greater confidence you have in rejecting the null hypothesis.

POSITIVE PREDICTIVE VALUE: The probability of having a disease, given a positive test result.

POWER: The probability of finding a difference when, in fact, one exists. (Power = 1 - β)

PREVALENCE: The number of existing cases of a disease (at a particular time), divided by the population observed (at that particular time). A proportion; dimensionless. Often referred to as ‘prevalence rate,’ though not a true rate.

RELATIVE RISK: the risk (or incidence rate) of disease among subjects exposed to a risk factor, divided by the risk (or incidence rate) in subject not exposed. Used in cohort studies; often used to provide clues to etiology.

SENSITIVITY: The probability of having a positive test result in a population with a particular disease.

SPECIFIC RATE: The rate in a particular portion of a population (e.g. a sex-specific incidence rate of hypertension in women would be the number of new cases of hypertension in women divided by the number of women at risk for developing hypertension [i.e. without hypertension at the beginning of the study period]).

SPECIFICITY: The probability of having a negative test result among people who do not have the particular disease in question.

STANDARD DEVIATION: A measure of the average variation of individual observations from the mean. Useful for describing the spread of data in the study sample. Calculated as the square root of [the sum of squared differences from the mean, divided by n-1 (where n is the number of observations)].

STANDARD ERROR: A measure that allows calculation of how much your observed mean from your sample is likely to differ from the true population mean. Useful for comparing groups, calculating confidence intervals. Calculated as SD (standard deviation) divided by the square root of n (the sample size).

STATISTICALLY SIGNIFICANT: The difference observed in the study, after statistical testing, is judged to exceed what would be expected by chance variation. The conventional significance level of P < 0.05 (see P-value) is purely convention.
TYPE I ERROR: The probability of erroneously concluding that an observed difference, or association, is real. (see P-value; alpha error)

TYPE II ERROR: The probability that a study will fail to find a difference (or association) when a specific difference (or association) actually exists. (see beta error)