NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS – SPOTTING THEM AND AVOIDING THEM
Part I: Fundamentals

MAY 24 – 27, 2010

Instructor: Martha K. Smith

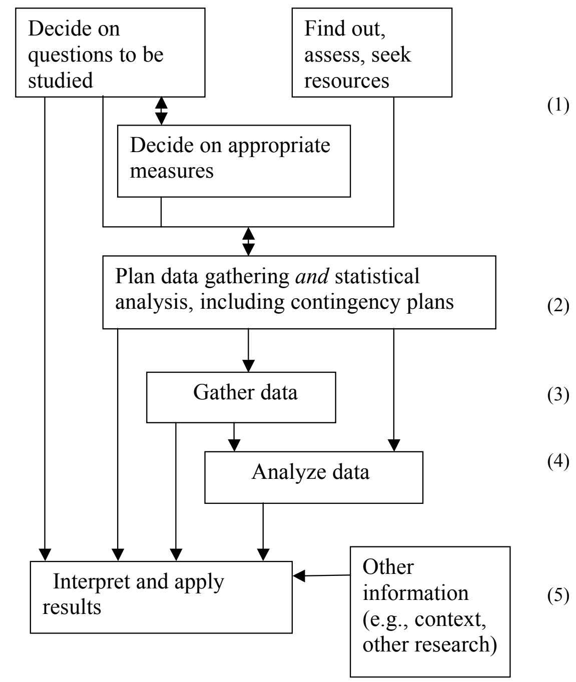
Course Description: We often hear results of studies that appear to contradict studies that were widely publicized just a couple of years ago. Medical researcher, John P. Ioannidis has claimed that, in fact, most claimed research findings are false. Most of his arguments involve the misunderstanding and misuse of statistics. Sometimes misunderstandings are passed down from teacher to student or from colleague to colleague. In some cases, policies based on these misunderstandings have become institutionalized. This workshop will discuss some of these misunderstandings and misuses, and offer suggestions for what teachers, readers, researchers, reviewers, and editors can do to deal with this fact of life and to help improve the situation.

Prerequisites: Familiarity with the basics of statistical inference (random variable, sampling distribution, hypothesis testing, confidence interval, simple linear regression, transformations of random variables, especially the logarithm). Some acquaintance with Analysis of Variance and Multiple Regression would be helpful but not necessary.

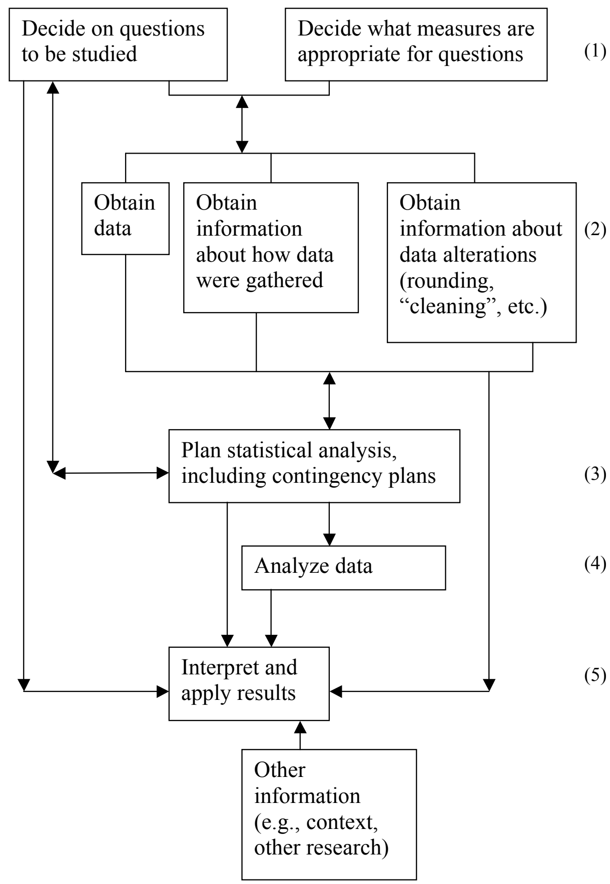
CONTENTS OF PART I

OUTLINE FLOW CHART OF PROCESS FOR RESEARCH INVOLVING STATISTICS

I. Research including data gathering



II. Research using existing data



COMMON MISTAKE:
EXPECTING TOO MUCH CERTAINTY

If it involves statistical inference, it involves uncertainty!

> *Humans may crave absolute certainty; they may aspire to it; they may pretend ... to have attained it. But the history of science … teaches that the most we can hope for is successive improvement in our understanding, learning from our mistakes, … but with the proviso that absolute certainty will always elude us.*
> Astronomer Carl Sagan, *The Demon-Haunted World: Science as a Candle in the Dark (1995), p. 28.*

> *... to deal with uncertainty successfully we must have a kind of tentative humility. We need a lack of hubris to allow us to see data and let them generate, in combination with what we already know, multiple alternative working hypotheses. These hypotheses are then modified as new data arrive. The sort of humility required was well described by the famous Princeton chemist Hubert N. Alyea, who once told his class, "I say not that it is, but that it seems to be; as it now seems to me to seem to be."*
> Statistician Howard Wainer, last page of *Picturing the Uncertain World (2009)*

(More quotes on website!)

General Recommendations Regarding Uncertainty

Recommendation for reading research that involves statistics:

- Look for sources of uncertainty.

Recommendations for planning research:

- Look for sources of uncertainty.
- Wherever possible, try to reduce or take into account uncertainty.

Recommendations for teaching and writing:

- Point out sources of uncertainty.
- Watch your language to be sure you don't falsely suggest certainty.
  Example: Do not say that a result obtained by statistical inference is true or has been proved; instead say, e.g., that the data support the result.

Recommendation for research supervisors, reviewers, editors, and members of IRB's:

- Look for sources of uncertainty.
- Has the researcher followed the recommendations above?

Other words that indicate uncertainty:

Random
Variability/variation
Noise
Probably/probability/probable/improbable
Possibly/possible/possibility
Plausibly/plausible

Caution: Assuming that words mean the same thing in different contexts is a **common mistake** in using statistics as well as in other fields.

Examples of different meanings of words relevant to uncertainty:

1. Some people (especially many in environmental studies) distinguish between "uncertainty" and "variability":

- *Variability* refers to natural variation in some quantity
- *Uncertainty* refers to the degree of precision with which a quantity is measured.

Example:
- The amount of a certain pollutant in the air is *variable*, since it varies from place to place and from time to time.
- The *uncertainty* in the amount of that pollutant present in a particular place at a particular time depends on the quality (and presence or absence) of the instruments used to measure it.

Other people consider both of these as instances of uncertainty.

2. The everyday and technical meanings of "random" are different. (More later.)

COMMON MISTAKES INVOLVING UNCERTAINTY AND CAUSALITY

1. **Confusing correlation and causation**.

Examples:

i. Elementary school students' shoe sizes and their scores on a standard reading exam are correlated. Does having a larger shoe size *cause* students to have higher reading scores?

ii. Suppose research has established that college GPA is related to SAT score by the equation

$$GPA = \alpha + \beta SAT,$$

and $\beta > 0$. Can we say that an increase of one point in SAT scores causes, on average, an increase of β points in college GPA?

2. **Interpreting causality deterministically when the evidence is statistical**.

After pointing out problems such as confusing correlation and causation, most statistics textbooks include a statement such as:

> "To establish causality, we need to use a randomized experiment."

Suppose a well planned, well implemented, carefully analyzed randomized experiment concludes that a certain medication is effective in lowering blood pressure. Would this be justification for telling someone, "This medication will lower your blood pressure?

WHAT IS PROBABILITY?

"*It is his knowledge and use of the theory of probability that distinguishes the statistician from the expert in chemistry, agriculture, bacteriology, medicine, production, consumer research, engineering, or anything else.*"

Statistician W. Edwards Deming

Uncertainty can often be "quantified"

- i.e., we can talk about *degrees* of certainty or uncertainty.
- This is the idea of probability: a higher probability expresses a higher degree of certainty/a lower degree of uncertainty that something will happen.
- Statistical inference techniques are based on probability.

Dictionary definition:

- American Heritage Dictionary Definition 3: "*Math*. A number expressing the likelihood of occurrence of a specific event, such as the ratio of the number of experimental results that would produce the event to the total number of results considered possible."
- AHD Definition 1 of Likelihood: "The state of being likely or probable; probability."

Compare:

- What is time?
- What is a point?

Terminology: The "something's" we consider the probabilities of are called ***events***.

Examples:

- The event that the number showing on a die we have rolled is 5.
- The event that it will rain tomorrow.
- The event that someone in a certain group will contract a certain disease within the next five years.

Probability of an Event: Three Perspectives

- Classical ("A priori" or "theoretical")
- Empirical ("A posteriori" or "Frequentist")
- Subjective

Classical ("A Priori" or "Theoretical") Perspective

- Situation: a non-deterministic process ("random process") with n *equally likely* outcomes.
- e.g., toss a fair die: Six equally likely outcomes,
- P(A) is defined to be m/n, where A is satisfied by exactly m of the n outcomes
- e.g., toss a fair die; A = an odd number comes up -> P(A) = 3/6.

Pros and Cons of Classical Probability

- Conceptually simple for many situations.
- Doesn't apply when outcomes are not equally likely.
- Doesn't apply when there are infinitely many outcomes

Empirical ("A Posteriori" or "Frequentist") Perspective

- Consider a process that we can imagine performing repeatedly (e.g., tossing a die); we consider an event A that can be described in terms of the results of the process (e.g., "the number that comes up is less than 4")
- P(A) is defined to be the limiting value, as we perform the process more and more times, of the ratio

$$\frac{\textit{Number of times A occurs}}{\textit{Number of times process is repeated}}$$

- E.g., toss a fair die; A = six lands up
- E.g., toss a die that is suspected of *not* being fair; A = six lands up.

Pros and Cons of Empirical Probability

- Covers more cases than classical.
- Intuitively, agrees with classical when classical applies.
- Repeating the identical experiment an infinite number of times (sometimes even twice) is physically impossible.
- How many times must we perform the process to get a good approximation to the limiting value?

The empirical view of probability is the one that is used in most commonly used statistical inference procedures. These are called *<u>frequentist</u>* *statistics.*

Subjective Perspective

- An individual's personal measure of belief that the event will occur.
- e.g., P(the stock market will go up tomorrow).
- Needs to be "coherent" to be workable.
    - e.g., P(stock market goes up tomorrow) = .6 and P(stock market goes down tomorrow) = .7 are inconsistent.

Pros and Cons of Subjective Probability

- Applicable in situations where other definitions are not.
- Fits intuitive sense of probability.
- Can be considered to extend classical and empirical views.
- Can vary from individual to individual.
- Requires "coherence" conditions; are people always that rational?

The subjective perspective of probability fits well with Bayesian statistics, which are an alternative to the more common frequentist statistical methods. (This course will mainly focus on frequentist statistics.)

Unifying Perspective: Axiomatic Model of Probability

- The coherence conditions needed for subjective probability can be proved to hold for the classical and empirical definitions.
- The axiomatic perspective codifies these coherence conditions, so can be used with any of the above three perspectives.

Terminology:

- The *certain event* is the event "some possibility occurs."
    - For example, in rolling a die, the certain event is "One of 1, 2, 3, 4, 5, 6 comes up."
    - In considering the stock market, the certain event is "The Dow Jones either goes up or goes down or stays the same."

- Two events are called *mutually exclusive* if they cannot both occur simultaneously.
    - For example, the events "the die comes up 1" and "the die comes up 4" are mutually exclusive (assuming we are talking about the same toss of the same die).

- The *union* of events is the event that at least one of the events occurs.
    - For example, if E is the event "a 1 comes up on the die" and F is the event "an even number comes up on the die," then the union of E and F is the event "the number that comes up on the die is either 1 or even."

The ***axiomatic model of probability*** says that probability is a function (i.e., a rule; we'll call it P) that assigns a number to each event, and satisfies the three conditions (axioms; coherence conditions) below. (Just what constitutes events will depend on the situation where probability is being used.)

The three axioms of probability:

I. 0 is the smallest allowable probability and 1 is the largest allowable probability (i.e., $0 \leq P(E) \leq 1$).

II. P(certain event) = 1

III. P(union of mutually exclusive events) = sum of P of individual events

Example: If we have a fair die, the axioms of probability *require* that each number come up with probability 1/6.

Proof: Since the die is fair, each number comes up with the same probability.

Since the outcomes "1 comes up," "2 comes up," ... "6 come up" are mutually exclusive and their union is the certain event, Axiom III says that

P(1 comes up) + P(2 comes up) + ... + P(6 comes up)

= P(the certain event),

which is 1 (by Axiom II).

Since all six probabilities on the left are equal, that common probability must be 1/6

MISUNDERSTANDINGS INVOLVING PROBABILITY

> *... misunderstanding of probability, may be the greatest of all general impediments to scientific literacy*.
>
> Stephen Jay Gould, *Dinosaur in a Haystack*

Common misunderstanding: If there are only two possible outcomes, and you don't know which is true, the probability of each of these outcomes is ½.

Possible contributing cause: Many students only see the Classical perspective, where outcomes have equal probabilities.

Teachers take warning!

Common misunderstanding: Confusing the "reference category"

Example (Gigerenzer et al, 2007): A physician may tell a patient that if he takes a certain antidepressant, his chance of developing a sexual problem is 30% to 50%.

- The patient may interpret that as saying that in 30% to 50% *of the occasions on which he wishes to have sex*, he will have a problem.
- But the physician means that 30 to 50% *of patients who take the medication* develop a sexual problem.

The intended "reference category" (or "population") is "patients who take the medication," but the patient heard "occasions on which he wishes to have sex."

Suggestions:

- In reading, be careful to interpret the reference category from context – and remain uncertain if you can't.
- In writing, be very careful to make the reference category clear. In particular, avoid saying, e.g., "his chance …" when the reference category is a population of people.

Common source of misunderstanding: Different uses of the word "risk."

Everyday meaning: risk = danger

Technical meanings: A number quantifying a danger.

1. Risk as a probability (*absolute risk*)

   Example: "the risk that a U.S. resident dies from a heart attack is about 25%."

   Note: Risk may be misunderstood if the reference category is not understood.

2. *Relative risk* (also called *risk ratio*).
   - A method of comparing the risk for one group with the risk for another.
   - E.g., one group might be people with a certain condition (or receiving a certain treatment) and the other group people without that condition (or not receiving the treatment). Or the groups might be men and women. Or smokers and non-smokers; etc.
   - *Relative risk* is *the ratio of the risks for the two group*s.
   - Three possible source of confusion:

     i. What are the two groups?

     ii. Which group's risk is in the numerator and which group's risk is in the denominator?

     iii. A relative risk is difficult to interpret without knowing the absolute risks.

Example:

You are told that a certain treatment will reduce your risk of contracting a certain disease by 25%.

This means that the relative risk of those having the treatment compared to those who don't have the treatment is 0.75.

Scenario 1: The absolute risk for those not using the treatment is 40% (i.e., 4 out of 10).

- A risk reduction of 25% reduces the risk to 3 out of 10, giving absolute risk 30%.
- So 10% of people who use the treatment benefit from it.
- This is substantial, but not as substantial as "25% risk reduction" might sound.

Scenario 2: The absolute risk for those not using the treatment is 0.0004 % (i.e., 4 out of 1,000,000).

- A risk reduction of 25% reduces the risk to 3 out of 1,000,000.
- This is not a substantial reduction.

3. Risk = Probability X Consequences

MISUNDERSTANDINGS INVOLVING CONDITIONAL PROBABILITIES

Conditional probability: A probability with some condition imposed.

Examples:

- The probability that a person with low bone density will have a hip fracture in the next five years. ("Low bone density" is the condition.)
- The probability that a person who scores below 400 on the SAT Math subject area exam will pass Calculus I. ("Scores below 400 on the SAT Math subject area exam" is the condition.)

Notation:

P(Event | Condition), read
"The probability of Event given Condition"

Conditional probabilities are very common. For example, we may talk about the probability of having a heart attack in the next five years for various conditions, such as:

Men
Women
Men over 65
Women with high cholesterol

Note: In these and many other examples, a conditional probability can be thought of as restricting interest to a certain *population*.

Common misunderstanding: Ignoring the condition

Example: A study of a cholesterol-lowering medication includes only men between the ages of 45 and 65 who have previously had a heart attack. The results give an estimate of the probability of the effectiveness of the medication for people in the group studied – that is, the conditional probability P(medicine effective| male between the ages of 45 and 65 who have previously had a heart attack)

How helpful would this study be in deciding whether or not to prescribe the medication to a woman who is 75 years old and has no previous record of heart attacks?

Note: Ignoring the condition is one form of *extrapolation*: applying or asserting a result beyond the conditions under which it has been studied. (More on this later.)

Common misunderstanding: Confusing a conditional probability and the reverse (also called inverse) probability.

In the notation P(E|F), the condition F is also an event, so it often makes sense to talk about P(F|E). However, these are different.

One situation where this confusion is particularly common is in reference to medical diagnostic tests. These are usually not perfect, so results are called "positive" and "negative" rather than "has disease" and "does not disease". It is then important to consider conditional probabilities such as

Sensitivity = P(tests positive | has the disease)
i.e., the probability that a person tests positive if the disease is present

and

Positive predictive value = P(has the disease | tests positive)
i.e., the probability that someone has the disease if they test positive

These are *not* usually the same. In fact, the sensitivity for a test can be very high (e.g., 95% or 99%), but the positive predictive value for that same test can be very low (e.g., 40% or less).

Which is of most interest to a patient who tests positive?

Note: To figure out the positive predictive value if you know the sensitivity, you also need to know the *specificity*
P(tests negative | does not have the disease)
and the *prevalence rate*
P(has the disease),
which itself might be a conditional probability – for example,
P(infected with HIV| intravenous drug user), or
P(infected with HIV | age over 80)

CONFUSIONS INVOLVING THE WORD "RANDOM"

The word "random" has various related but not identical *technical* meanings in statistics.

- The technical meaning may depend on the context.
- In some cases the exact technical meaning is hard to define precisely without getting so technical as to lose many people.
- In some cases, the everyday meaning is a pretty good guide, *whereas in other cases, it can cause misunderstandings.*
- The common element is that there is some degree of *uncertainty* (in particular, indeterminacy) involved.

The *everyday* (first) definitions of "random" from a couple of dictionaries:

> "*Having no specific pattern or objective; haphazard*" (*The American Heritage Dictionary*, Second College Edition, Houghton Mifflin, 1985)

> "*Proceeding, made, or occurring without definite aim, reason, or pattern* (*Dictionary.Com*, http://dictionary.reference.com/browse/random, accessed 11/19/09)

Specific technical uses of "random" include the phrases:

I. Random process
II. Random sample
III. Random variable

I. ***Random Process***

A random process may be thought of as a process where the outcome is probabilistic (also called *stochastic*) rather than deterministic in nature; that is, where there is uncertainty as to the result.

Examples:

1. Tossing a die – we don't know in advance what number will come up.
2. Flipping a coin – if you carefully devise an apparatus to flip the coin, it will always come up the same way. However, normal flipping by a human being can be considered a random process.
3. Shaking up a collection of balls in a hat and then pulling out one without looking.

Comment: All the examples above may appear to be situations where the outcomes have equal probabilities. But consider

1. A die that is not fair – e.g., 2 comes up twice as often as 3
2. A coin that is not fair – e.g., heads comes up twice as often as tails
3. A collection of balls not all of the same size or weight – you are more likely to pick out large balls than small ones, or light ones than heavy ones.

II. ***Random Sample***

A. Definition and Common Misunderstandings
B. Simple Random Sample, Part I
C. Other Types of Random Samples

A. A random sample can be defined as one that is chosen by a random process. Thus, *"random sample" really means "randomly chosen sample".*

Common confusion: The everyday definition of random prompts many people to believe that a random sample does not have a pattern.

- *This is false* – random samples may indeed display patterns.
- For example, it is possible for a random sequence of six coin tosses to have the pattern HTHTHT, or the pattern HHHTTT, etc.
- In fact, *there is no way we can tell from looking at the sample whether or not it qualifies as a random sample.*

Common myth: Many people believe that a random sample is representative of the population from which it is chosen.

- *This is false.*
- For example, a random sample of five people from a group might turn out to consist of the tallest five people in the group.

B. *Simple Random Sample, Part I*:

The following definition (from Moore and McCabe, *Introduction to the Practice of Statistics*) is good enough for many practical purposes:

> "*A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.*"

Here, *population* refers to the collection of people, animals, locations, etc. that the study is focusing on.

Examples:

1. In a medical study, the population might be all adults over age 50 who have high blood pressure.
2. In another study, the population might be all hospitals in the U.S. that perform heart bypass surgery.
3. If we were studying whether a certain die is fair or weighted, the population would be all possible tosses of the die.

In Example 3, it is fairly easy to get a simple random sample: Just toss the die n times, and record each outcome.

Selecting a simple random sample in examples 1 and 2 is much harder.

A good way to select a simple random sample for Example 2:

1) Obtain or make a list of all hospitals in the U.S. that perform heart bypass surgery. Number them 1, 2, ... up to the total number M of hospitals in the population. (Such a list is called a ***sampling frame***.)
2) Use a random number generating process (i.e., equivalent to the process of selecting balls used in some lotteries) to obtain a simple random sample of size n from the population of integers 1, 2, ... , M.
3) The simple random sample of hospitals would consist of those hospitals in the list corresponding to the numbers in the SRS of numbers.

In theory, the same process could be used in Example 1.

- However, obtaining the sampling frame would be much harder -- probably impossible.
- So some compromises may need to be made.
- Unfortunately, these compromises can easily lead to a sample that is *biased* (more later) or otherwise not close enough to random to be suitable for the statistical procedures used.

Cautions:

- Even the sampling procedure described above is a compromise and may not be suitable in some situations, described in the next section.
- We will discuss simple random samples more in a later section.

C. *Other types of Random Samples*

There are various types of random samples (also called ***probability samples***) besides simple random samples.

- These may be appropriate in some studies.
- But when they are used, *the correct method of statistical analysis will differ from the method for a simple random sample*.

Examples:

1. ***Stratified random sample***:

- The population is first classified into groups (called *strata*) with similar characteristics.
- Then a simple random sample is chosen from each stratum separately.
- These simple random samples are combined to form the overall sample.

Examples of characteristics on which strata might be based include: gender, state, school district, county, age.

Reasons to use a stratified rather than simple random sample include:

- The researchers may be interested in studying results by strata as well as overall. Stratified sampling can help ensure that there are enough observations within each stratum to be able to make meaningful inferences by strata.
- Statistical techniques can be chosen taking the strata into account to allow stronger conclusions to be drawn.
- Practical considerations may make it impossible to take a simple random sample.

2. ***One-stage cluster sample***:

- The population is also divided into groups, called *clusters*.
- Instead of sampling within each cluster, a simple random sample of *clusters* is selected.
- The overall sample consists of all individuals in the clusters that constitute this simple random sample of clusters.

Example: If the purpose of the study is to find the average hourly wage of convenience store employees in a city, the researcher might randomly select a sample of convenience stores in the city and find the hourly wages of all employees in each of the stores in the sample.

Note: The results from cluster samples are not as reliable as the results of simple random samples or stratified samples, so they should only be used if practical considerations do not allow a better sample scheme. For example, in the convenience store example, it may be practically speaking impossible to draw up a list of all convenience store employees in the city, but it would be much less difficult to draw up a list of all the convenience stores in the city.

III. **Random Variables**

In most applications, a ***random variable*** can be thought of as *a variable that depends on a random process*.

Examples:

1. Toss a die and look at what number is on the side that lands up.
    - Tossing the die is an example of a random process;
    - The number on top is the value of the random variable.

2. Toss two dice and take the *sum* of the numbers that land up.
    - Tossing the dice is the random process;
    - The sum is the value of the random variable.

3. Toss two dice and take the *product* of the numbers that land up.
    - Tossing the dice is the random process;
    - The product is the value random variable.

Examples 2 and 3 together illustrate:

> *The same random process can be involved in two different random variables*.

4. Randomly pick (in a way that gives each student an equal chance of being chosen) a UT student and measure their height.

- Picking the student is the random process.
- The student's height is the value of the random variable.

5. Randomly pick (in a way that gives each student an equal chance of being chosen) a student in a particular UT class and measure their height.

- Picking the student is the random process.
- The student's height is the value of the random variable.

Examples 4 and 5 illustrate:

> *Using the same variable* (in this case, height) *but different random processes* (in this case, choosing from different populations) *gives different random variables*.

Confusing two random variables with the same variable but different random processes is a **common mistake**.

6. Measure the height of the third student who walks into the class in Example 5.

- In all the examples before this one, the random process was done deliberately.
- In Example 6, the random process is one that occurs naturally.
- *Because Examples 5 and 6 depend on different random processes, they are different random variables*.

7. Toss a coin and see whether it comes up heads or tails.

- Tossing the coin is the random process.
- The variable is heads or tails.
- Example 7 shows that *a random variable doesn't necessarily have to take on numerical values*.

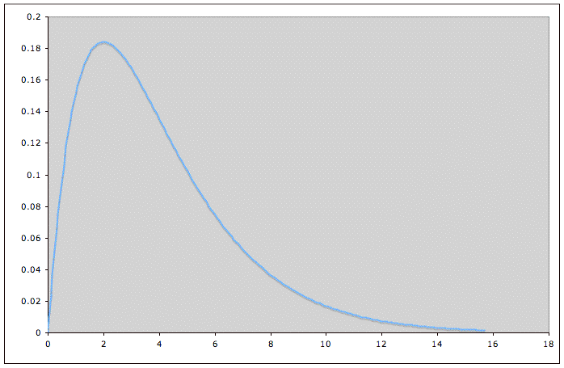
Probability Distributions:

*Recall***:**

- "Random" indicates uncertainty.
- Probability quantifies uncertainty

For random variables, probability enters as a *probability distribution*:

- Typically, some values (or ranges of values) of a random variable occur more frequently than others.
- For example, if we are talking about heights of university students, heights of around 5' 7" are much more common that heights of around 4' or heights around 7'.
- In other words, some values of the random variable occur with higher probability than others.
- This can be represented graphically by the ***probability distribution*** of the random variable.
- For example, a random variable might have a probability distribution that looks like this:



- The possible values for the random variable are along the horizontal axis.
- The height of the curve above a possible value roughly tells how likely the nearby values are.
- This particular distribution tells us that values of the random variable around 2 (where the curve is highest) are most common, and values greater than 2 become increasingly less common, with values greater than 14 (where the curve is lowest) very uncommon.
- More precisely, *the area under the curve between two values a and b is the probability that the random variable will take on values between a and b.*
- In this example, we can see that the value of the random variable is much more likely to lie between 2 and 4 (where the curve is high, hence has a lot of area under it) than between 12 and 14 (where the curve is low, and hence has little area under it).

BIASED SAMPLING AND EXTRAPOLATION

A sampling method is called ***biased*** if it *systematically* favors some outcomes over others.

Bias can be intentional, but *often is unintentional*. The following example shows how *a sample can be biased, <u>even though there is some randomness in the selection</u> of the sample.*

Example: Telephone sampling is common in marketing surveys.

- A simple random sample might be chosen from the sampling frame consisting of a list of *telephone numbers* of people in the area being surveyed.
- This method does involve taking a simple random sample, but it is *not* a simple random sample of *the target population* (consumers in the area being surveyed.)
- It will miss people who do not have a phone.
    - It may also miss people who only have a cell phone that has an area code not in the region being surveyed.
    - It will also miss people who do not wish to be surveyed, including those who monitor calls on an answering machine and don't answer those from telephone surveyors.
- *Thus the method systematically excludes certain types of consumers in the area.*

Inferences from a biased sample are not as trustworthy as conclusions from a truly random sample.

Common Sources and Consequences of Bias:

1. **Convenience samples**:
- Sometimes it is not possible or not practical to choose a random sample.
- In those cases, a ***convenience sample*** might be used.
- Sometimes it's plausible that a convenience sample could be considered as a random sample, but often a convenience sample is biased.
- *If a convenience sample is used, inferences are not as trustworthy as if a random sample is used.*

2. Voluntary response samples:
- If the researcher appeals to people to voluntarily participate in a survey, the resulting sample is called a "voluntary response sample."
- *Voluntary response samples are always biased*:
  - They only include people who choose volunteer,
  - whereas a random sample would need to include people whether or not they choose to volunteer.
- Often, voluntary response samples over-sample people who have strong opinions and under-sample people who don't care much about the topic of the survey.
- Thus *inferences from a voluntary response sample are not as trustworthy as conclusions based on a random sample of the entire population under consideration*.

Lack of Blinding: When two "treatments" are compared (e.g., drugs; surgical procedures; teaching methods), bias can sometimes be introduced by the human beings involved, despite their best efforts to be objective and unbiased.
- Thus it is important in these situations to try to make sure that no one who might, even unintentionally, influence the results knows which treatment each subject is receiving.
- This is called *blinding*.

Examples:

1. If two *drugs* are being compared (or a drug and a placebo), blinding involves the following (and possibly more):

- The two pills need to look alike, so the patient and the attending medical personnel don't know which drug the patient is taking.
- If a drug has noticeable side effects and is being compared with a placebo, the placebo should have the same side effects.
- The person arranging the randomization (i.e., which patient takes which drug) should have no other involvement in the study, and should not reveal to anyone involved in the study which patient is taking which drug.
- Anyone evaluating patient outcomes (e.g., examining the patient or asking the patient about their symptoms) should not know which drug the patient is taking.

2. Now suppose that two *surgical treatments* are being compared.

- *It is impossible to prevent the surgeons from knowing which surgical treatment they are giving*.
- Thus, total blinding is not possible, and there is the possibility that the surgeon's knowledge of which treatment is being given might influence the outcome.
- Sometimes the researchers can partially get around this by using only surgeons who genuinely believe that the technique they are using is the better of the two.
    - But this may introduce a confounding of technique and surgeon characteristics:
    - For example, the surgeons preferring one technique might be, as a group, more skilled or more experienced or more careful than the surgeons preferring the other, or have different training that affects the outcome regardless of the surgical method.

***Extrapolation*:** In statistics, drawing a conclusion about something beyond the range of the data is called *extrapolation*.

- Drawing a conclusion from a biased sample is one form of extrapolation:
    - Since the sampling method systematically excludes certain parts of the population under consideration, the inferences only apply to the subpopulation that has actually been sampled.
- Extrapolation also occurs if, for example, an inference based on a sample of university undergraduates is applied to older adults or to adults with only an eighth grade education.
- *Extrapolation is a* **common error** *in applying or interpreting statistics*.
- Sometimes, because of the difficulty or impossibility of obtaining good data, extrapolation is the best we can do, but it always needs to be taken with at least a grain of salt – i.e., with a large dose of uncertainty.

PROBLEMS INVOLVING CHOICE OF MEASURE

I. ***Choosing Outcome (and Predictor) Measures (Variables)***
II. ***Using Questionnaires***
III. ***Choosing summary statistics***

I. *Choosing Outcome (and Predictor) Measures (Variables)*

Example 1: A study is designed to measure the effect of a medication intended to reduce the incidence of osteoporotic fractures. Subjects are randomly divided into two groups. One group takes the new medication, the other, a placebo or an existing medication. What should be measured to compare the two groups?

Bone density?
Percent of subjects who have hip fractures?
Average number of hip fractures?
Percent of subjects who have vertebral fractures?
Average number of vertebral fractures?
Percent of subjects who experience any fracture?
Average number of fractures of all kinds?
More than one of these?

Note:

- All of these involve *random variables* – e.g, bone density, number of hip fractures, mobility
- Measures such as bone density and body mass index are sometimes called *markers* or *proxy measures*.

Example 2: The official US Unemployment Rate is defined as "Total unemployed persons, as a percent of the civilian labor force."

- This measure of unemployment depends on the definitions of "unemployed" and "civilian labor force".
- For example, "employed persons" includes "All persons who did at least 15 hours of unpaid work in a family-owned enterprise operated by someone in their household."
- Other countries use different definitions of unemployment.
- In 1976, the U.S. department introduced several "Alternative measures of labor underutilization" and regularly publishes these other measures of unemployment rate.

(For more examples and references for the various "measures of labor underutilization," see http://www.ma.utexas.edu/users/mks/statmistakes/Outcomevariables.html)

Comments: Choice of measure is not always easy; it may involve compromises. For example:

- A good measure may be harder to obtain than a proxy measure; the researchers need to weigh the expense and benefits of each choice.
- Changing a measure that has been used in past research because a better measure is feasible may prevent comparisons of trends in the past and future.

Suggestions for researchers:
- Think carefully about the measures you use.
- Be sure to give a precise definition of your measures.
- Explain *why* you chose the measures you did.
- State clearly any cautions needed in using your measures.

Suggestion for reviewers, supervisors, editors, etc:
- Have the researchers done all the above? Or have they left the reader with more uncertainty than they should?

Suggestions for consumers of research:
- Carefully read the definitions of measures.
  - They may not be what you might think (e.g., unemployment rate).
- Think whether the measures used really measure what you are interested in (e.g., bone density vs. incidence of fractures)
- Be cautious in drawing conclusions involving more than one study – if the measures are not the same, comparisons in results may not be valid.

II. *Using Questionnaires*

Asking people questions can raise many problems. Two main types of problems:
- Questions may be ambiguous.
  - The responder may interpret them differently than the questioner.
  - Different responders may have different interpretations.

- The wording or sequencing of questions can influence the answers given.

Examples:

1. The developers of a questionnaire to study the incidence of disabilities tried to write a question to detect psychosis. They tried, “Do you see things other people don't see or hear things other people don't hear?” But in testing the question, they found that non-psychotic people would answer yes to the question, explaining that they had unusually good vision, or excellent hearing.

2. In developing a survey on housing demand, researchers found that if they asked questions about specific amenities before asking a question on overall satisfaction, the overall satisfaction rating was lower than if these more specific questions were asked later than the overall satisfaction question.

Devising good questions is much more complicated and subtle than it may initially appear; whole books have been written about wording questions; entire courses are devoted to survey design; there are labs devoted to testing out questions for surveys. See http://www.ma.utexas.edu/users/mks/statmistakes/wordingquestions.html and the links and references therein for more examples and resources.

Suggestions for researchers using questionnaires:
- Educate yourself about the problems involved in designing good survey questions.
- Test out your tentative questions on subjects *similar to those you plan to survey*.
- Then modify questions and re-test as needed.
- When reporting results of a survey, be sure to provide access to the exact questions asked, so others can verify whether or not the questions are likely to be ambiguous or influential.
- In reporting results, discuss any questions that turned out to be problematical.
    - Take the problems into account in interpreting analysis of the data.

Suggestions for reviewers, editors, etc.:
- Have the researchers done the above?

Suggestions for consumers of research:
- Be cautious in interpreting the results of a survey.
- In particular, try to find the exact questions asked and check them over for possible ambiguity or other problems in wording.
- If the authors of the survey are not willing to reveal the questions, be doubly cautious in making interpretations.

Suggestions for teachers:
- Even though you cannot include thorough coverage of the topic of wording in a general statistics course, be sure to
    - Mention it.
    - Give some examples
    - Ask some questions on it on exams, and
    - Have students pay attention to question wording if they are designing or carrying out a study.

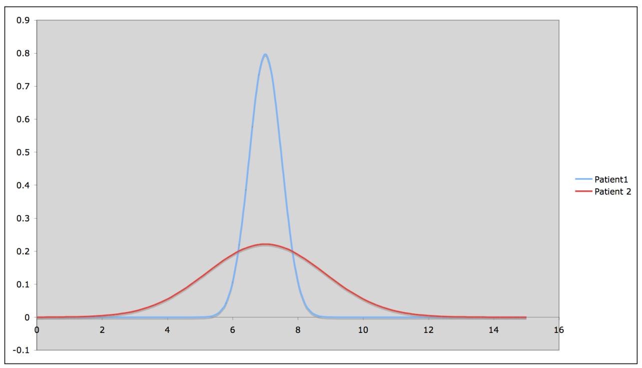
III. ***Choosing Summary Statistics***

- Many of the most common statistical techniques (e.g., one and two sample t-tests, linear regression, analysis of variance) concern the mean.
- In many circumstances, focusing on the mean is appropriate.
- But there are also many circumstances where focusing on the mean can lead us astray. Some types of situations where this is the case:
    - A. *When Variability Is Important*
    - B. *Skewed Distributions*
    - C. *Unusual Events*

A. *When Variability Is Important*

Example: The target range for blood glucose (BG, in millimoles per liter) is 3.9 to 10. The graph below shows the distribution of blood glucose for two hypothetical patients.

- Both patients have *mean* BG 7.
- The distribution for Patient 1 (blue) has standard deviation 0.5.
- The distribution for Patient 2 (red) has standard deviation 1.8.
- Patient 1 is in good shape.
- Patient 2 is in trouble.



Standard deviation is one common measure of variability.

- Depending on the situation, other measures of variability may be more appropriate, as discussed below.

Focusing just on the mean and ignoring variability is a **common mistake**, particularly in applying results.

B. *Skewed Distributions*

A ***skewed distribution*** is one that is bunched up on one side and has a "tail" on the other.

Measure of Center

When we focus on the mean of a variable, we're usually trying to focus on what happens "on average," or perhaps "typically".

- The mean does this well when the distribution is symmetrical, and especially when it is "mound-shaped," such as a normal distribution.
  - For a symmetrical distribution, the mean is in the middle.
  - If the distribution is also mound-shaped, then values near the mean are typical.
- But *if a distribution is skewed, then the mean is usually not in the middle.*
  - *Example*: The mean of the ten numbers 1, 1, 1, 2, 2, 3, 5, 8, 12, 17 is 55/10 = 5.2.
  - Seven of the ten numbers are less than the mean, with only three of the ten numbers greater than the mean.
  - A better measure of the center for this distribution would be the *median*, which in this case is (2+3)/2 = 2.5.
  - Five of the numbers are less than the median 2.5, and five are greater.

- Notice that in this example, the mean (5.2) is *greater* than the median (2.5).
- This is common for a distribution that is *skewed to the right* (that is, bunched up toward the left and with a "tail" stretching toward the right).

- Similarly, a distribution that is *skewed to the left* (bunched up toward the right with a "tail" stretching toward the left) typically has a mean *smaller* than its median.
  - See von Hippel 2005 for discussion of exceptions.
- *Note*: For a symmetrical distribution, such as a normal distribution, the mean and median are the same.

Many distributions that occur in practical situations are skewed, not symmetric.

Example: Suppose a friend is considering moving to Austin and asks you what houses here typically cost.

- Would you tell her the mean or the median house price?
- [Hint: Think Dellionaires]

In fact, blood glucose typically has a skewed (to the right) distribution rather than the normal distribution shown in the example above.

See Limpert and Stahel (1998) for more examples.

Measure of Spread

For a *normal* distribution, the standard deviation is a very appropriate measure of variability (or spread) of the distribution.

- If you know a distribution is normal, then knowing its mean and standard deviation tells you exactly which normal distribution you have.

But for *skewed* distributions, the standard deviation *gives no information on the asymmetry.*

- For a skewed distribution, it's usually better to use the *first and third quartiles*, since these will give some sense of the asymmetry of the distribution.

Other Problems Arising with Skewed Distributions

- Most standard statistical techniques focus on the mean *and* also assume that the random variable in question has a distribution that is normal. *(More later)*
- Many still give pretty accurate results if the random variable has a distribution that is not too far from normal.
- But *many common statistical techniques are **not valid** for strongly skewed distributions.*
- Applying techniques intended for normal distributions to strongly skewed distributions is a **common mistake**.

Possible remedies:

i. Always plot the data before applying a statistical test assuming a normal distribution

- If the data are strongly skewed, use one of the techniques below.
- See more details later in these notes

ii. Consider taking logarithms or applying another transformation to the original data

- Many skewed random variables that arise in applications are *lognormal*.
- This means that the logarithm of the random variable is normal.
- Hence most common statistical techniques *can* be applied to the logarithm of a lognormal (or approximately lognormal) variable.
- However, doing this *may require some care in interpretation*. There are three common routes to interpretation when dealing with logs of variables. (For more details, see http://www.ma.utexas.edu/users/mks/statmistakes/skeweddistributions.html)
- For some skew distributions that are not lognormal, another transformations (e.g., square root) can yield a distribution that is close enough to normal to apply standard techniques. However, interpretation will depend on the transformation used.

iii. Try non-parametric techniques

iv. If regression is appropriate, try quantile regression

- Standard regression estimates the *mean* of the conditional distribution (conditioned on the values of the predictors) of the response variable.
- *Quantile regression* is a method for estimating conditional quantiles (i.e., percentiles), including the median.
- For more on quantile regression, see Roger Koenker's Quantile Regression website.

C. *Unusual Events*

If the research question being studied involves unusual events, the mean or median is *not* adequate as a summary statistic.

Examples:

1. If you are deciding what capacity air conditioner you need, the average yearly (or even average summer) temperature will not give you guidance in choosing an air conditioner that will keep your house cool on the hottest days.
   - Instead, it would be much more helpful to know the highest temperature you might encounter, or how many days you can expect to be above a certain temperature.

2. Traffic safety interventions typically are aimed at high-speed situations.
   - So the average speed is not as useful as, say, the 85th percentile of speed.

3. Pregnancy interventions are often aimed at reducing the incidence of low birth weight babies.
   - The mean or median birth weights in the intervention and non-intervention group do not give you this information.
   - Instead, we need to focus on percentage of births in the low weight category.
   - This might be defined in absolute terms (e.g., weight below a certain specific weight) or in relative terms (e.g., below the median or below the first quartile.)

4. If two medications for lowering blood pressure have been compared in a well-designed, carefully carried out randomized clinical trial, and the average drop in blood pressure for Drug A is more than that for Drug B, we cannot conclude just from this information alone that Drug A is better than Drug B. We also need to consider the incidence of undesirable side effects.
   - One might be that for some patients, Drug A lowers blood pressure to dangerously low levels.
   - Or it might be the case that for some patients, Drug A actually increases blood pressure.
   - Thus in this situation, we need to consider extreme events in *both* directions.
   - Note that this is another example where variability is important.

5. Unusual events such as earthquakes and extreme behavior of the stock market can have large effects, so are important to consider.

Many techniques have been developed for studying unusual events.
   - However, these techniques are not usually mentioned in introductory courses in statistics.
   - And, like other statistical techniques, they are not "one-size-fits-all."
   - See http://www.ma.utexas.edu/users/mks/statmistakes/unusualevents.html for some references.

REFERENCES

Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, Steven Woloshin (2007), "Helping Doctors and Patients Make Sense of Health Statistics," *Psychological Science in the Public Interest*, vol. 8, No. 2, pp. 53 - 96. Download from http://www.psychologicalscience.org/journals/index.cfm?journal=pspi&content=pspi/8_2

Von Hippel, Paul (2005), "Mean, Median, and Skew: Correcting a Textbook Rule," *Journal of Statistics Education, v. 13 no. 2,* **www.amstat.org/publications/jse/v13n2/vonhippel.html**

Koenker, Roger, Quantile Regression website, **http://www.econ.uiuc.edu/~roger/research/rq/rq.html** Includes references to books, articles, and software**.**

Limpert, E. and Stahel, W. (1998) Life is Lognormal!, http://www.inf.ethz.ch/personal/gut/lognormal/brochure.html