NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS –
SPOTTING THEM AND AVOIDING THEM

May 26 - 29, 2015

Instructor: Martha K. Smith

Day 1: Fundamental Mistakes and Misunderstandings

Course Description: We often hear results of research studies that contradict earlier studies. In 2005, medical researcher John P. Ioannidis asserted that most claimed research findings are false. In 2011, psychologists Simmons, Nelson and Simonsohn brought further attention to this topic by using methods common in their field to "show" that people were almost 1.5 years younger after listening to one piece of music than after listening to another.

Both articles highlight the frequency and consequences of misunderstandings and misuses of statistical inference techniques. These misunderstandings and misuses are often passed down from teacher to student or from colleague to colleague. Some practices based on these misunderstandings have become institutionalized. This course will discuss some of these misunderstandings and misuses.

Topics covered include:

- Mistakes involving uncertainty, probability, or randomness
- Biased sampling
- Problematical choice of measures
- Misinterpretations and misuses of p-values
- Mistakes involving statistical power
- The File Drawer Problem (AKA Publication Bias)
- Multiple Inference (AKA Multiple Testing, Multiple Comparisons, Multiplicities, or The Curse of Multiplicity)
- Data Snooping.
- Ignoring model assumptions.

To aid understanding of these mistakes, about half the course time will be spent deepening understanding of the basics of statistical inference beyond what is typically covered in an introductory statistics course.

Course notes and supplemental materials are available at the Canvas course website, and at the website at http://www.ma.utexas.edu/users/mks/CommonMistakes2015/commonmistakeshome2015.html.

The supplemental materials provide:

- Elaboration of some items discussed only briefly in class
- References cited in the class notes
- Specific suggestions for what teachers, readers of research, researchers, referees, reviewers, and editors can do to avoid or deal with these mistakes.
- Additional references

Additional information on this general topic is available at the instructor's website *Common ~~Misteaks~~ Mistakes in Using Statistics* at http://www.ma.utexas.edu/users/mks/statmistakes/TOC.html (or just google: misteaks statistics)

CONTENTS OF DAY I:
Fundamental Mistakes and Misunderstandings

I. MISTAKES INVOLVING UNCERTAINTY

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)*

> *One of our ongoing themes when discussing scientific ethics is the central role of statistics in recognizing and communicating uncertainty. Unfortunately, statistics—and the scientific process more generally—often seems to be used more as a way of laundering uncertainty, processing data until researchers and consumers of research can feel safe acting as if various scientific hypotheses are unquestionably true.*
>
> Statisicians Andrew Gelman and Eric Loden, "The AAA Tranche of Subprime Science," the *Ethics and Statistics* column in *Chance Magazine 27.1*, 2014, 51-56, available at **http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics10.pdf**

(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.

  Better alternatives:

  _____

  _____

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

- Look for sources of uncertainty.
- If the researcher has not followed the recommendations above, send the paper or proposal back for appropriate revisions.

Terminology Inspired Confusions Involving Uncertainty:

Many words are used to indicate uncertainty, including:

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

Moreover, these and other words indicating uncertainty may be used with different meanings in different contexts.

Examples:

1. Some people (e.g., in environmental studies) distinguish between "uncertainty" and "variability":

- *Variability* refers to natural variation in some quantity
  - May be called *aleatory* (from Lat. *aleator*, gambler) *uncertainty* in some fields
- *Uncertainty* refers to the degree of precision with which a quantity is measured.
  - May be called *epistemic uncertainty* or *fuzziness* in some fields

Environmental example:

- The amount of a certain pollutant in the air is *variable*, because _____
- The amount of a certain pollutant in the air is *uncertain*, because
  _____

  _____

Other people may consider both of these as instances of uncertainty, or as instances of variability.

2. *Noise* is sometimes used to mean something similar to the use of "uncertainty" in Example 1, but is sometimes used to refer to variability. The latter use can often be confusing.

Example: In neural imaging, MRI scans produce waveforms that are used to obtain information about what is happening in a person's brain.

- Extraneous factors (such as the person's slight body movements, or vibration of the machine) produce *noise* in the waveform.
- But there is also *variability* from person to person that is reflective of different brain activity.

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

For more examples of terminology-inspired confusions in statistics, see Wainer (2011)

Common Mistakes Involving Causality and Uncertainty

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?

Note: The confusion in Example (ii) is partly fostered by confusing terminology: The coefficient β of SAT is called an "effect" – but ***in this statistical use, effect does not imply causality.***

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?

II. MISTAKES INVOLVING 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 of probability:

- 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?

Differing Perspectives on Probability

Some confusions involving probability and statistics involve confusing three perspectives on probability:

1. Classical ("A priori" or "theoretical")
2. Empirical ("A posteriori" or "frequentist" or, confusingly, "classical")
3. Subjective

Terminology for all perspectives: The things 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.

Perspective 1: Classical ("A Priori" or "Theoretical")

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

Pros and Cons of Classical Probability

Pro or Con?

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

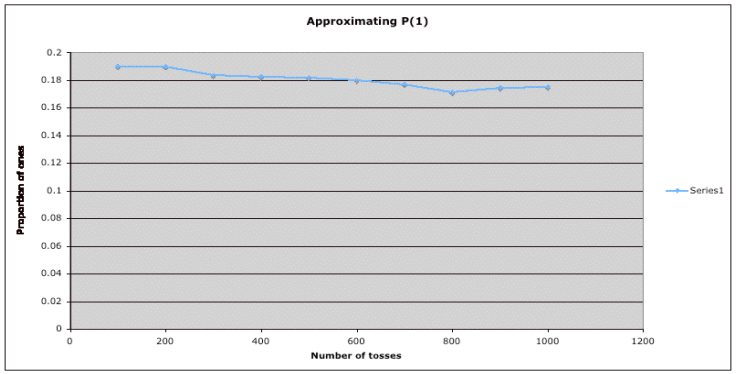
Perspective 2: Empirical ("A Posteriori" or "Frequentist" or, confusingly, "Classical")

- Consider a process that we can imagine performing repeatedly.
  - e.g., tossing a (not necessarily fair) die
- Now 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

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

- Compare and contrast:
  1. Toss a fair die; A = 4 lands up
  2. Toss a die that is suspected of *not* being fair; A = 4 lands up.

Illustration of the Empirical Perspective of Probability: The graph shows the results of a simulation of tossing a fair die 1000 times, recording after each 100 tosses the proportion of times "1" comes up on the (simulated) die.



- The horizontal axis shows the number of tosses of the fair die.
- The vertical axis shows the proportion of those tosses that came up 1.
- The *trend* of the graph is that as the number of tosses increases, the proportion of ones approaches the true probability of 1/6 = 0.16666... .
- Notice that the zeroing in on the true value is not steady -- in this particular simulation, there is some moving upward from 800 to 1000.
- If we increased the number of tosses to 2000, 3000, etc., we would expect the calculated proportions to vary less and less from the true value.

Pros and Cons of Empirical Probability

Pro or Con?

- 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 frequentist statistics.*

Perspective 3: Subjective

- An individual's subjective probability of an event is the 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

Pro or Con?

- 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.***

A 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.
- It is used to prove the mathematical theorems on which statistical inference relies.
- See Appendix for more details

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 heed!

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., "your chances …" when the reference category is a population of people.

Misunderstandings Involving Conditional Probabilities

Conditional probability: A probability with some condition imposed.

Note: The condition describes an event.

Examples:

1. The probability that a randomly chosen person with low bone density will have a hip fracture in the next five years.

   The condition is _____

2. The probability that a person who scores below 400 on the SAT Math subject area exam will pass Calculus I.

   The condition is _____

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

In Example 1, the subpopulation is _____

In Example 2, the subpopulation is _____

Notation:

P(Event | Condition), read as

"The probability of Event given Condition"

e.g.,

In Example 1: _____

In Example 2: _____

Conditional probabilities are very common.

Example: We could talk consider the probability of having a heart attack in the next five years for various conditions (subpopulations), such as:

Male
Female
Male over 65
Female with high cholesterol

Common mistake: 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 has 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.

Recall that 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 concepts.*

Example: 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 have 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 (PPV) = P(has the disease | tests positive)

> i.e., the probability that someone has the disease if they test positive

Note that *these are reverse conditional probabilities*.

These conditional probabilities are usually *not* 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 of these two conditional probabilities is of most interest to a patient who tests positive?

(More detail in appendix)

III. CONFUSIONS INVOLVING THE WORD "RANDOM"

Dictionary vs Technical Meanings

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)
>
> ***These uses are often misleading in a technical context.***

Specific technical uses of "random" include the phrases:

A. Random process
B. Random sample
C. Random variable

A. Random Processes

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 enough 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.

Caution: 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 of different sizes or weights – you are more likely to pick out large balls than small ones, or light ones than heavy ones.

B. Random Samples

Definition and Common Misunderstandings:

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

i.e., it's the *process*, not the *result* that makes the sample random.

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 (i.e., chosen by a random process) 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* - a random sample might, by chance, turn out to be anything but representative.
- For example, a random sample of five people from a group might turn out to consist of the tallest five people in the group.
- If you find a book or web page that claims that random samples are representative, apply some healthy skepticism to other things it claims!

We'll discuss later why random samples are important.

Preliminary Definition of Simple Random Sample

The following definition (from Moore and McCabe, *Introduction to the Practice of Statistics, 5th edition*) 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's 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.

Difficulties in Obtaining a Simple Random Sample

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 used in some lotteries of drawing balls from a container) 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*.

Caution: Even the sampling procedure described above is a compromise and may not be suitable in some situations.

Other Types of Random Samples (briefly)

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*. **Using a method requiring a simple random sample with a different type of random sample is a common mistake in using statistics.**
- Examples include *stratified random samples* and *cluster samples*.
    - *See the Appendix for further details*

C. 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 the 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.
   - The random process is _____
   - The value of the random variable is _____

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

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.
   - The random process is _____
   - The value of the random variable is _____

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.
   - The random process is _____
   - The value of the random variable is _____

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.*

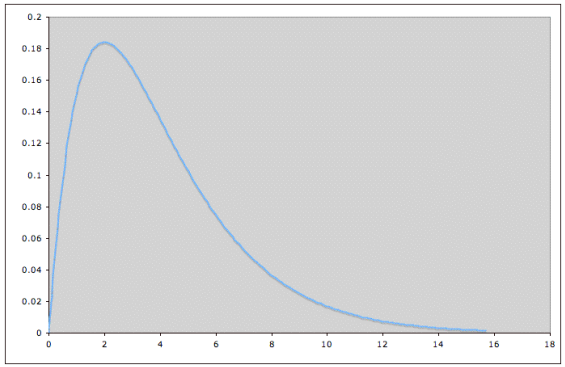
7. Toss a coin and see whether it comes up heads or tails.
- The random process is _____
- The value of the random variable is _____
- Example 7 shows that *a random variable doesn't necessarily have to take on numerical values.*

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're 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.

Example:



- 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*.

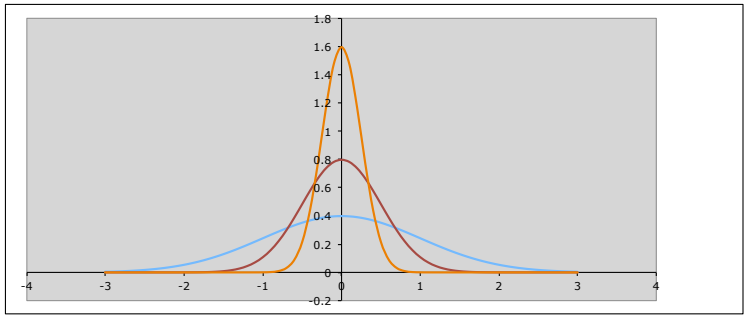
- In this example, we can see that the value of the random variable is much more likely to lie between 2 and 4 (where the curve is high, hence has a lot of area under it) than between 12 and 14 (where the curve is low, and hence has little area under it).

Question: What is the total area under a probability distribution curve?

Why?

Common mistake: Confusing different normal distributions.

Example 1: Three different normal distributions, same scale



What are the means of the three distributions?

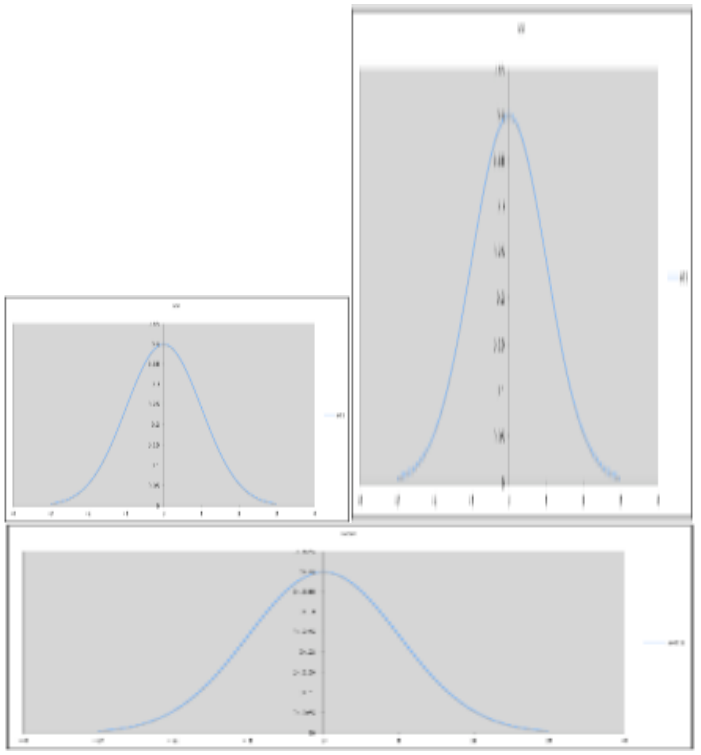
How do you know?

One of the distributions has standard deviation 1, one has standard deviation 0.5, and one has standard deviation 0.25.

Which distribution has which standard deviation?

How do you know?

Example 2: The same normal distribution (mean 0, standard deviation 1), but shown with three different "aspect ratios."



IV. BIASED SAMPLING AND EXTRAPOLATION

A sampling method is called ***biased*** (or often, ***systematically biased***) if it *systematically* favors some outcomes over others.

- Systematic bias can be intentional, but *often happens unintentionally*.
- Systematic bias is sometimes called ***ascertainment bias***, especially in medical or biological studies.

Inferences from a systematically biased sample are not as trustworthy as conclusions from a truly random sample, so need to be taken with a large grain of salt.

Some Randomness Does Not Ensure Lack of Systematic Bias

The following examples show how *a sample can be systematically biased, even though there is some randomness in the selection 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 (of telephone numbers), but it is *not* a simple random sample of *the target population* (households or consumers in the area being surveyed.)
- It will miss:
  - People who do not have a phone.
  - People who only have a cell phone that has an area code not in the region being surveyed.
  - People who do not wish to be surveyed, including those who monitor calls on an answering machine and don't answer calls from telephone surveyors.
- *Thus the method systematically excludes certain types of consumers in the area.*

Common Sources of Systematic Bias:

1. **Convenience samples**:
   - Sometimes it's 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 systematically 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 systematically biased*:
     - They only include people who choose to volunteer, whereas
     - A random sample would need to include people whether or not they choose to volunteer.
   - In addition, voluntary response samples typically 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.*

3. Lack of Blinding: When two "treatments" are compared (e.g., drugs; surgical procedures; teaching methods), systematic bias can sometimes be introduced by the human beings involved, despite their best efforts to be objective and unbiased.
   - Thus it's important in such 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:

A. 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.

B. 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.

Miscellaneous sources of sampling bias: Sampling bias may occur for many reasons, so vigilance is needed!

Example 1: Studies of human genetic variation typically use DNA microchips to identify variation in certain genes that are known to have different versions. But if the microchip is created to assess only certain genes known to vary in a particular population, the study will not pick up genes that do not vary in that population but vary between that population and others, or within some other populations.

For example, a study using a microchip based on genes known to vary in European populations may miss variation between European and Asian populations and between different Asian populations. (Jobling and Tyler-Smith, 2003, Box 1, p. 600)

Example 2: Some clinical trials have a "pre-randomized run-in period" to identify patients who are not adherent, or who respond to the placebo being used, or who do not tolerate or do not respond to the drug being tested. These patients are then excluded from the group that is randomized to treatment. (Pablos-Mendez et al, 1998)

How does this cause sampling bias?

How is this sampling bias likely to affect the results of the study?

Extrapolation

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

- Drawing a conclusion from a systematically 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 mistake** *in applying or interpreting statistics*.
- Because of the difficulty or impossibility of obtaining good data, extrapolation is sometimes 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.

V. PROBLEMS INVOLVING CHOICE OF MEASURES

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?
Number of subjects having hip fractures?
Number of hip fractures experienced by subjects?
Number of subjects who have vertebral fractures?
Number of vertebral fractures experienced by subjects?
Number of subjects who experience any fracture?
Number of fractures of all kinds experienced by subjects?
More than one of these?
Something else?

Note:

- All of these involve *random variables* – e.g, bone density, number of hip fractures
- Measures such as bone density and body mass index are sometimes called *markers* or *proxy measures* or *surrogates*.
- For an example of how poor proxy measures can go awry, see Distrust Your Data, https://source.opennews.org/en-US/learning/distrust-your-data/

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, the official definition of "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 of Labor introduced several "Alternative measures of labor underutilization" and regularly publishes these other measures of unemployment rate.
- Recently more and more people are holding more than one part time job, sometimes serially taking temporary jobs. How are they/should they be counted?

(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 often difficult; 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.
- Statistical properties of the measure are also relevant; see Senn and Julious (2009) for discussion of this for measures used in clinical trials, and for further references.

Suggestions for consumers of research:

- Carefully read the definitions of measures.
    - They may not be what you might think (e.g., unemployment rate).
- Think about whether the measures used really measure what you're 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 of results may not be valid*.

Suggestions for researchers:

- Think carefully about the measures you use. Ask others to critique your choices and reasoning.
- Be sure to give a precise definition of each measure you use.
- 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 of the above? Or have they left the reader with more uncertainty than they should?

Asking Questions (*if time permits; more in Appendix*)

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?"
    - 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.

(More details and suggestions in Appendix)

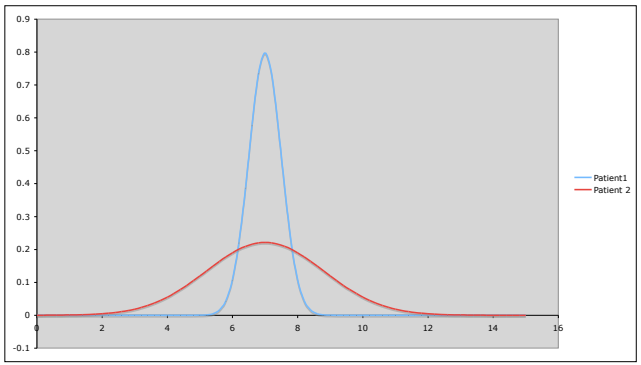
Choosing Summary Statistics

- Many of the most common statistical techniques (e.g., one and two sample t-tests, linear regression, analysis of variance) concern means.
- 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:
    1. *When Variability Is Important*
    2. *Skewed Distributions*
    3. *Ordinal Random Variables*
    4. *Unusual Events*

1. *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.
- Would you rather be Patient 1 or Patient 2?



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.

2. *Skewed Distributions* (*as time permits*)

A ***skewed distribution*** is one that is bunched up on one side and has a "tail" on the other. When a distribution is skewed, care needs to be given to choosing both an appropriate *measure of center* and an appropriate *measure of spread*.

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 a symmetrical 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*.
    - In this case, the *median* is a better measure of "typical".
    - See appendix for elaboration.
    - *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.

Examples:
1. 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]

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

3. See Limpert and Stahel (1998) for more examples.

Caution: There are applications where the mean of a distribution is used to estimate a population total (e.g., total crop yield or total health care burden) or as a proxy for that total. *In these situations, it is appropriate (and important) to estimate the mean, even if the distribution is skewed.*

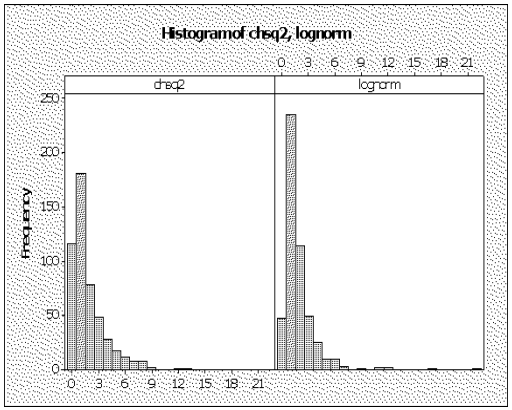
Measure of spread

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*.

Example: Both samples are from distributions with mean 2 and standard deviation 2 but the distributions are noticeably different.



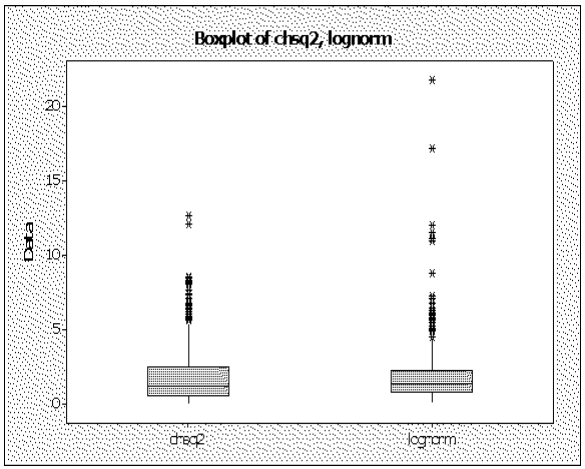
Common Mistake: Automatically giving *just* mean and standard deviation for a variable, without considering whether the variable might have a skewed distribution

- For a skewed distribution, it's usually better to use the *first and third quartiles*, as well as the medians, since these will give some sense of the asymmetry of the distribution.
- Boxplots typically give this information.
- But remember that some applications might require the mean, and perhaps the standard deviation. So give those as well.

In the example above, the quartiles are:

Graph	*First quartile*	*median*	*third quartile*
chsq2	0.5552	1.1895	2.5038
lognorm	0.8012	1.3428	2.2566

A boxplot:



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.
- 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**.
- This is one example of a more general **common mistake** of ignoring what are called *model assumptions*. (*More on this tomorrow.*)
- But remember the *caution:* In situations where a mean is used to estimate a total (or as a proxy for a population total), *it is important to give the mean and standard deviation, since these are important in estimating the population total.*

Suggestions for dealing with skewed distributions:

i. Always plot the data before applying a statistical test that assumes the variable has a normal distribution

- If the data are strongly skewed, use one of the suggestions below or another technique that does not require normality.

ii. If the goal is to estimate or give a proxy for a population total, be sure to follow guidelines of sampling theory in estimation.

- These will usually require use of the mean and standard deviation.
- They may also require careful planning of the type of sampling.
- If estimation of a population total is not the goal, consider the following possibilities.

iii. 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.
    - See Limpert and Stahel (1998) for examples and elaboration.
    - 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 transformation (e.g., square root, reciprocal) 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

- These include a variety of *permutation tests* (or *randomization tests*) as well as some standard named tests such as the Wilcoxon signed-rank test.
- Permutation tests (or a variation called Bootstrap tests) can also sometimes be "made to order" when an unusual test statistic is more appropriate than a standard one for the question under study.
- For more information, see Moore (2010), Eddington (1995), Good (2005)

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 at http://www.econ.uiuc.edu/~roger/research/rq/rq.html

3. *Ordinal Variables* (*if time permits*)

An *ordinal* variable is a categorical variable for which the possible values are ordered. Ordinal variables can be considered "in between" categorical and quantitative variables.

Example: Educational level might be categorized as

1: Elementary school education
2: High school graduate
3: Some college
4: College graduate
5: Graduate degree

- In this example (and for many ordinal variables), *the quantitative differences between the categories are uneven, even though the differences between the labels are the same*.
- *Thus it does not make sense to take a mean of the values*.
- ***Common mistake***: Treating ordinal variables like quantitative variables without thinking about whether this is appropriate in the particular situation at hand.
- For example, the "floor effect" can produce the appearance of interaction when using Least Squares Regression, when no interaction is present.
    - See Agresti (2010) for this example and for some methods that are appropriate for ordinal data.
- Permutation tests (or randomization tests) can also be used on ordinal data (See references on previous page.)

4. *Unusual Events* (*if time permits*)

If the research question being studied involves unusual events, *neither* the mean nor median is 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. 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.)

(See the Appendix for more examples and references involving unusual events).