6_misuses_of_stats.pdf	Michael Hallstone, Ph.D. hallston@hawaii.edu

Lecture 6: Misuses of Statistics

This lecture will cover a few of the ways in which statistics are misused.

BIAS OBSTACLE

With this problem the issue is not with the numbers themselves but the way in which the numbers are gathered, the way in which the data are collected. The data will be "junk" if they are collected badly. Many times there are biases inherent in the data collection process that produce poor data.

Bias Obstacle #1: Poorly Written Survey Questions Creating Biased Data

One way to have a bias inherent in the data collection process is to ask misleading questions in a survey. So the idea here is that poorly worded questions bias the responses that will be given on a survey. So a "bias obstacle" occurs because poorly worded questions ***create*** biased data.

There are literally many ways to write poor survey questions, but here are a few examples:

Sanders and Smidt (2000) write that members of congress have asked people if they agree with questions like:

"Do you favor elimination of waste in the defense budget?" Obviously the member got a very high "yes" response rate to that. The answers might have been much different if the member of congress asked people *which defense programs* they considered wasteful.

Likewise many voters support cutting spending to balance the budget – meaning they want government spending to match the amount of money the government collects in taxes. So if a poll asked, "Do you support a balanced budget?" it would get high levels of agreement to this question. In order to truly reduce the US deficit one must cut social security, Medicare, and/or defense – our largest expenditures – or increase taxes. When asked about cutting spending on any one of these individual programs you get far less agreement. So Americans want it both ways. They want low taxes and generally do not want to see cuts in defense, social security, or Medicare spending – thus the US borrows money each year.

"Loaded" or "leading" questions

Some words or phrases are "loaded" in certain cultures. So if you use a "loaded word" in a survey question, it will influence the responses you receive. You can get completely different responses about the same concept depending upon how you word the question. Try to use neutral language. People are more likely to support "government payments to the poor" than "welfare," "universal health insurance" than "socialized medicine." Also, people are less likely to "forbid" something than to "not allow" something (Neuman: 244).

prestige bias

Do not associate a particular answer with a prestigious individual or organization. People are likely to provide an answer that agrees with the authority figure.

"Do you agree with the Surgeon General that smoking should be against the law" will get a higher level of agreement than "Do you agree that smoking should be against the law."

"Do you agree with the president that we should raise taxes to lower the national debt" will receive a higher level of agreement than "do you agree that we should raise taxes to lower the national debt."

social desirability response

Social desirability response occurs when people are "unwilling to admit or to report accurately various behaviors or opinions because these are not considered to be socially acceptable. The challenge for researchers is to ask questions about socially sensitive issues in ways that illicit honest answers" (Foltz, 1996: 90). There are simply some questions where there is a "politically correct" answer. Do you go to church? Did you contribute to charity last year? How much do you weigh? How tall are you? How many alcoholic drinks do you consume each week? Did you vote in the last election? How many hours of TV do you allow your elementary school aged kids to watch each day? Have you ever taken office supplies home with you? All of these question will receive less-than-honest responses.

For example, research have compared actual voting records with survey data following an election have found discrepancies of about 15% -- typically non-voters claiming they voted. In plain English: survey researchers asked people if they voted and then compared their answers to actual voting records and found that about 15% who said "yes I voted" actually did not vote!

Bias Obstacle #2: Poor Sampling Creates Poor Data

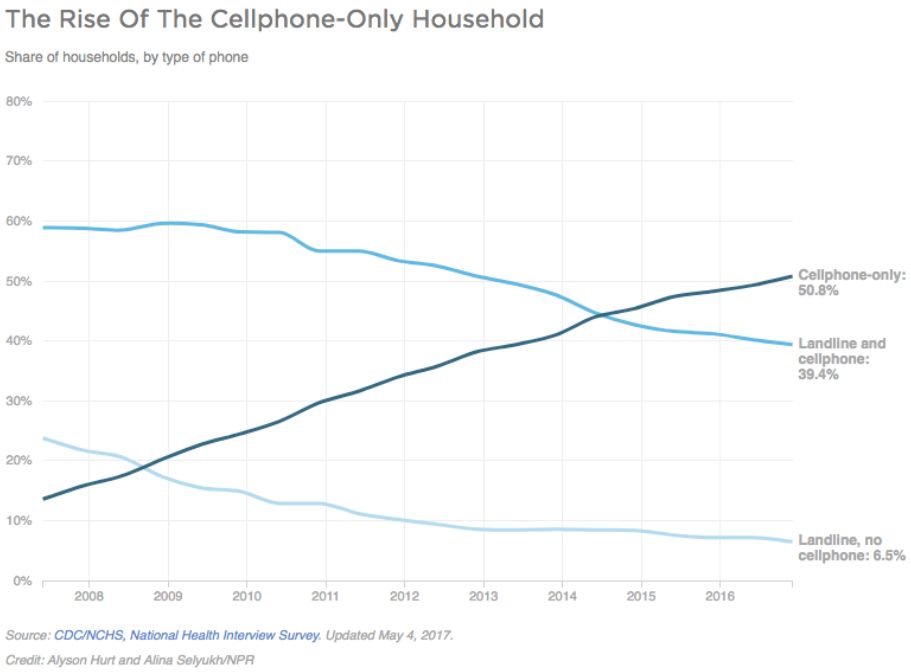
If the researcher uses a sampling frame that is a poor match to the population being studied, then sample itself can be biased – and again create biased data.

Sampling frame of people in alcohol treatment

What if I used a sampling frame of people in alcohol treatment to reach a population of "all alcohol users?" By definition people who are in treatment for any kind of drug treatment are going to have had very heavy usage patterns compared to the general population. I would expect questions like "how many drinks a day do you have?" to be quite skewed.

Sampling frame that excludes cell phones in 2008 election

Back in the day, about 95% of the household population had landline telephones. Therefore the books used to teach that landlines are a pretty good sampling frame. That has changed.



So imagine using landlines to reach the population of registered voters? What would be wrong with that? Well in 2008 when President Obama was elected some polls did that. As late as 2017 it was illegal for pollsters to use computerized random digit dialing to reach cell phones. You could use it to reach landlines, but to reach a cell phone you had to have a real person dial the number. [Hint: it is more expensive to pay a person than a computer to dial phone numbers.] So if you had a sample of registered voters that came from landlines in 2008, what type of voters would you systematically

miss? The young. Why? Because at that time the young were more likely to have completely dropped landline service and use a cell phone as their only phone.

Bias Obstacle #3: Poor Definitions of Key Variables

The idea here is that if a study uses "poor definitions" to operationalize (or define) a key variable, the data that is produced by that variable will be bad.

For example, during the 2008 presidential campaign, John McCain said something like a "rich" person was someone making over 100 million a year [let us hope he probably misspoke and did not mean it]. But pretend you worked for the Department of Labor and did a study to find out how many Americans were "rich" and you defined (or operationalized) the variable "rich" using that definition:

Rich = someone making 100 million a year or more

Not Rich = someone making less than 100 million

Well it would not take a rocket scientist to see that such a [poor] definition would produce data that "proves" there are *very few rich Americans*.

Using a real academic study example, one study found that "crack users living on the streets committed a lot of predatory crime. However, their definition of "crime" included drug sales! If you excluded the drug sales, the "level of crime" for the sample was much lower because a lot of "street users" sell drugs simply to earn enough money to feed their habit. Believe it or not selling drugs is not a predatory crime (it's a mutually agreed upon transaction) – we tend to think of violence or property crime when we think of predatory crime.

I will give you one more real example from best study on drug use in the US. The federal government defines "daily" drug use as….used 20 of 30 days out of the last month. Now certainly that would include those who used 30 out of last 30 days ("real" daily users in my opinion), but by also including those who used 2/3 of the days (2/3 of the days users in my opinion) INFLATES the number of "daily" drug users the data "produces."

To summarize, the idea here is that a study's operational definition can be bad and that will bias the data.

Bias Obstacle Conclusion

To recap, a bias obstacle is when something inherent in the data collection process creates "biased" or "poor" data. It can be either poorly worded questions or poor sampling, but the underlying problem remains the same: something in the data collection process produces poor data.

AGGRAVATING AVERAGES

Recall there are 3 "averages" in statistics : mean, median, and mode. As noted in the "measures of central tendency" there are assumptions when each should be used (for example only the mode can be used for a nominal level variable and you must have at least an interval level variable to compute the mean), but the biggest problem is that extreme scores distort the mean.

First I will give a common sense example, then some actual data. Pretend I wanted to know the "average income" of our class and for some strange reason Bill Gates [he started the software company Microsoft that sells Windows and is one of the richest men in the world] was in our class. His annual income is probably in the hundreds of millions of dollars and it would inflate our class mean income greatly: so the class "average income" would be like 45 million a year [when all other members of class make far less than $100,000 a year!]. So his extreme income would inflate the class mean income and create an "aggravated average." Below is an example with some numbers to help you see how this happens.

Example of extreme scores distorting the mean

data point	age	$ in wallet
X1	2	100
X2	3	110
X3	4	120
X4	4	120
X5	5	130
X6	5	130
X7	5	130
X8	6	180
X9	36	2000
MEAN	7.77777	335.555
MODE	5	130
MEDIAN	5	130

In both cases, there is an extreme value (in the x9 place) and you can see that the extreme score "inflates" the mean. (By the way extremely small scores would "deflate" the mean.) For *age* 7.7 is a really bad "single number used to describe all the numbers" and the same thing can be said for *$ in wallet:* $335.5 does a very poor job of representing all of the numbers in the sample. These are examples of the mean being distorted by extreme scores. Below is an example of a data set without extreme scores.

Example of data without extreme scores

data point	age	$ in wallet
X1	2	100
X2	3	110
X3	4	120
X4	4	120
X5	5	130
X6	5	130
X7	5	130
X8	6	180
Mean	4.25	127.5
median	4.5	125
mode	5	130

Note that in this case the mean is not distorted and probably does a fairly good job of representing the scores.

So in summary, if you choose the wrong "average" in statistics or use the mean that is distorted by extreme scores in the data set, you have the problem of "aggravating averages."

DISREGARDED DISPERSIONS

This concept is related to aggravated averages. The two often "go together" and happen at the same time."

The idea here is that averages do not tell the whole story – measures of dispersion must also be considered along with averages so we can get a true appreciation for "how the data looks." Recall an average is a "single number that is used to describe a group of numbers" and that single number cannot tell us how much variation [or dispersion] exists in the group of numbers. In order to appreciate whether or not the average is doing a good job of representing the whole group of numbers *we also need to have a measure of dispersion to tell us how much the group of numbers varies, or is dispersed, or "how spread out" the group is.*

Average Income if Bill Gates Were in this class

data point	income
X1	$10,000
X2	$15,000
X3	$25,000
X4	$20,000
X5	$24,000
X6	$25,000
X7	$40,000
X8	$100,000,000
Mean	$12,519,875
Median	$22,000

Above is the "Bill Gates" example I mentioned above. In this case I could say that the average income of students in my class is 12 million a year! Clearly, the mean is a poor average to use. It shows how we also need to know how much variation or dispersion exists in a data set to know whether or not the average is doing a good job of representing the data. Disregarded dispersions sort of "lead to" or "cause" aggravated averages. See how Disregarded Dispersions and Aggravated Averages are related?

Below is an example from investing. Mutual fund companies try to get you to invest in their funds by providing their average percentage return over time. Below is an example of $100 invested over five years when the five-year average was 1%. One might think this means the person "makes" 1% on their money over five years (or 1% compounded), but this is not the case. The mean return is 1%, but the great variability (or dispersion) causes the person to actually lose money.

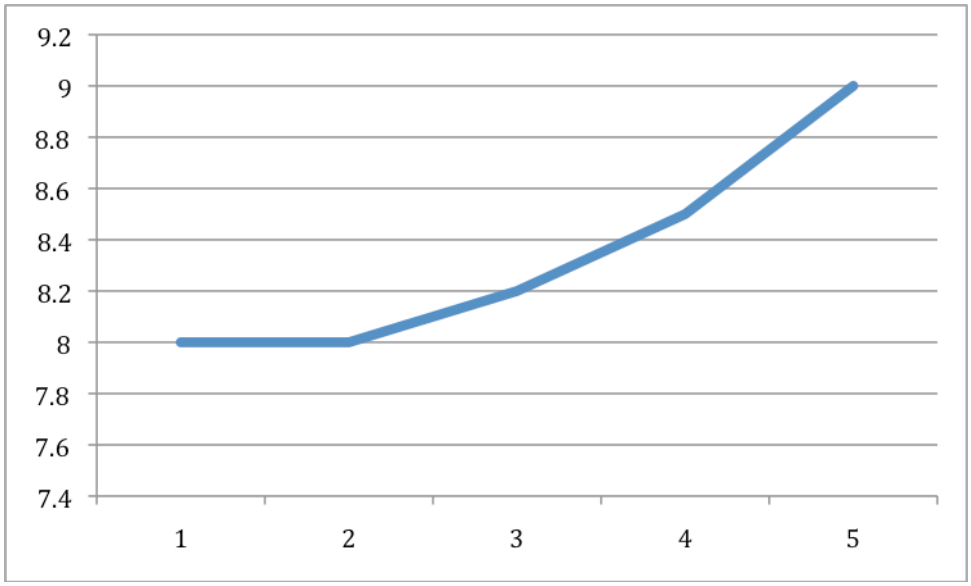
		Start with	$100.00
data point	% return	actual return	1% compounded
X1	10%	$110.00	$101.00
X2	10%	$121.00	$102.21
X3	10%	$133.10	$103.54
X4	10%	$146.41	$105.01
X5	-35%	$95.17	$105.96
mean	1%	(lost ~ 5%)	(gained ~5%)

In this case there is a whole lot of dispersion [or variability] over the five years: the fund did quite well for four years (returning 10% a year) but lost a ton of money in the fifth year. However a person might look at the "mean return over 5 years" which is 1% and think "Well, at least I would not have lost money; I would have received 1% a year for five years (or 1% compounded), giving me about $105.95 at the end of five years." But the great variability (or dispersion) shows that the actual return was negative. The person actually LOST money over five years. They started with $100.00 and ended up with $95.17. *So while the average is part of the picture, we also need to consider how spread out (or dispersed) the group of numbers is*.

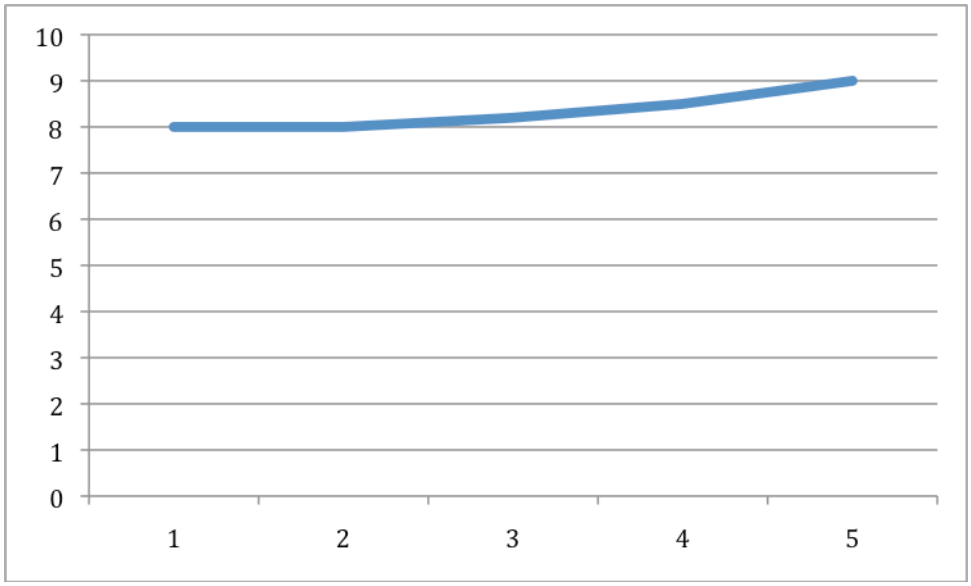
To summarize, by ignoring the dispersion that exist in a data set, we can be misled by "disregarded dispersions" [which is also related to aggravated averages].

THE PERSUASIVE ARTIST

If a person distorts the x or y-axis one can make the data appear dramatic when no such drama exists. Basically if you mess with the x and y-axis you can create the impression you want.



Sales are flat!



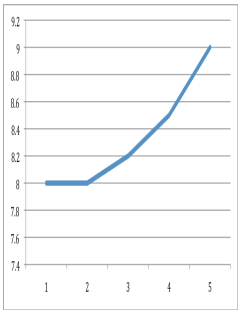
So if you do not look closely at the two charts, you can see that the “sales are up” chart appears to show a dramatic increase while the “sales are flat” chart seems to indicate nothing is going on. When you start the y-axis at something besides zero you are supposed to put to slash lines [//] across the y axis to indicate that you are using it “out of scale.”

But both charts were made using the same exact data. In the one above I compressed the y axis (vertical axis) and started at around 7% and in the other I did not distort the y-axis and started at zero. ***But each chart was made using the exact same data:***

% sales
8
8
8.2
8.5
9

Sales went up 1 percentage point in five years but the top chart creates a misleading impression.

By the way if you were to also compress x-axis in the "sales are up" chart you can make it seem even more dramatic. (The picture is poor but you get the idea.)



There are many ways to distort figures to be a "persuasive artist" and create a MISLEADING impression.

POST HOC ERGO PROPER HOC

(Correlation does not equal causation trap)

Just because A happens before B does not mean that A caused B. So when you commit this error you distort the dubious relationship between association (or correlation) and causation. Since "b follows a therefore a must cause b." The truth of the matter is that just because two things are associated or correlated it does NOT follow logically that there is a causal relationship.

There can be correlation, but not causation. And when this occurs typically there is a "third" (or several) intervening variables that create the causation.

"fruit sales and race riots"

This is the most obvious and easy to explain example. Apparently when fruit sales rise, so do the number of race riots. So does it mean that eating more fruit causes people to suddenly protest racism in their society? None of us would say so, but the two variables are associated or correlated. [By the way I have a common sense explanation for the correlation between these two variables. Fruit sales traditionally spike in the summer as that is when most fruit become ripe. Summer is also a time when it is hot and humid in the areas of the US where large urban ghettos exist. Poor people in these neighborhoods are out in the street because they are too poor to afford air conditioning and hot humid weather makes you cranky! But imagine January when it is say 10 degrees! No one is out on the streets because it is so cold so who the heck want to protest out in the streets when you are getting hypothermia! Now if you are already out in the streets because your apartment is hot as a furnace…well you get the idea.]

"drug use and crime"

It is true that most people sitting in prison used drugs at sometime in their lives, but it does not always mean that "drug use turns otherwise normal people into criminals." I will not bore you with the criminological methodological issues [like I do in "criminology"] but suffice it to say that creating a causal theory using the methodological rules of causation is very difficult in this instance. At the very least, for a causal relationship to exist, a person must have never engaged in any criminal activity prior to first using drugs. And this is to say nothing of the fact that criminal activity is sufficiently complex that there is no "one thing" that causes it. For example, the *overwhelming majority* of people who use drugs [even dangerous drugs], never engage in any type of criminal activity.

"gateway drugs"

Most people have heard of the "marijuana is a gateway drug" theory. The idea is that using marijuana "leads" people into using other "harder" drugs such as cocaine, heroin, methamphetamine, etc. The idea is that if they had never used marijuana then they could have avoided using all of those other drugs. It is true that marijuana is correlated with use of harder drugs later in life. But it does not logically follow that the marijuana use caused the person to use harder drugs. It is most likely that

there are a number of intervening variables that are *also closely associated with marijuana use* that are the causal factor. So for example, it would also be true that these hard drug users probably started their drug careers with tobacco and/or alcohol – *meaning that they used these before they ever used marijuana.* Very few people say tobacco or alcohol are "gateway drugs." So, there is a probably a set of intervening variables (like risk taking or liking extreme intoxication that are *also closely associated with marijuana use* that) that "cause" people to use harder drugs in life.

Wikipedia Example: shoe size and reading ability

Wikipedia had a great example: people with larger shoe sizes also score better on reading tests. Obviously, big feet do not cause some one to be a better reader. Can you think of the third or intervening variable that is also correlated with big feet and reading skill? Age. People become better readers as they age due to practice – and what happens to one's shoe size as one ages?

Conclusion of "correlation does not automatically mean causation"

Regardless of how it occurs, the error of "post hoc egro propter hoc" occurs when a person assumes that because A happens before B that means that A caused B. Remember two things can be associated or correlated, but that does NOT automatically mean that there is a causal relationship. (There could be a causal relationship, but the case is not "automatic" just because two things occur together.)

THE TREND MUST GO ON!

The idea here is that just cause something happened in the past it must continue to happen. Well that is not the case. Things that occur in the real world fluctuate. Just because something was true in the past 5 years, does NOT automatically mean that "thing" will continue to happen for the next 5 years.

The stock market is the perfect example. The late 1990's were one of the best times in recorded history to be invested in the stock market. People were routinely making 20%, 30%, 40% (and more!) a year. If you were to look at the trend in the stock market from say 1996-2000 you would see a steep upward curve. However the trend reversed dramatically in 2000-2001. This is why if you look at the fine print in stock market adds you see something like "past performance is not a guarantee of future performance." The upward trend did not continue.

The same happened with the “housing bubble” of 2007-2008. From about 2003-2007 housing prices increased dramatically in the US, but the trend did not continue: 2007-2008 saw housing prices drop up considerably in the US. The upward trend did not continue.

A few alcoholic drinks might increase the sociability of a person but as drinks increase to the point of being drunk then we would see sociability decrease. The same can be said for caffeine and mental capability: a little caffeine increases it, but you can not just keep drinking coffee and become “super smart” in fact as anyone who has had too many lattes at Starbucks can tell you, too much caffeine lowers mental capabilities. The same is true with wave height and # of surfers in water. We would expect the number of surfers in the water tor rise to a certain extent, but say after 12-15 ft (Hawaiian scale)the number of surfers in the water actually drops off because not many folks have the courage and or ability to go out in truly huge surf. There are many examples of “curvilinear relationships” in the social world. IN PERSON CLASS DRAW THESE GRAPHS:

So the trend does NOT automatically continue!

Practice

Everything that appears in these lecture notes is fair game for the test. They are the best “study guide” I can provide. It is impossible to provide a “list” that is more comprehensive than the lecture notes above. However, here are a few additional practice exercises or practice concepts.

For this lecture you will need to be able to identify each of the following “misuses of statistics” when provided an example: bias obstacle, aggravating averages, disregarded dispersions, persuasive artist, *post hoc ergo proper hoc*, the trend must go on.