Real Statistics: Your Antidote to "Stat 101"

Norm Matloff
Department of Computer Science
University of California at Davis
http://heather.cs.ucdavis.edu/matloff.html

Walnut Creek Library
April 26, 2011
These slides available at
`http://heather.cs.ucdavis.edu/realstat.pdf`.

Goals

Goals

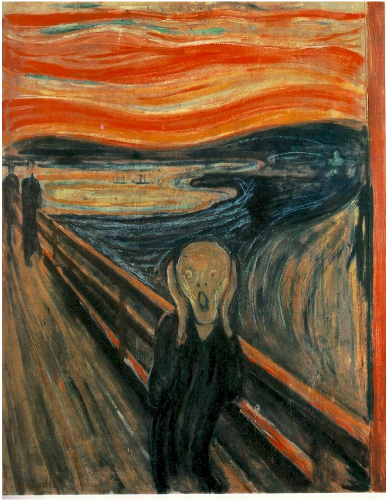
GOAL I: Demolish most people's images of statistics:

Goals

GOAL I: Demolish most people's images of statistics:

Goals

GOAL I: Demolish most people's images of statistics:



Goals, cont'd.

Goals, cont'd.

GOAL II: Show modern uses of statistics.

Goals, cont'd.

GOAL II: Show modern uses of statistics.

GOAL III: Expose common statistical fallacies

Goals, cont'd.

GOAL II: Show modern uses of statistics.

GOAL III: Expose common statistical fallacies—<u>especially</u> in Stat 101.

Goals, cont'd.

GOAL II: Show modern uses of statistics.

GOAL III: Expose common statistical fallacies—especially in Stat 101.

GOAL IV: Show how you can do your own statistics, using the Web and free software.

Goals, cont'd.

GOAL II: Show modern uses of statistics.

GOAL III: Expose common statistical fallacies—especially in Stat 101.

GOAL IV: Show how you can do your own statistics, using the Web and free software.
Not a methods course. Suggestions later.

History of Statistics: the Elevator Speech

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.
- Least-squares fitting of lines to data, 1794, Gauss.

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.
- Least-squares fitting of lines to data, 1794, Gauss.
- Agricultural research, Sir Ronald Fisher, 1920s.

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.
- Least-squares fitting of lines to data, 1794, Gauss.
- Agricultural research, Sir Ronald Fisher, 1920s.
- Modern mathematical era developed by many in the 1950s, 60s, with Jerzy Neyman of UC Berkeley arguably the pioneer.

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.
- Least-squares fitting of lines to data, 1794, Gauss.
- Agricultural research, Sir Ronald Fisher, 1920s.
- Modern mathematical era developed by many in the 1950s, 60s, with Jerzy Neyman of UC Berkeley arguably the pioneer.
- Space race, medical research give the field a big boost, 1970s.

History of Statistics: the Elevator Speech

- Analysis of gambling, 1700s, e.g. Demoivre.
- Least-squares fitting of lines to data, 1794, Gauss.
- Agricultural research, Sir Ronald Fisher, 1920s.
- Modern mathematical era developed by many in the 1950s, 60s, with Jerzy Neyman of UC Berkeley arguably the pioneer.
- Space race, medical research give the field a big boost, 1970s.
- "New" applications (e.g. social network analysis), very fast/cheap computers radically changing things today.

Statistics, Old and New

Statistics, Old and New

Old applications:

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

Statistics, Old and New

New applications:

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.
- Machine speech recognition, computer vision.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.
- Machine speech recognition, computer vision.
- Search: Google, Jeopardy playing computer, etc.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.
- Machine speech recognition, computer vision.
- Search: Google, Jeopardy playing computer, etc.
- Marketing, e.g. Amazon recommendation system.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.
- Machine speech recognition, computer vision.
- Search: Google, Jeopardy playing computer, etc.
- Marketing, e.g. Amazon recommendation system.
- Analysis of social networks.

Statistics, Old and New

Old applications:

- Compare 4 varieties of wheat.
- Formalize obscure academic research studies.
- Economic forecasting.
- Medical research.

New applications:

- Mapping human genome; genetic counseling.
- Machine speech recognition, computer vision.
- Search: Google, Jeopardy playing computer, etc.
- Marketing, e.g. Amazon recommendation system.
- Analysis of social networks.
- (Some of this stuff is scary.)

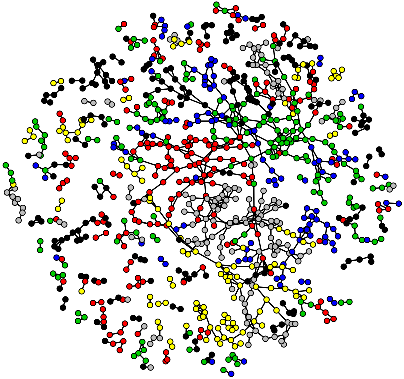
Impact of Having Fast (and Cheap) Computers

Impact of Having Fast (and Cheap) Computers

Example: *Exponential random graph model* of social relations at a high school. (Sorry, no details here.)

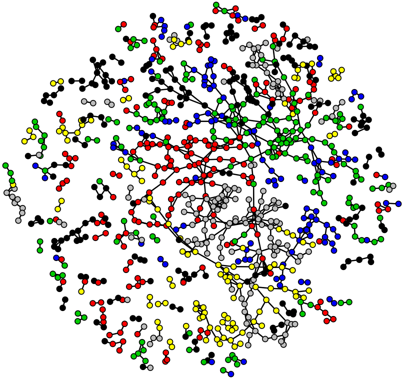
Impact of Having Fast (and Cheap) Computers

Example: *Exponential random graph model* of social relations at a high school. (Sorry, no details here.)



Impact of Having Fast (and Cheap) Computers

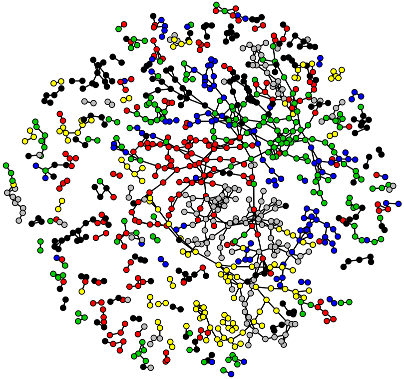
Example: *Exponential random graph model* of social relations at a high school. (Sorry, no details here.)



Took only about 30 seconds to do complex compute and graph.

Impact of Having Fast (and Cheap) Computers

Example: *Exponential random graph model* of social relations at a high school. (Sorry, no details here.)



Took only about 30 seconds to do complex compute and graph.

Same methodology used for protein molecular analysis, etc.

Computation for the Masses

Computation for the Masses

- You can do big-data statistics.

Computation for the Masses

- You can do big-data statistics.
- Even the cheapest PC is far more powerful than the old mainframes.

Computation for the Masses

- <u>You</u> can do big-data statistics.
- Even the cheapest PC is far more powerful than the old mainframes.
- Sophisticated, professional software is free: R, discussed later.

Computation for the Masses

- <u>You</u> can do big-data statistics.
- Even the cheapest PC is far more powerful than the old mainframes.
- Sophisticated, professional software is free: R, discussed later.
- Interesting real data is abundant on the Web.

Computation for the Masses

- You can do big-data statistics.
- Even the cheapest PC is far more powerful than the old mainframes.
- Sophisticated, professional software is free: R, discussed later.
- Interesting real data is abundant on the Web.
- Why are the high schools still teaching statistics on pocket calculators?

Even the Old Is New!

Even the Old Is New!

Example: Heritage Health Prize

Even the Old Is New!

Example: Heritage Health Prize

- Develop algorithm to predict who will need a hospital stay during the next year. This is an **old** application.

Even the Old Is New!

Example: Heritage Health Prize

- Develop algorithm to predict who will need a hospital stay during the next year. This is an **old** application.
- This is a statistics problem, though most contestants will be using **new** statistical methods.

Even the Old Is New!

Example: Heritage Health Prize

- Develop algorithm to predict who will need a hospital stay during the next year. This is an **old** application.
- This is a statistics problem, though most contestants will be using **new** statistical methods.
- $3 million prize to the winner. **This is new!**

Even the Old Is New!

Example: Heritage Health Prize

- Develop algorithm to predict who will need a hospital stay during the next year. This is an **old** application.
- This is a statistics problem, though most contestants will be using **new** statistical methods.
- $3 million prize to the winner. **This is new!**
- Anyone can enter, `http://www.heritagehealthprize.com/c/hhp`—sign up today!

Even the Statistics Contests Are a Business!

Even the Statistics Contests Are a Business!

- There are so many of these contests that Australian Anthony Goldbloom started a company, Kaggle, to manage them.

Even the Statistics Contests Are a Business!

- There are so many of these contests that Australian Anthony Goldbloom started a company, Kaggle, to manage them.
- Check out the contests, `www.kaggle.com`, and *Forbes* article on Kaggle, http://blogs.forbes.com/tomiogeron/2011/04/04/kaggles-predictive-data-contest-aims-to-fix-health-care/

Even the Statistics Contests Are a Business!

- There are so many of these contests that Australian Anthony Goldbloom started a company, Kaggle, to manage them.
- Check out the contests, `www.kaggle.com`, and *Forbes* article on Kaggle, http://blogs.forbes.com/tomiogeron/2011/04/04/kaggles-predictive-data-contest-aims-to-fix-health-care/
- Chris Raimondi, self taught in machine learning by watching YouTube (!), beat out a team from IBM Research for first place in one contest.

Much That Looks New Is Not Really

Much That Looks New Is Not Really

These days there are various "new" fields that are really statistics:

Much That Looks New Is Not Really

These days there are various “new” fields that are really statistics:

- Machine learning (automatic prediction).

Much That Looks New Is Not Really

These days there are various “new” fields that are really statistics:

- Machine learning (automatic prediction).
- Data mining (statistical fishing expedition).

Much That Looks New Is Not Really

These days there are various “new” fields that are really statistics:

- Machine learning (automatic prediction).
- Data mining (statistical fishing expedition).
- Analytics (anything business finds useful, often for marketing).

Much That Looks New Is Not Really

These days there are various "new" fields that are really statistics:

- Machine learning (automatic prediction).
- Data mining (statistical fishing expedition).
- Analytics (anything business finds useful, often for marketing).
- Methods are more specialized, and much more computationally intensive, but basically variations on old ones.

Real Statistics

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

- significance testing (a Bad Thing), confidence intervals

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

- significance testing (a Bad Thing), confidence intervals
- covariates

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

- significance testing (a Bad Thing), confidence intervals
- covariates

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

- significance testing (a Bad Thing), confidence intervals
- covariates

Really, everything else is just variations on a theme.

Real Statistics

Being able to **UNDERSTAND**—not just know formulas—and use statistics boils down to just two main concepts:

- significance testing (a Bad Thing), confidence intervals
- covariates

Really, everything else is just variations on a theme.
But one must really understand these two concepts.

Statistical Pitfalls

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data <u>seems</u> to indicate new drug is better.

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data <u>seems</u> to indicate new drug is better.
- But could it be a sampling accident? (E.g. the new drug happened to be assigned to healthier patients.)

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data seems to indicate new drug is better.
- But could it be a sampling accident? (E.g. the new drug happened to be assigned to healthier patients.)
- Computer calculates *p-value* (defined below), say 0.02.

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data seems to indicate new drug is better.
- But could it be a sampling accident? (E.g. the new drug happened to be assigned to healthier patients.)
- Computer calculates *p-value* (defined below), say 0.02.
- You then say (more or less),

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data seems to indicate new drug is better.
- But could it be a sampling accident? (E.g. the new drug happened to be assigned to healthier patients.)
- Computer calculates *p-value* (defined below), say 0.02.
- You then say (more or less),

  > If the two drugs were equally effective, there would only be a 2% chance of getting the data we have. So we doubt that they are equally effective, and conclude that **they are significantly different**.

Statistical Pitfalls

First, the Mother of All Statistical Fallacies—*significance testing*.

- Example: Compare old, new drugs for hypertension.
- Suppose data <u>seems</u> to indicate new drug is better.
- But could it be a sampling accident? (E.g. the new drug happened to be assigned to healthier patients.)
- Computer calculates *p-value* (defined below), say 0.02.
- You then say (more or less),

  > If the two drugs were equally effective, there would only be a 2% chance of getting the data we have. So we doubt that they are equally effective, and conclude that **they are significantly different**.

- This is **the very core of statistics**—yet it's a Bad Thing.

History of Objections to Significance Testing

History of Objections to Significance Testing

- Significance testing very old, developed by Sir Ronald Fisher in the 1920s.

History of Objections to Significance Testing

- Significance testing very old, developed by Sir Ronald Fisher in the 1920s.
- *"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"*—Paul Meehl, professor of psychology and the philosophy of science, 1978

History of Objections to Significance Testing

- Significance testing very old, developed by Sir Ronald Fisher in the 1920s.
- *"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"*—Paul Meehl, professor of psychology and the philosophy of science, 1978
- There was opposition even during Sir Fisher's time.

History of Objections to Significance Testing

- Significance testing very old, developed by Sir Ronald Fisher in the 1920s.
- *"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"*—Paul Meehl, professor of psychology and the philosophy of science, 1978
- There was opposition even during Sir Fisher's time.
- But...Knights prevail, right? :-)

History of Objections to Significance Testing

- Significance testing very old, developed by Sir Ronald Fisher in the 1920s.
- *"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"*—Paul Meehl, professor of psychology and the philosophy of science, 1978
- There was opposition even during Sir Fisher's time.
- But...Knights prevail, right? :-)
- So, it is widely recognized as problematic today—yet solidly entrenched.

So, What's Wrong with Significance Testing?

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.
- So, the consultant is 95% confident (details later) that Obama's support is currently between 47% and 83%.

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.
- So, the consultant is 95% confident (details later) that Obama's support is currently between 47% and 83%.
- The consultant will be thrilled! Granted, part of that interval is below 50%, but most of it is well above 50%.

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.
- So, the consultant is 95% confident (details later) that Obama's support is currently between 47% and 83%.
- The consultant will be thrilled! Granted, part of that interval is below 50%, but most of it is well above 50%.
- And yet...

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.
- So, the consultant is 95% confident (details later) that Obama's support is currently between 47% and 83%.
- The consultant will be thrilled! Granted, part of that interval is below 50%, but most of it is well above 50%.
- And yet... a significance test would find "There is no statistically significant difference in support between Obama and X."

So, What's Wrong with Significance Testing?

To see the problem, picture a consultant to Obama's campaign in the 2012 election. His opponent is X.

- The results of a small poll are just in: 65% favor Obama, with a margin of error of 18%.
- So, the consultant is 95% confident (details later) that Obama's support is currently between 47% and 83%.
- The consultant will be thrilled! Granted, part of that interval is below 50%, but most of it is well above 50%.
- And yet... a significance test would find "There is no statistically significant difference in support between Obama and X."
- Do you really believe that???? The test is leading us astray.

What's Wrong, cont'd.

What's Wrong, cont'd.

The opposite situation is disturbing too:

What's Wrong, cont'd.

The opposite situation is disturbing too:

- Say the interval is 50.2% to 50.7%.

What's Wrong, cont'd.

The opposite situation is disturbing too:

- Say the interval is 50.2% to 50.7%.
- The significance test says, "Obama has significantly more support than X."

What's Wrong, cont'd.

The opposite situation is disturbing too:

- Say the interval is 50.2% to 50.7%.
- The significance test says, "Obama has significantly more support than X."
- Should the consultant be thrilled? No! Obama's support in this situation is razor-thin. It could change tomorrow.

What's Wrong, cont'd.

The opposite situation is disturbing too:

- Say the interval is 50.2% to 50.7%.
- The significance test says, "Obama has significantly more support than X."
- Should the consultant be thrilled? No! Obama's support in this situation is razor-thin. It could change tomorrow.
- Once again, the test has fooled us.

What Went Wrong?

What Went Wrong?

- The math theory underlying testing is fine.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.
- In the second example above, the significance test is addressing the question whether Obama's support is $> 50\%$ *by any amount at all, large or small*.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.
- In the second example above, the significance test is addressing the question whether Obama's support is $> 50\%$ *by any amount at all, large or small*.
- Its answer there—Yes—was highly misleading.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.
- In the second example above, the significance test is addressing the question whether Obama's support is $> 50\%$ *by any amount at all, large or small*.
- Its answer there—Yes—was highly misleading. It didn't tell us that the support was *just barely* above 50%.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.
- In the second example above, the significance test is addressing the question whether Obama's support is $> 50\%$ *by any amount at all, large or small*.
- Its answer there—Yes—was highly misleading. It didn't tell us that the support was *just barely* above 50%.
- In the first example the answer—No—didn't tell us that Obama's support could be huge.

What Went Wrong?

- The math theory underlying testing is fine.
- But the test isn't answering the real question of interest.
- In the second example above, the significance test is addressing the question whether Obama's support is $> 50\%$ *by any amount at all, large or small*.
- Its answer there—Yes—was highly misleading. It didn't tell us that the support was *just barely* above 50%.
- In the first example the answer—No—didn't tell us that Obama's support could be huge.
- Also: That word "significant" should NOT be taken as meaning "important."

So, What to Do?

So, What to Do?

People want simple answers—even if wrong ones.

So, What to Do?

People want simple answers—even if wrong ones.
“Preponderance of evidence.”

Significance Tests Shouldn't Be Used at All

Significance Tests Shouldn't Be Used at All

Significance tests are simply the wrong way to go.

Significance Tests Shouldn't Be Used at All

Significance tests are simply the wrong way to go.

- At worst highly misleading, at best underinformative.

Significance Tests Shouldn't Be Used at All

Significance tests are simply the wrong way to go.

- At worst highly misleading, at best underinformative.
- Reporting a *confidence interval*—the point estimate plus/minus the margin of error—is much better. (E.g. 65% $\pm$ 18% above.)

Significance Tests Shouldn't Be Used at All

Significance tests are simply the wrong way to go.

- At worst highly misleading, at best underinformative.
- Reporting a *confidence interval*—the point estimate plus/minus the margin of error—is much better. (E.g. 65% $\pm$ 18% above.)
- Though, of course in some cases one is "forced" to use significance tests, say by a government agency.

Meaning of Confidence Level

Meaning of Confidence Level

- A margin of error is usually given at the 95% confidence level.

Meaning of Confidence Level

- A margin of error is usually given at the 95% confidence level.
- It's necessary to have a confidence level necessary because one is dealing with samples.

Meaning of Confidence Level

- A margin of error is usually given at the 95% confidence level.
- It's necessary to have a confidence level necessary because one is dealing with samples.
- The 95% means that, in 95% of all possible samples, your sample estimate will be within the margin of error of the true population value.

Next Big Pitfall: the Effects of Covariates

Next Big Pitfall: the Effects of Covariates

- No "primrose path" remarks here; everyone agrees about the importance of covariates.

Next Big Pitfall: the Effects of Covariates

- No "primrose path" remarks here; everyone agrees about the importance of covariates.
- Say you are studying some variable Y. It may be necessary to bring in one or more other variables in order to properly study Y.

Next Big Pitfall: the Effects of Covariates

- No "primrose path" remarks here; everyone agrees about the importance of covariates.
- Say you are studying some variable Y. It may be necessary to bring in one or more other variables in order to properly study Y.
- Or, say you are studying the relation between variables Y and X. To properly study the relation, you may need to bring in a third variable, or more.

Next Big Pitfall: the Effects of Covariates

- No "primrose path" remarks here; everyone agrees about the importance of covariates.
- Say you are studying some variable Y. It may be necessary to bring in one or more other variables in order to properly study Y.
- Or, say you are studying the relation between variables Y and X. To properly study the relation, you may need to bring in a third variable, or more.
- Those other variables are called *covariates*.

Example: Kaiser Consulting

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.
- So, measuring the relation between Y and X here means comparing the 4 hospitals in terms of heart attack survival rates.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.
- So, measuring the relation between Y and X here means comparing the 4 hospitals in terms of heart attack survival rates.
- But 1 of the 4 served an area with a lot of elderly patients.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.
- So, measuring the relation between Y and X here means comparing the 4 hospitals in terms of heart attack survival rates.
- But 1 of the 4 served an area with a lot of elderly patients. Thus direct comparison of the 4 hospitals would be unfair.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.
- So, measuring the relation between Y and X here means comparing the 4 hospitals in terms of heart attack survival rates.
- But 1 of the 4 served an area with a lot of elderly patients. Thus direct comparison of the 4 hospitals would be unfair.
- Thus need to bring in a covariate, $Z =$ age.

Example: Kaiser Consulting

My first consulting project, evaluating 4 LA Kaiser hospitals.

- Here Y was survival after a heart attack. $Y = 1$ means survive, $Y = 0$ means not.
- X was the hospital ID, numbered say from 1 to 4.
- So, measuring the relation between Y and X here means comparing the 4 hospitals in terms of heart attack survival rates.
- But 1 of the 4 served an area with a lot of elderly patients. Thus direct comparison of the 4 hospitals would be unfair.
- Thus need to bring in a covariate, Z = age. I.e., measure the relation between Y and X, holding Z constant.

Why Are Covariates So Important?

A correlation between variables Y and X can change from positive to negative, or *vice versa*, once a covariate Z is accounted for.

Why Are Covariates So Important?

A correlation between variables Y and X can change from positive to negative, or *vice versa*, once a covariate Z is accounted for. Known as "Simpson's Paradox."

Example of Simpson's Paradox

Example UC Berkeley gender bias claim.[1]

[1] Adapted from http://www.math.upenn.edu/ kazdan/210/gradadmit.html

Example of Simpson's Paradox

Example UC Berkeley gender bias claim.[1]

dept.	M app.	M admit.	F app.	F admit.
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
total	2318	51%	1494	35%

- In every department, F admission rate similar to or $>$ M rate.

[1] Adapted from http://www.math.upenn.edu/ kazdan/210/gradadmit.html

Example of Simpson's Paradox

Example UC Berkeley gender bias claim.[1]

dept.	M app.	M admit.	F app.	F admit.
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
total	2318	51%	1494	35%

- In every department, F admission rate similar to or > M rate.
- Yet <u>overall</u> F rate much lower than M.

[1] Adapted from http://www.math.upenn.edu/ kazdan/210/gradadmit.html

Example of Simpson's Paradox

Example UC Berkeley gender bias claim.[1]

dept.	M app.	M admit.	F app.	F admit.
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
total	2318	51%	1494	35%

- In every department, F admission rate similar to or $>$ M rate.
- Yet <u>overall</u> F rate much lower than M.
- Reason: Fs applied to tougher departments than Ms.

[1] Adapted from http://www.math.upenn.edu/ kazdan/210/gradadmit.html

Example of Simpson's Paradox

Example UC Berkeley gender bias claim.[1]

dept.	M app.	M admit.	F app.	F admit.
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
total	2318	51%	1494	35%

- In every department, F admission rate similar to or $>$ M rate.
- Yet overall F rate much lower than M.
- Reason: Fs applied to tougher departments than Ms.
- The point: Doing an analysis that did NOT account for the department covariate would have been misleading.

[1] Adapted from http://www.math.upenn.edu/ kazdan/210/gradadmit.html

The R Statistical Language

The R Statistical Language

The R Statistical Language

We are fortunate to have a professional quality, **FREE** (open source) statistical language available—R.

The R Statistical Language

We are fortunate to have a professional quality, **FREE** (open source) statistical language available—R.

You can use the same software used at Google, NIH etc.!

The R Statistical Language

We are fortunate to have a professional quality, **FREE** (open source) statistical language available—R.

You can use the same software used at Google, NIH etc.!

You do NOT have to be a programmer to use it; just be patient and learn a bit at a time.

A Short R Example

A Short R Example

Can only just scratch the surface here...

A Short R Example

Can only just scratch the surface here...

Example: Data on forest fires in Portugal.

A Short R Example

Can only just scratch the surface here...

Example: Data on forest fires in Portugal.

Read in data from Web, find CI for the mean temperature, plot area burned versus temperature, and do regression prediction of area burned from temperature, humidity and wind.

A Short R Example

Can only just scratch the surface here...

Example: Data on forest fires in Portugal.

Read in data from Web, find CI for the mean temperature, plot area burned versus temperature, and do regression prediction of area burned from temperature, humidity and wind.

(Plot, prediction output not shown.)

```
> frs <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/for
> t.test(frs$temp)
...
95 percent confidence interval:
 18.38747 19.39087
...
> plot(frs$temp,frs$area)
> lm(frs$area ~ frs$temp + frs$RH + frs$wind)
```

Where to Go From Here?

Where to Go From Here?

Some resources:

Where to Go From Here?

Some resources:

- *Introductory Statistics with R*, by Peter Dalgaard. Thin paperback. Learn stat and R, gently. I recommend Chapters 2-6, 8, 10, 11, 13.

Where to Go From Here?

Some resources:

- *Introductory Statistics with R*, by Peter Dalgaard. Thin paperback. Learn stat and R, gently. I recommend Chapters 2-6, 8, 10, 11, 13.
- *Reference Guide on Statistics*, by D. Kaye and D. Freedman. Free, on Web at `ftp.resource.org/courts.gov/fjc/sciam.0.stats.pdf`. Commissioned by U.S. Supreme Court to educate judges. Statistically correct! (Many books are not.)

Where to Go From Here?

Some resources:

- *Introductory Statistics with R*, by Peter Dalgaard. Thin paperback. Learn stat and R, gently. I recommend Chapters 2-6, 8, 10, 11, 13.
- *Reference Guide on Statistics*, by D. Kaye and D. Freedman. Free, on Web at `ftp.resource.org/courts.gov/fjc/sciam.0.stats.pdf`. Commissioned by U.S. Supreme Court to educate judges. Statistically correct! (Many books are not.)
- *Statistics*, by D. Freedman, R. Purves, R. Pisani. Also statistically correct, and engaging. But $113?

Where to Go From Here?

Some resources:

- *Introductory Statistics with R*, by Peter Dalgaard. Thin paperback. Learn stat and R, gently. I recommend Chapters 2-6, 8, 10, 11, 13.
- *Reference Guide on Statistics*, by D. Kaye and D. Freedman. Free, on Web at `ftp.resource.org/courts.gov/fjc/sciam.0.stats.pdf`. Commissioned by U.S. Supreme Court to educate judges. Statistically correct! (Many books are not.)
- *Statistics*, by D. Freedman, R. Purves, R. Pisani. Also statistically correct, and engaging. But $113?
- *The Art of R Programming*, by N. Matloff, NSP, forthcoming.

Where to Go From Here?

Some resources:

- *Introductory Statistics with R*, by Peter Dalgaard. Thin paperback. Learn stat and R, gently. I recommend Chapters 2-6, 8, 10, 11, 13.
- *Reference Guide on Statistics*, by D. Kaye and D. Freedman. Free, on Web at `ftp.resource.org/courts.gov/fjc/sciam.0.stats.pdf`. Commissioned by U.S. Supreme Court to educate judges. Statistically correct! (Many books are not.)
- *Statistics*, by D. Freedman, R. Purves, R. Pisani. Also statistically correct, and engaging. But $113?
- *The Art of R Programming*, by N. Matloff, NSP, forthcoming.
- *The Numbers Guy*, by Carl Bialik. Excellent weekly column on statistics in the *Wall Street Journal*.