# Data Acquisition and Statistical Analysis

Facts are stubborn, but statistics are more pliable.
- Mark Twain

Laura Saba, PhD
Research Assistant Professor
Department of Pharmaceutical Sciences
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Denver
Laura.Saba@ucdenver.edu

# Outline

- Data Acquisition
  - Collection
  - Storage
  - Ownership/Sharing
- Statistical Analysis
  - Outliers and other annoyances
  - Suitable, better, and best methods for analysis
  - Display and interpretation of results

# Reproducible Research

"the idea that the ultimate product of research is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. necessary for reproduction of the results and building upon the research"

- *Wikipedia*

# Reproducible Research

## Reproducible Results

- Begley, CG and Ellis, LM Nature 29 Mar 2012
  - Of 53 "landmark" studies, only 6 (11%) were reproducible
  - "In studies for which findings could be reproduced, authors paid close attention to controls, reagents, investigator bias and describing the complete data set"

## Reproducible Research

- Ioannidis, JPA et al Nature Genetics Feb 2009
  - Replicated analysis of 18 microarray studies in Nature Genetics '05-'06
    - 2 replicated in principle
    - 6 partially replicated
    - 10 not replicated

# Data Acquisition

Data Collection

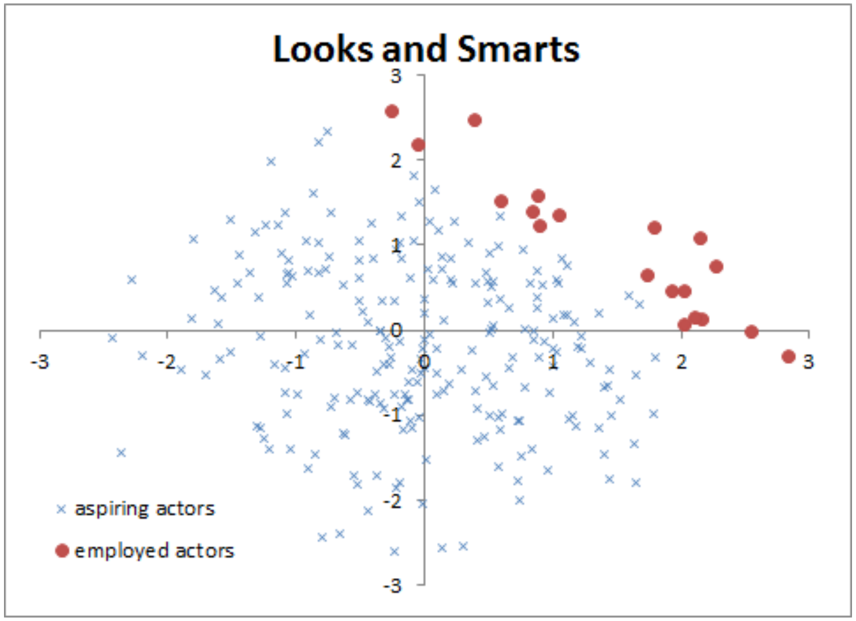
# Data Collection

1. Appropriate Methods
   - Garbage in, garbage out (biased data collection, e.g., sample selection, biased results)
2. Attention to detail
   - Accuracy in recording, interpretation, publications
3. Authorized
   - HIPAA, hazardous materials, copyrights, etc.
4. Recording
   - **Hard copy evidence** should be entered into a numbered, bound notebook
   - **Electronic evidence** should be validated in some way to assure that it was actually recorded on a particular date and not changed at some later data
   - Not only should data derived from the research be accurately recorded, but also detailed information on procedures including materials used, e.g., chemical agents.

Taken from *ORI Introduction to RCR* (http://ori.hhs.gov/education/products/

# Example – Appropriate Data Collection

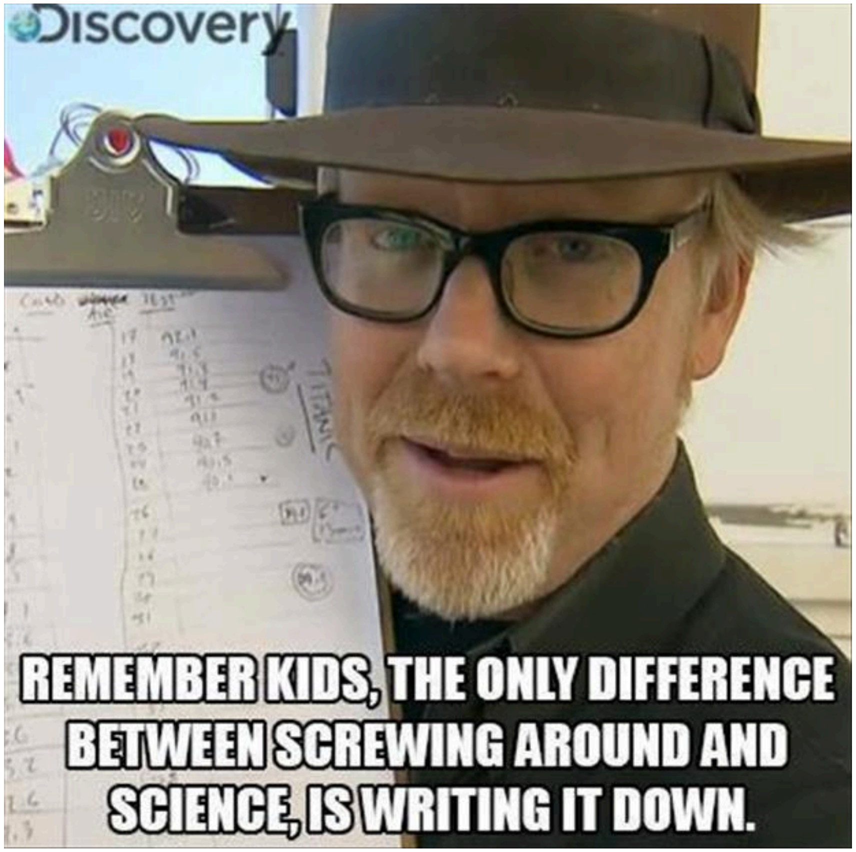
- If good looks and smarts are distributed normally,
- If good looks and smarts **have nothing to do with each other**
- If movie producers want both smarts and looks Then, by observing **employed actors** we'll assume that looks and smarts have a **negative correlation**



http://euri.ca/2012/youre-probably-polluting-your-statistics-more-than-you-think/index.html
http://www.theatlantic.com/business/archive/2012/05/when-correlation-is-not-causation-but-something-much-more-screwy/256918/

# Reasons to keep accurate records

- Reproducibility
- Future analyses
- Investigations of misconduct
- Proving ownership of intellectual properties
- Others?


Discovery
REMEMBER KIDS, THE ONLY DIFFERENCE BETWEEN SCREWING AROUND AND SCIENCE, IS WRITING IT DOWN.

# Case Study

## from *Responsible Conduct of Research*

- Dr. Z is mentoring a “promising” medical student over the summer in his research lab
- Student’s project:
  - cancer cell line that requires 3 weeks to grow in order to test for a specific antibody
  - the student has already written a short paper on his work
- Dr. Z’s dilemma:
  - after going over the raw data, some data were on pieces of yellow pads without clear identification from which experiment the data came
  - some of the experiments were repeated several times without explanation as to way
  - Dr. Z is not happy about the data, but doesn’t want to discourage the student from pursuing a career in research

# Case Study

## from *Responsible Conduct of Research*

- Dr. Z is mentoring a "promising" medical student over the summer in his research lab
- Student's project
  - cancer cell line that requires 3 weeks to grow in order to test for a specific antibody
  - the student has already written a short paper on his work
- Dr. Z's dilemma:
  - after going over the raw data, some data were on pieces of yellow pads without clear identification from which experiment the data came
  - some of the experiments were repeated several times without explanation as to way
  - she is not happy about the data, but doesn't want to discourage him to pursue a career in research

- What is the primary responsibility of the mentor?
- Should the mentor write a short paper and send it for publication?
- Should the student write a short paper and send it for publication?
- If you were the mentor, what would you do?

# Data Acquisition

Data Storage

# Data Storage

"Over time, data, as the currency of research, become an investment in research. If the data are not properly protected, the investment, whether public or private, could become worthless"

– *ORI Introduction to RCR*

# Considerations When Storing Data/ Research

- Catastrophe
  - Lab notebooks are in a "safe" place
  - Electronic data are backed up and stored in a separate location
  - Samples are stored properly to avoid contamination
- Confidentiality
  - Information on human subject – see HIPAA guidelines
  - Information on intellectual property
- Period of retention
  - NIH generally requires 3 years after project end
  - Other agencies may require up to 7 years after project end
  - Other unforeseen uses…

Taken from *ORI Introduction to RCR* (http://ori.hhs.gov/education/products/RCRintro/)

# Data Acquisition

Data Ownership/Sharing

# Ownership/Data Sharing

Who owns the data?

- Researchers
- Funders
  - Grants vs. Contracts
- Research Institutions
  - e.g., "for the most part, NIH makes awards to institutions and not individuals" – *NIH Data Sharing Policy and Implementation Guidance*
- Data Sources
  - Subjects
  - Countries

Illustration by David Zinn

Taken from *ORI Introduction to RCR* (http://ori.hhs.gov/education/products/RCRintro/)

A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance

"**Final research data** are recorded factual material commonly accepted in the scientific community as necessary to document, support, and validate research findings."

# A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance

"NIH expects timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset."

# A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance

"For the most part, it is not appropriate for the initial investigator to place limits on the research questions and methods other investigators might pursue with the data."

# A few interesting quotes from the NIH Data Sharing Policy and Implementation Guidance

"It is also not appropriate for the investigator who produced the data to require co-authorship as a condition for sharing the data."

# Case Study

## from *Responsible Conduct of Research*

Drs. K and W are conducting a NIH-funded long-term (25 years), observational study of the health of pesticide applicators.

- Initial health assessment (health history, physical exam, blood and urine tests, DNA sample, and dust samples)
- Yearly health surveys and full health assessment every 4 years

After the first 15 years:

- Published more than a dozen paper from the database
- Require a elaborate data-sharing agreement before releasing the data

Drs K and W's dilemma is that they recently received requests for access to the database from:

- A pesticide company
- A competing research team
- A radical environment group with an anti-pesticide agenda

# Case Study

## from *Responsible Conduct of Research*

Drs. K and W are conducting a NIH-funded long-term (25 years), observational study of the health of pesticide applicators.

- Initial health assessment (health history, physical exam, blood and urine tests, DNA sample, and dust samples)
- Yearly health surveys and full health assessment every 4 years

After the first 15 years:

- Published more than a dozen paper from the database
- Require a elaborate data-sharing agreement before releasing the data

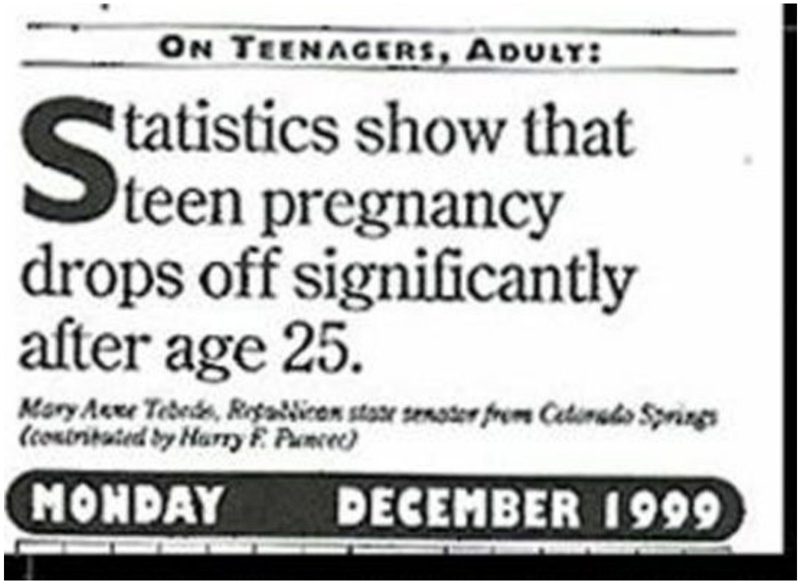
Drs Kessenbaum and Wilcox's dilemma is that they recently received requests for access to the database from:

- A pesticide company
- A competing research team
- A radical environment group with an anti-pesticide agenda

QUESTIONS

- How should Drs. K and W handle these requests to access their database?
- Is it ethical to require people who request data to sign elaborate data sharing agreements

# Statistical Analysis



ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

*Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Puncec)*

MONDAY DECEMBER 1999

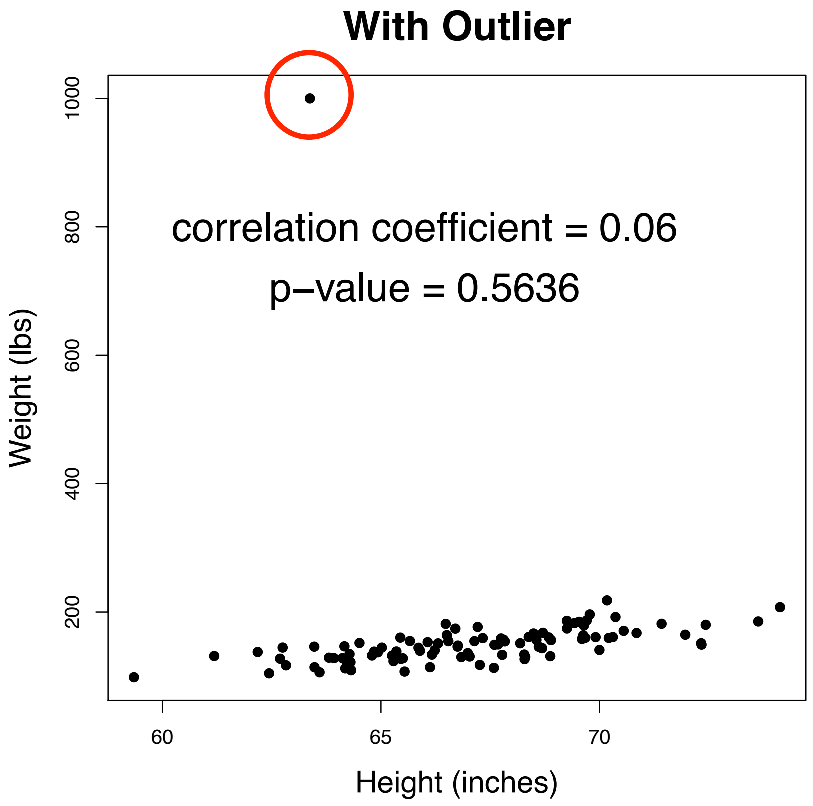
# Tips for Reproducible Statistical Analyses

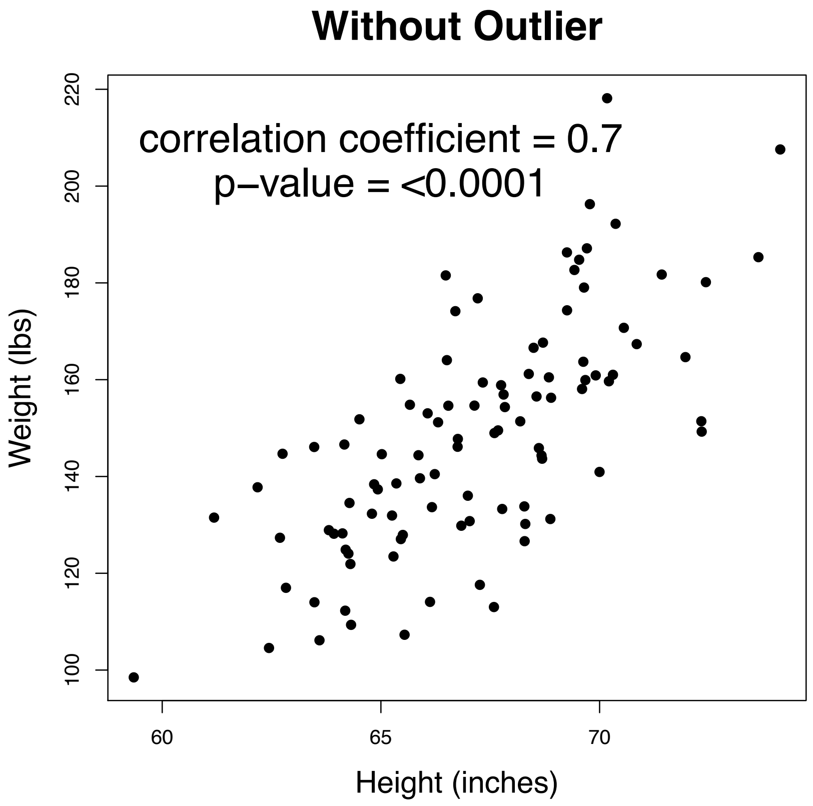
1. ALWAYS keep a version of the “most raw” data
   - Record when and where it was created, so you can easily tell if it has been changed since creation
2. Use a scripting language
   - Programs like R and SAS allow you to follow your steps <u>exactly</u> if you (or someone else) had to redo your analysis
   - Easily execute and document QC steps
   - Avoid copy/paste errors
3. Add comments/notes directly to program
   - Why are you doing this step?
   - What is the goal of this step?
4. Export precise tables/figures from program
   - Avoid transposition errors
   - Save time/energy where changes are requested in initial steps

# Statistical Analysis

Outliers

# Outliers


With Outlier
correlation coefficient = 0.06
p–value = 0.5636
Weight (lbs)
1000
800
600
400
200
60
65
70
Height (inches)


Without Outlier
correlation coefficient = 0.7
p–value = <0.0001
Weight (lbs)
220
200
180
160
140
120
100
60
65
70
Height (inches)

# Outlier Mitigation

1. Identify
   - 2 or 3 standard deviations
   - Unrealistic values
   - Inconsistent
2. Investigate
   - Was there a technical issue?  typo? etc?
   - Is it even a possible true value?
3. Remediate with DOCUMENTATION
   - Make a rule and write it down
4. Sensitivity analysis
   - What would have happened if you hadn't eliminated values?  Is you result robust?

# Case Study

## from *Responsible Conduct of Research*

Anonymous survey of college students on opinion about academic integrity

- 20 questions (Likert scale)
- 10 open-ended questions
- 480 surveys administered (320 responses)

<u>Issues:</u>

1. 8 surveys appear as practical jokes (obscenities, additional numbers added to scale,etc.)
   – Some questions appear usable but some are not
2. 35 respondents appear to be confused about scale
   – They answer “5” when “1” is more logical given their other answers
3. 29 surveys have names on them when respondents were instructed not to do so

# Case Study

## from *Responsible Conduct of Research*

Anonymous survey of college students on opinion about academic integrity

- 20 questions (Likert scale)
- 10 open-ended questions
- 480 surveys administered (320 responses)

<u>Issues:</u>

1. 8 surveys appear as practical jokes
   - Some questions appear usable but some do not
2. 35 respondents appear to be confused about scale
   - They answer "5" when "1" is more logical given their other answers
3. 29 surveys have names on them when respondents were instructed not to do so

QUESTIONS:

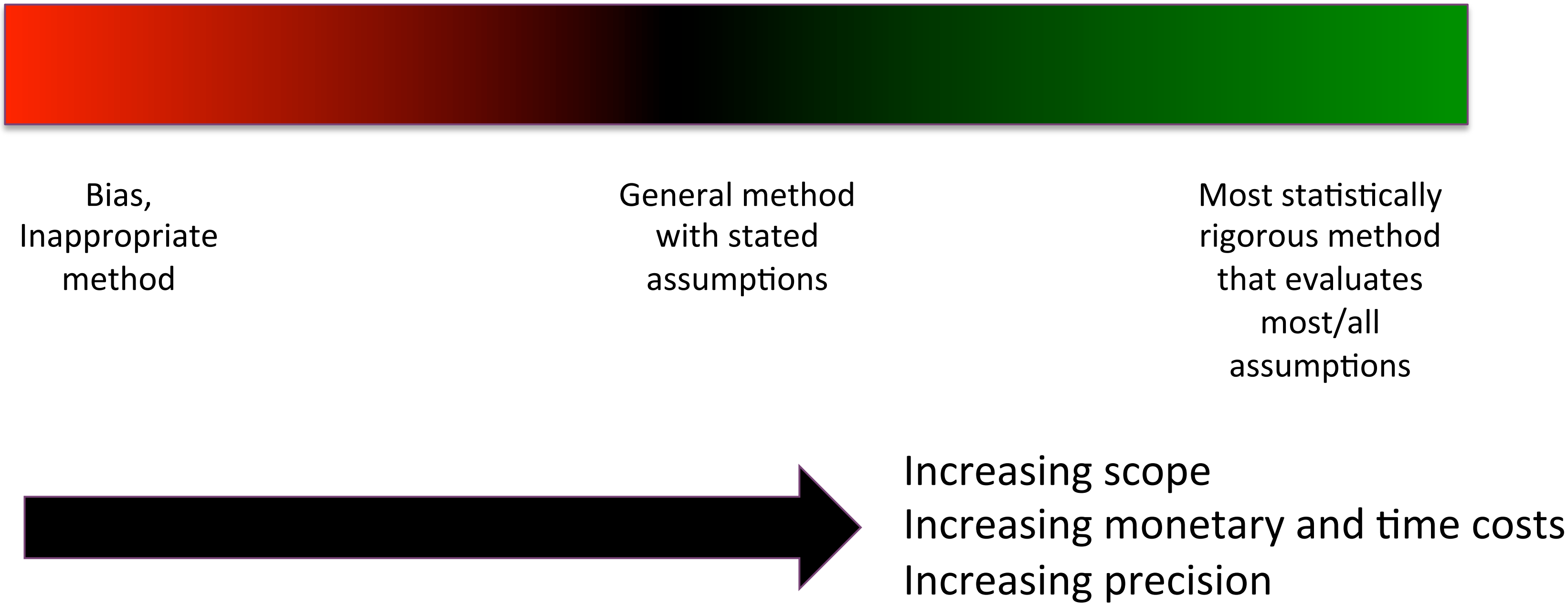
1. How should the researchers deal with theses issues with their data?
2. Should they try to edit/fix surveys that have problems?
3. Should they throw away any surveys? Which ones?
4. How might their decisions concerning the disposition of these surveys affect their overall results?

# Statistical Analysis

Suitable, better, and best methods for analysis

# Methods for Statistical Analysis

- What is the norm in the field?
- A spectrum of alternative statistical methods



Bias,
Inappropriate
method

General method
with stated
assumptions

Most statistically
rigorous method
that evaluates
most/all
assumptions

Increasing scope
Increasing monetary and time costs
Increasing precision

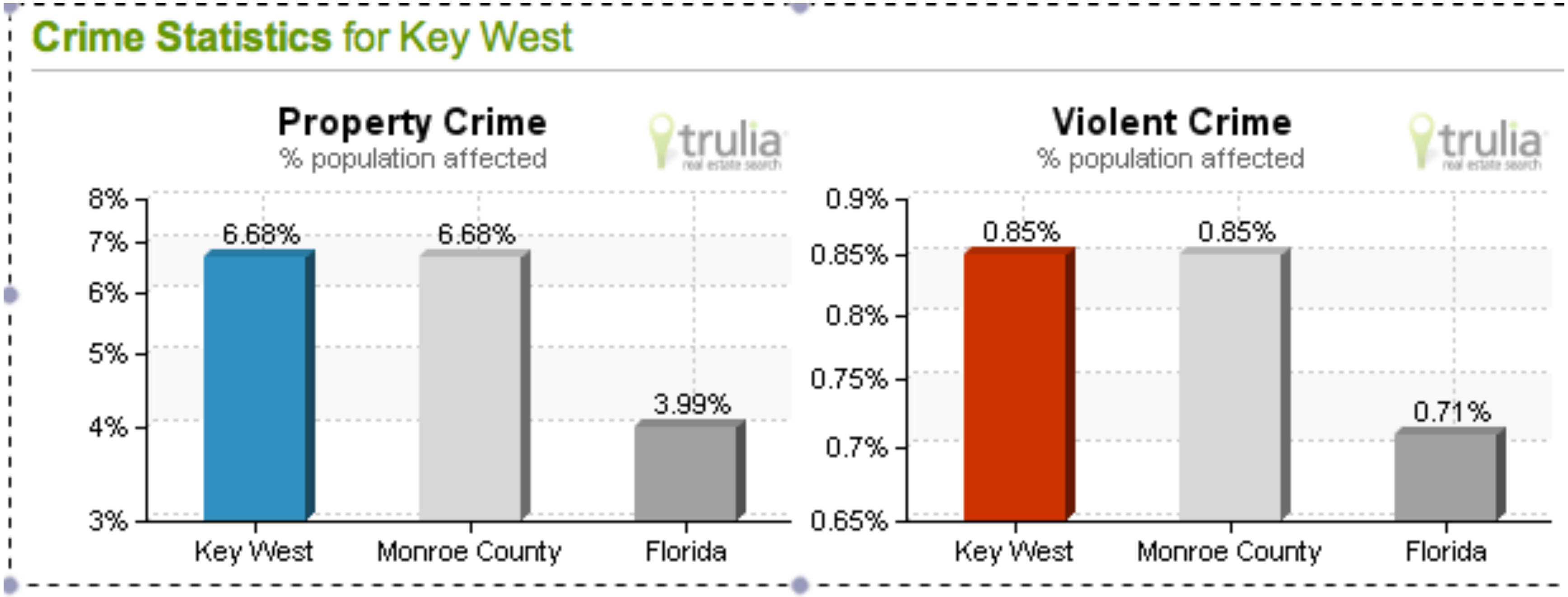
# Know the assumptions of any test

- Is the same subject/sample measured more than once?
- Are the data normally distributed?
- Is there equal variance in each group?
- Are subjects randomly assigned to a treatment? Are they matched?
- Is temporal order assumed?

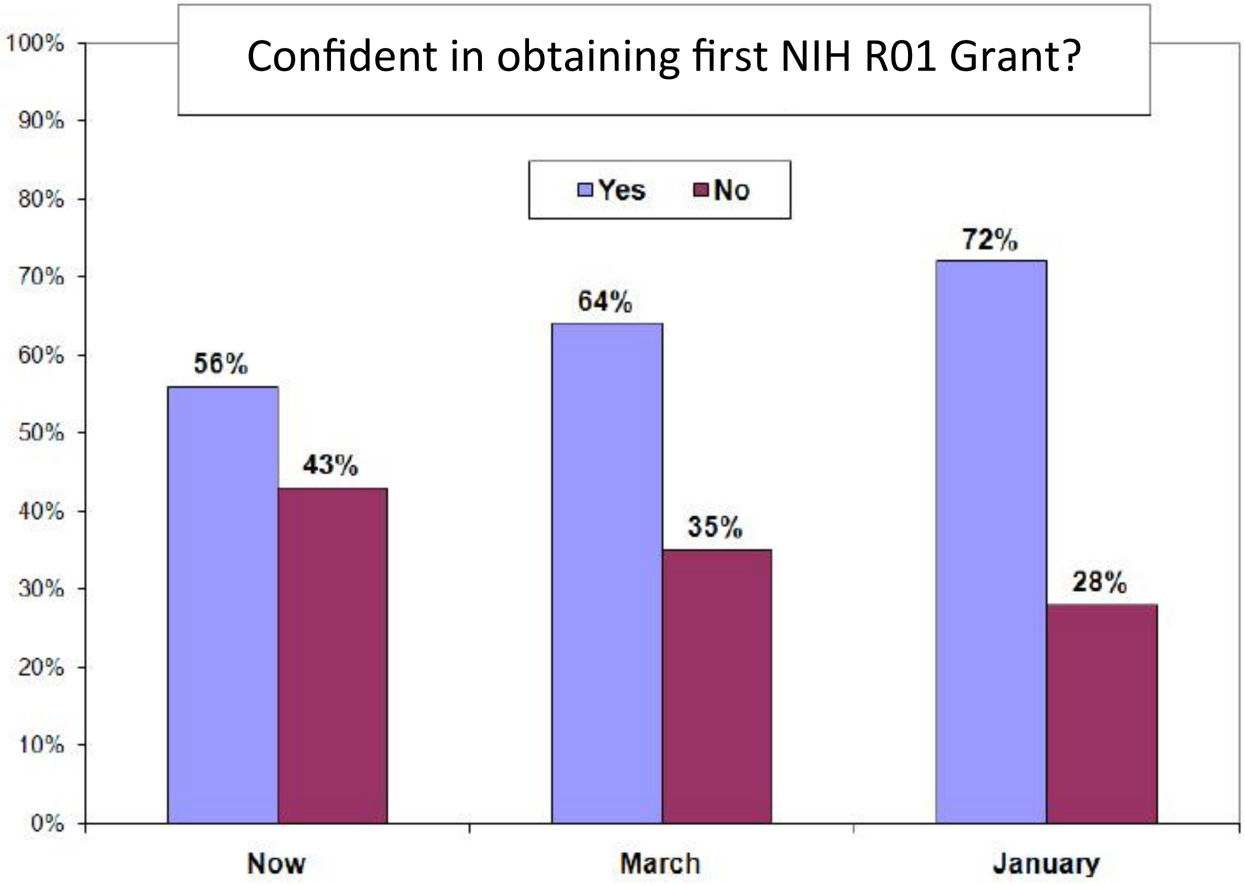
# Statistical Analysis

Display and interpretation of results

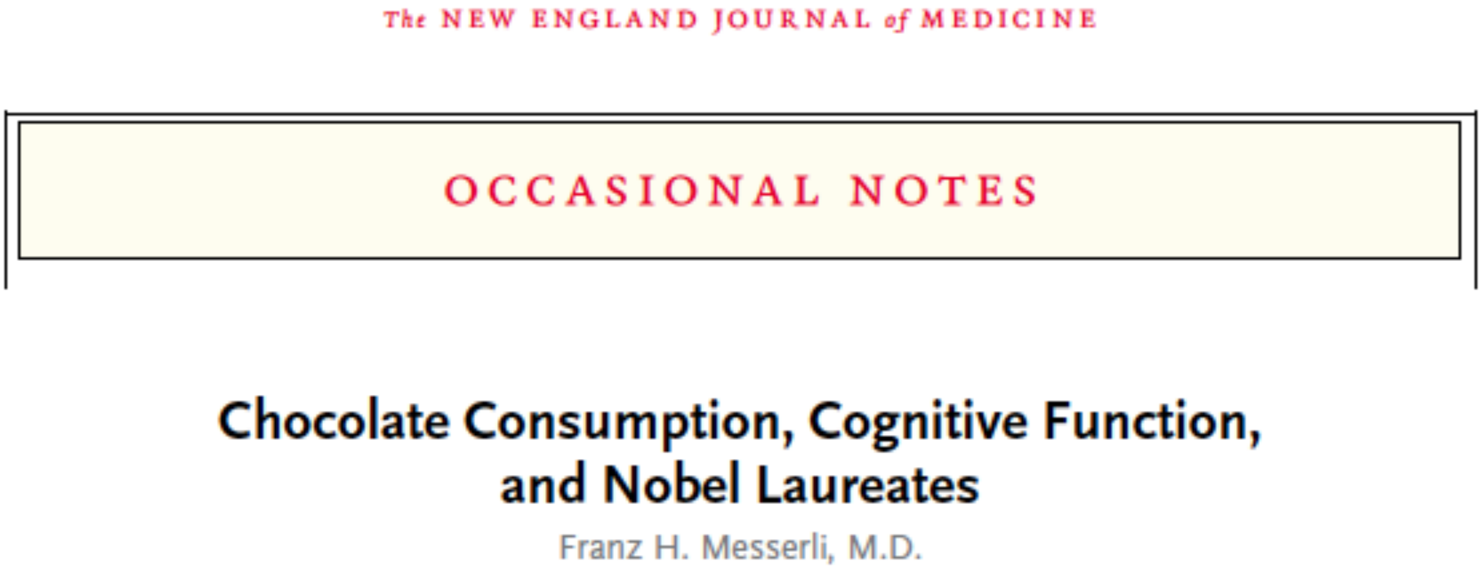
# Displaying Results



# Displaying Results



# Interpreting Results



The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

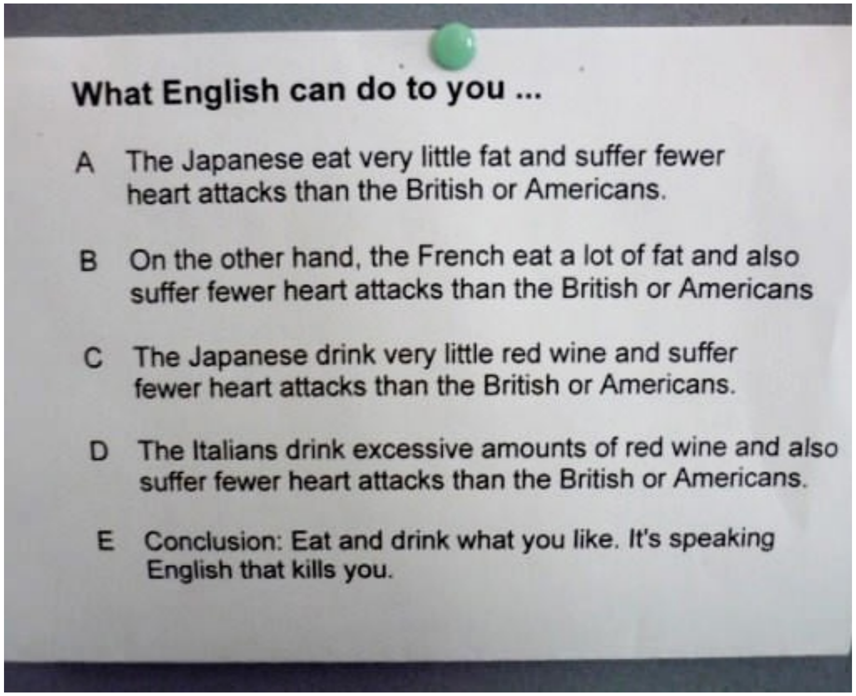
**Chocolate Consumption, Cognitive Function, and Nobel Laureates**

Franz H. Messerli, M.D.

“Switzerland was the top performer in terms of both the number of Nobel laureates and chocolate consumption. The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1. For the United States, that would amount to 125 million kg per year. The minimally effective chocolate dose seems to hover around 2 kg per year, and the dose–response curve reveals no apparent ceiling on the number of Nobel laureates at the highest chocolate-dose level of 11 kg per year.”

# Interpreting Results

- Association vs. Causation
  - Causation can only be proven in a carefully designed and carefully controlled prospective study
    - Eating more chocolate will not cause you to become a Nobel Laureate
- Potential Confounding Issues
  - Confounding variable – "extraneous variable in a statistical model that correlates with both the dependent variable and the independent variable" – *Wikipedia*
  - e.g., Coffee drinkers are more likely to get lung cancer
    - Smokers are more likely to be coffee drinkers and smokers are more likely to get cancer



**What English can do to you ...**

A The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.

B On the other hand, the French eat a lot of fat and also suffer fewer heart attacks than the British or Americans

C The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans.

D The Italians drink excessive amounts of red wine and also suffer fewer heart attacks than the British or Americans.

E Conclusion: Eat and drink what you like. It's speaking English that kills you.

# Highlights of Ethical Guidelines for Reporting Statistical Analysis/ Results in Publications

From American Statistical Association's *Ethical Guidelines for Statistical Practice*

1. Report statistical and substantive assumption made in the study.
2. Account for all data considered in a study and explain the sample(s) actually used
3. Report the sources and assessed adequacy of the data
4. Report the data cleaning and screening procedures used
5. Clearly and fully report the steps taken to guard validity. Address the suitability of the analytic methods and their inherent assumptions relative to the circumstances of the specific study

# Acknowledgements/References

- Dr. Paula Hoffman
- Dr. Brandie Wagner
- ORC and CCTSI

References

*Responsible Conduct of Research* by Adil E Shamoo and David B. Resnick. Second Ed. Oxford University Press, 2009.

*NIH Data Sharing Policy and Implementation Guidance* ( http://grants.nih.gov/grants/ppolicy/data_sharing/data_sharing_guidance.htm), March 5, 2003.

*Ethical Guidelines for Statistical Practic*e, American Statistical Association ( http://www.amstat.org/about/ethicalguidlines.cfm), August 7, 1999.

*Introduction to RCR – 6. Data Management Practices*, Office of Research Integrity, US Department of Health and Human Services (http://ori.hhs.gov/education/products/RCRintro/c06/0c6.html), September 2006