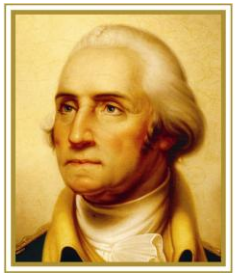
Introduction to Critical Statistical Thinking

Mary A. Foulkes, Ph.D.
George Washington University
April 29, 2012

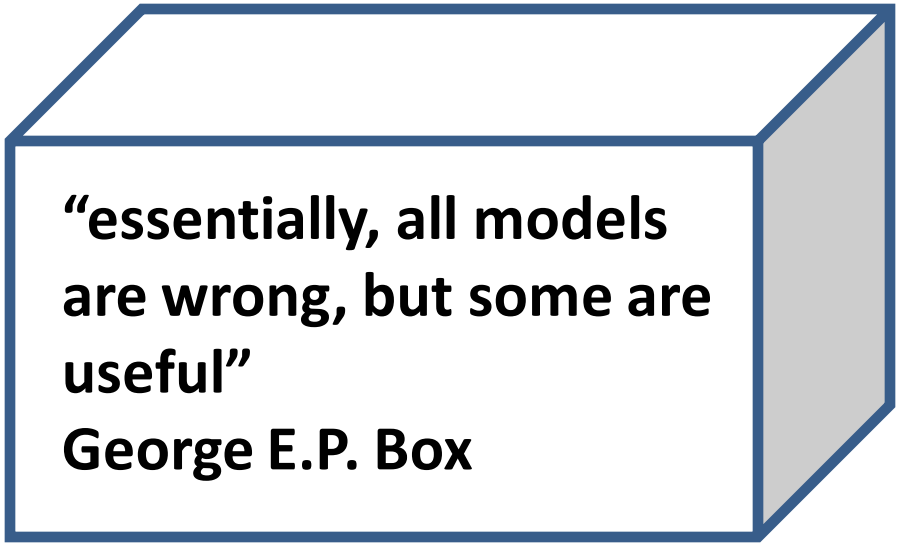
ASA
AMERICAN STATISTICAL ASSOCIATION



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

The Singular & Plural

- Statistics is the discipline of drawing conclusions from data
- Statistics is the science of uncertainty
- Statistics can provide a basis of policy-making




Chicago Daily Tribune
DEWEY DEFEATS TRUM

Failure vs. Promise
Opposite Conclusions

The New York Times

February 24, 2003

Large Trial Finds AIDS Vaccine Fails to Stop Infection

By ANDREW POLLACK with LAWRENCE K. ALTMAN

BRISBANE, Calif., Feb. 23 — The first AIDS vaccine ever to be tested in a large number of people has failed, over all, to protect them from infection with the virus that causes the disease, the company that makes it, VaxGen, said today.

The vaccine did, however, seem to significantly lower the infection rate among African-Americans and other non-Hispanic minorities participating in the trial, the company saic

Its researchers called this finding totally unexpected and said they were at a loss to explain why there would be ethnic differences in response to the vaccine. They concede that the findings, though statistically significant, might change if the vaccine were tested among more members of minorities, who were only a small fraction of the people in the trial.

February 24, 2003, *New York Times*

February 24, 2003, *USA Today*

SIAM/ASA Journal on

UNCERTAINTY QUANTIFICATION

Presenting significant mathematical, statistical, algorithmic, and application advances in uncertainty quantification and related fields

Coming in 2013

Publishes research articles presenting significant mathematical, statistical, algorithmic, and application advances in uncertainty quantification and related fields such as sensitivity analysis, model validation, model calibration, data assimilation, and code verification. The journal also solicits papers describing new ideas that could lead to significant progress in methodology for uncertainty quantification as well as review articles on particular aspects. The journal is dedicated to nurturing synergistic interactions between the mathematical, statistical, computational, and applications communities involved in uncertainty quantification and related areas.

Senate Subcommittee on Public Health – Feb 2002

- Don Berry, U Texas
- value of screening mammography
  - Lead-time bias, length bias
  - Unblinded assessment
  - Relative risk vs absolute risk

Berry's Recommendations

1. **Provide women with decision aids in which they are informed of the benefits and risks, including uncertainties, and helped to weigh them in making a decision.**
2. **Audit of the Swedish trials. If an audit of these trials examines the biases and confirms the recently announced 21% reduction in breast cancer mortality then I will agree that screening has a benefit.**
3. **Research progress will help us better understand the relationships between biological markers, early detection and treatment. Especially exciting are the genomics and bioinformatics revolutions.**

Not all evidence is equal

Hierarchy of Evidence

Anecdotal case reports
Case series without controls
Series with literature controls
Analyses using computer databases
"Case-control" observational studies
Series based on historical control groups
Single randomized controlled clinical trial
Confirmed randomized controlled clinical trial

Green and Byar, *Stat in Med* 1984

Definitions

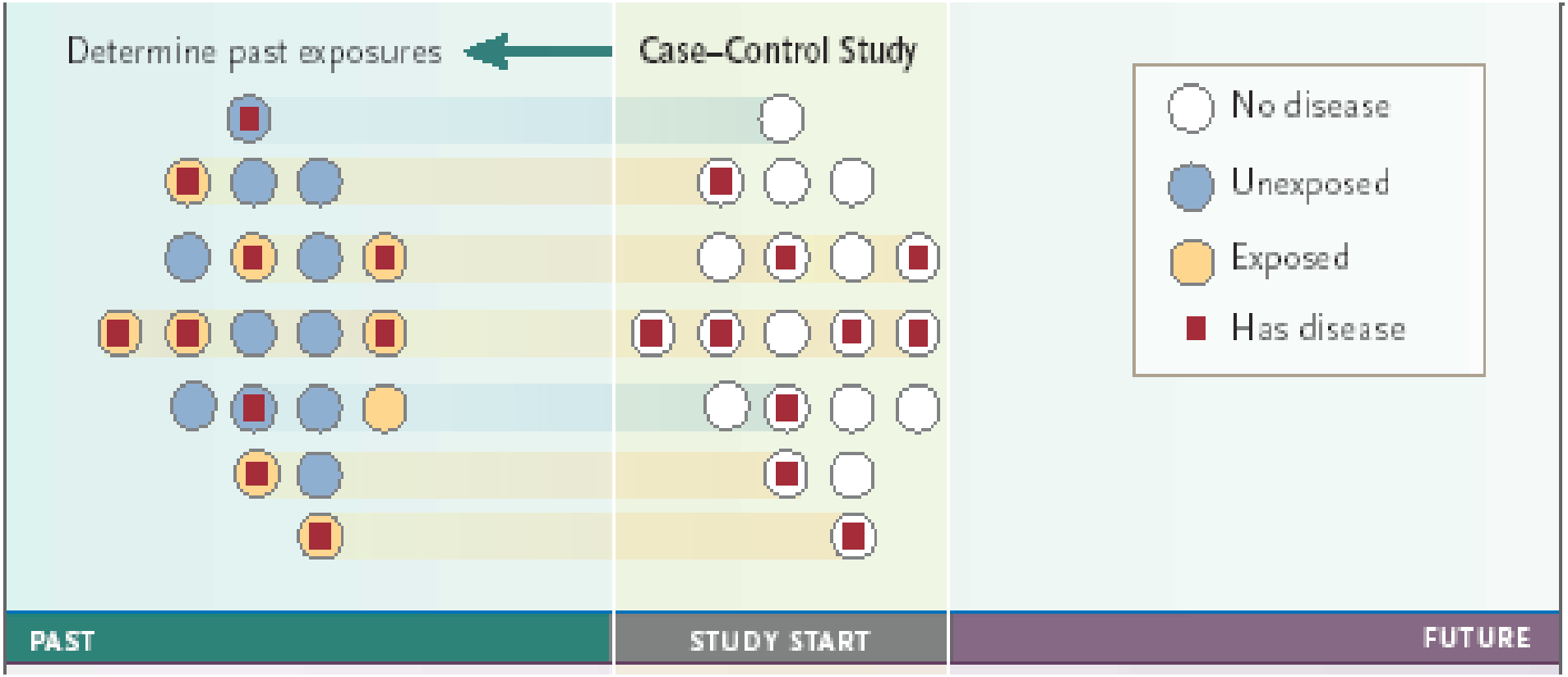
* Case-Control Study: A study that compares individuals affected by a disease (the cases) with individuals who do not have that disease (the controls) to seek possible causes or associations

* Historical Controls: Comparison of results in an experimental group with results in former, often old, reports or records

* Randomized Clinical Trial: A clinical trial that involves the formation of treatment groups by the process of random allocation

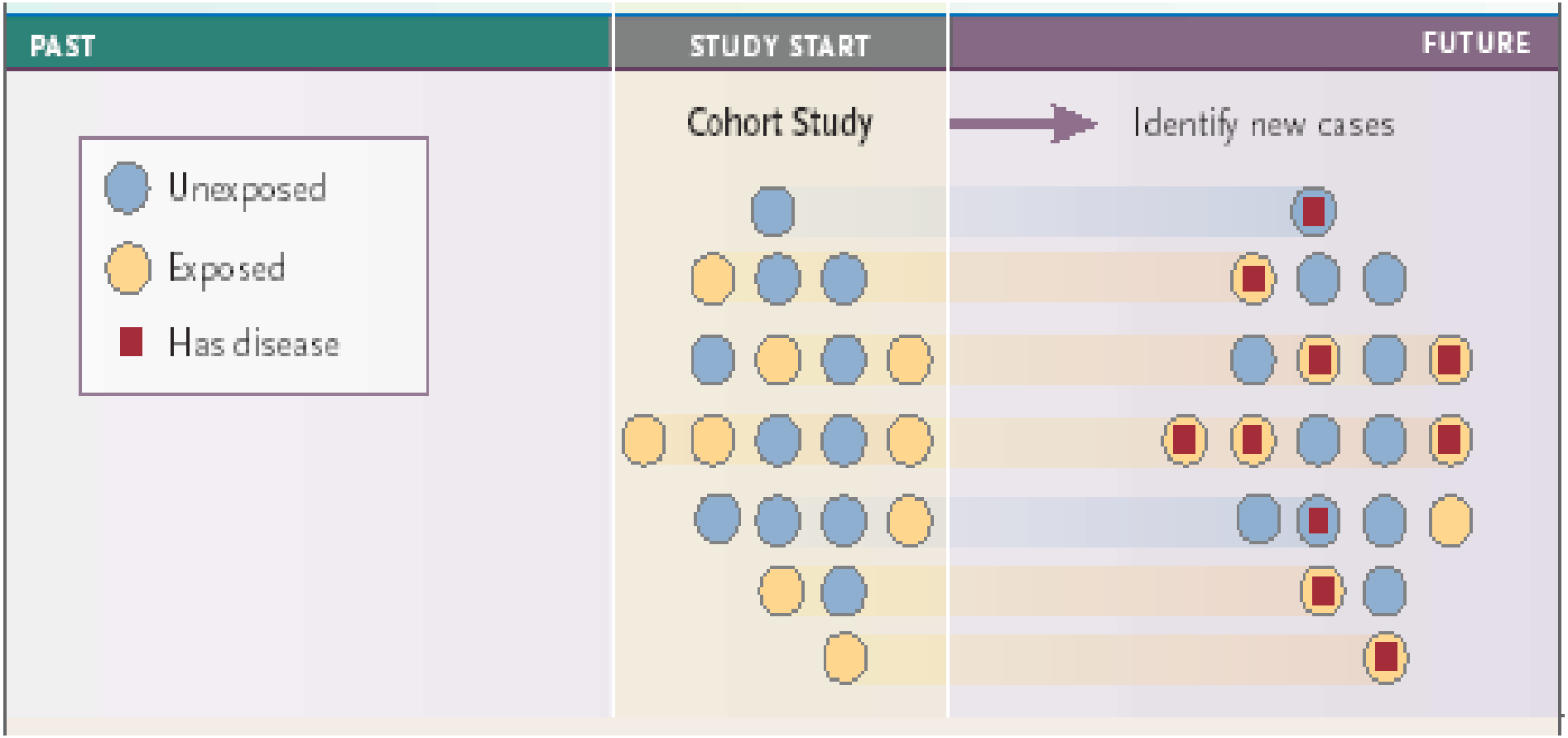
Cohn and Cope, *News & Numbers*, 2001
Cambridge Dictionary of Statistics, 1999

Case Control Study



Manolio, *NEJM* 2003

Cohort Study



Manolio, *NEJM* 2003

Rules of Thumb

- If too good to be true, probably isn't
- If biologically implausible, probably is
- Possibility ≠ Probability
- In vitro ≠ In vivo (gasoline kills HIV)
- Planned > Post Hoc
- More is more (beware of n = 1)
- Consider the source
- Garbage in, garbage out

Failure vs. Promise
Opposite Conclusions

The New York Times

February 24, 2003

Large Trial Finds AIDS Vaccine Fails to Stop Infection

By ANDREW POLLACK with LAWRENCE K. ALTMAN

BRISBANE, Calif., Feb. 23 — The first AIDS vaccine ever to be tested in a large number of people has failed, over all, to protect them from infection with the virus that causes the disease, the company that makes it, VaxGen, said today.

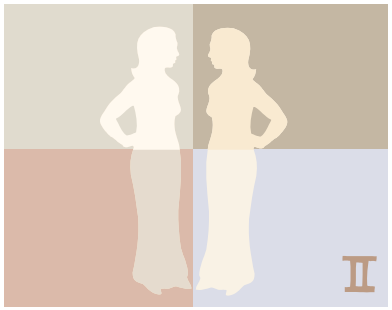
The vaccine did, however, seem to significantly lower the infection rate among African-Americans and other non-Hispanic minorities participating in the trial, the company said.

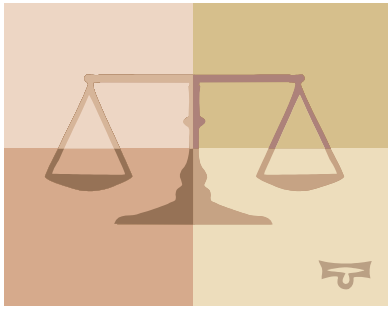
Its researchers called this finding totally unexpected and said they were at a loss to explain why there would be ethnic differences in response to the vaccine. They conceded that the findings, though statistically significant, might change if the vaccine were tested among more members of minorities, who were only a small fraction of the people in the trial.

February 24, 2003, *New York Times*

February 24, 2003, *USA Today*

Subgroup Analysis Problem

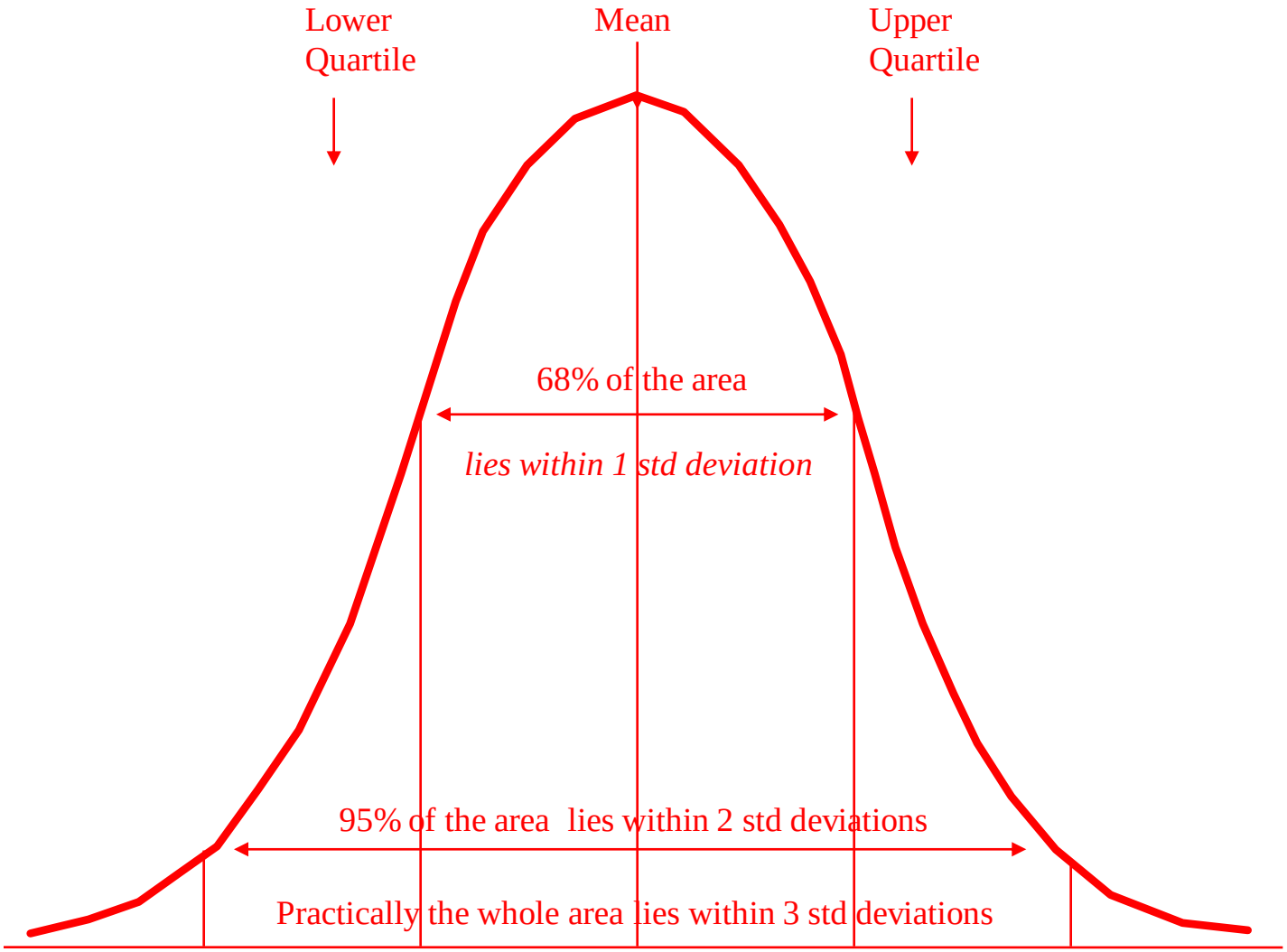




ISIS-2 (Lancet 1988):

Even when the overall effect is positive, subgroups can be identified in which the intervention is particularly ineffective: Gemini & Libra have a 9% increase in odds of death

The Normal Distribution



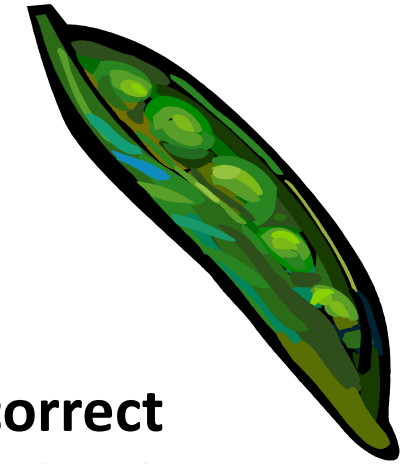
Statistical significance

- Researchers use probability ("p value*") to gauge significance of results
- Experiments compare status quo ("straw man") to new ideas

New York Times

- A finding is significant if new idea appears better, with low probability of result being due to chance
- "Significance" typically refers to $p < .05$, or more stringently, $p < .01$

p-values



A p-value is a measure of probability

- **It builds on the idea that the status quo is correct**
- **Given the status quo, how likely are the results that we see in this experiment?**

Some p-value definitions inherently are confusing:

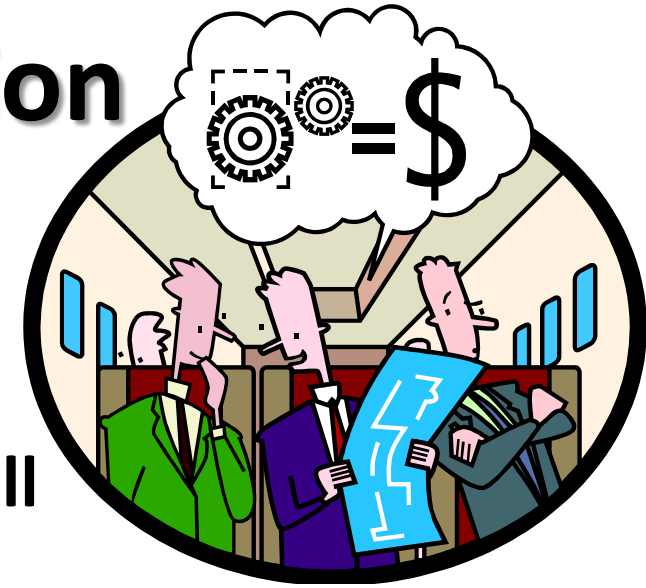
***p-value: The probability that an observed relationship or effect or result could have *seemed* to occur by chance *if there had actually been no real effect*.**

Cohn and Cope*. News & Numbers*, 2001

***p-value: The probability of the observed data (or data showing a more extreme departure from the null hypothesis) when the null hypothesis is true.**

***Cambridge Dictionary of Statistics*, 1999**

Lost in Translation



p-value often misinterpreted

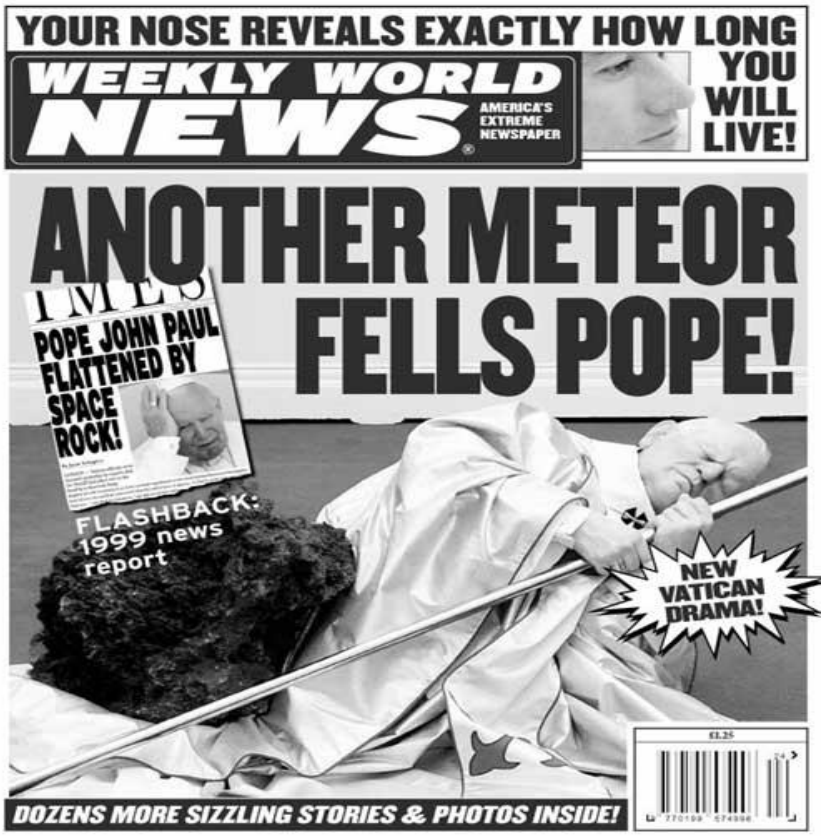
- **It is *not* the probability that the null hypothesis is true**
- **It is *not* the probability that the results arose by chance**

- **It is the probability that something could have happened under the status quo**
- **It is the probability that chance could have led to the results**

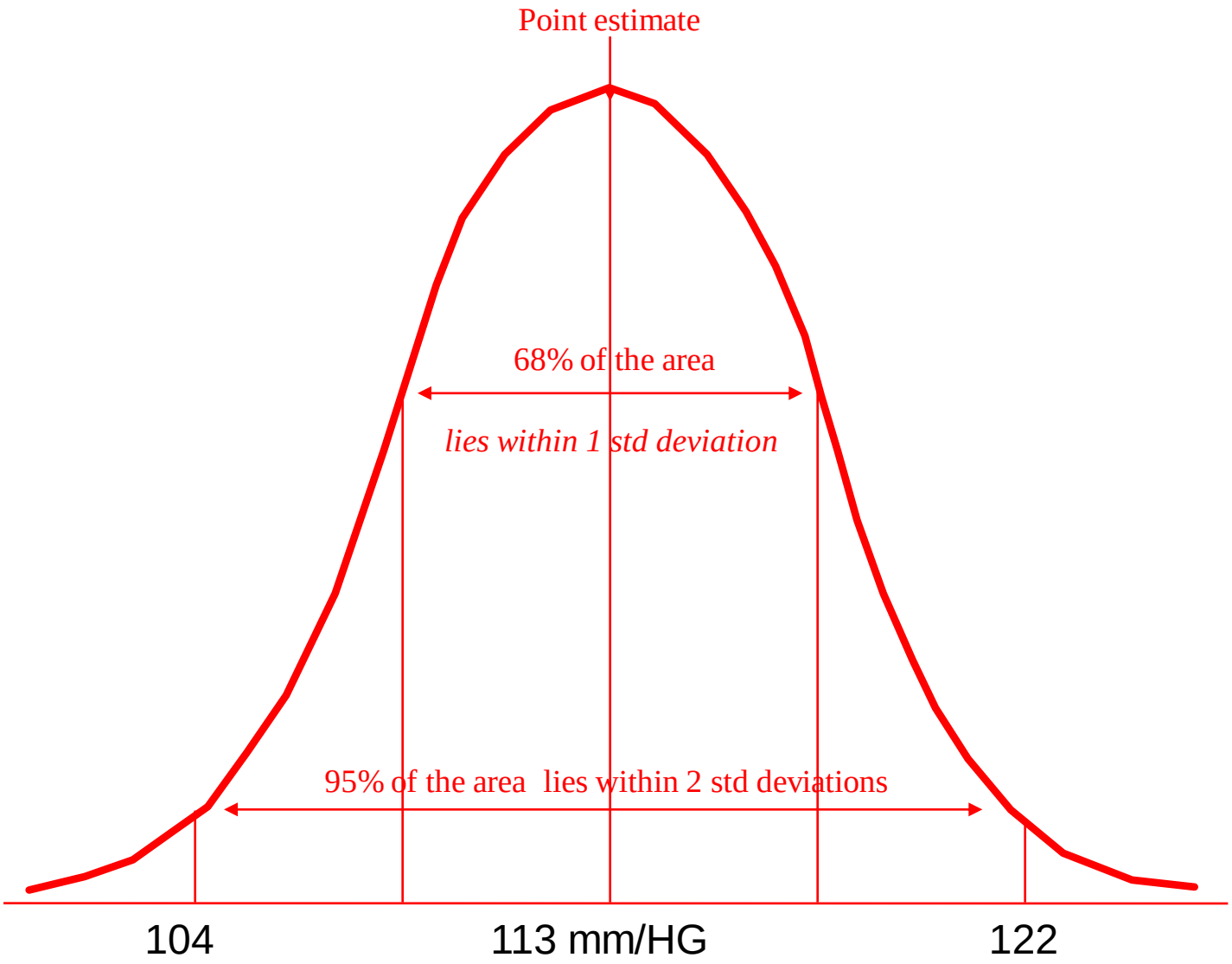
Two Types of Errors

*Type I: Disbelieving the truth

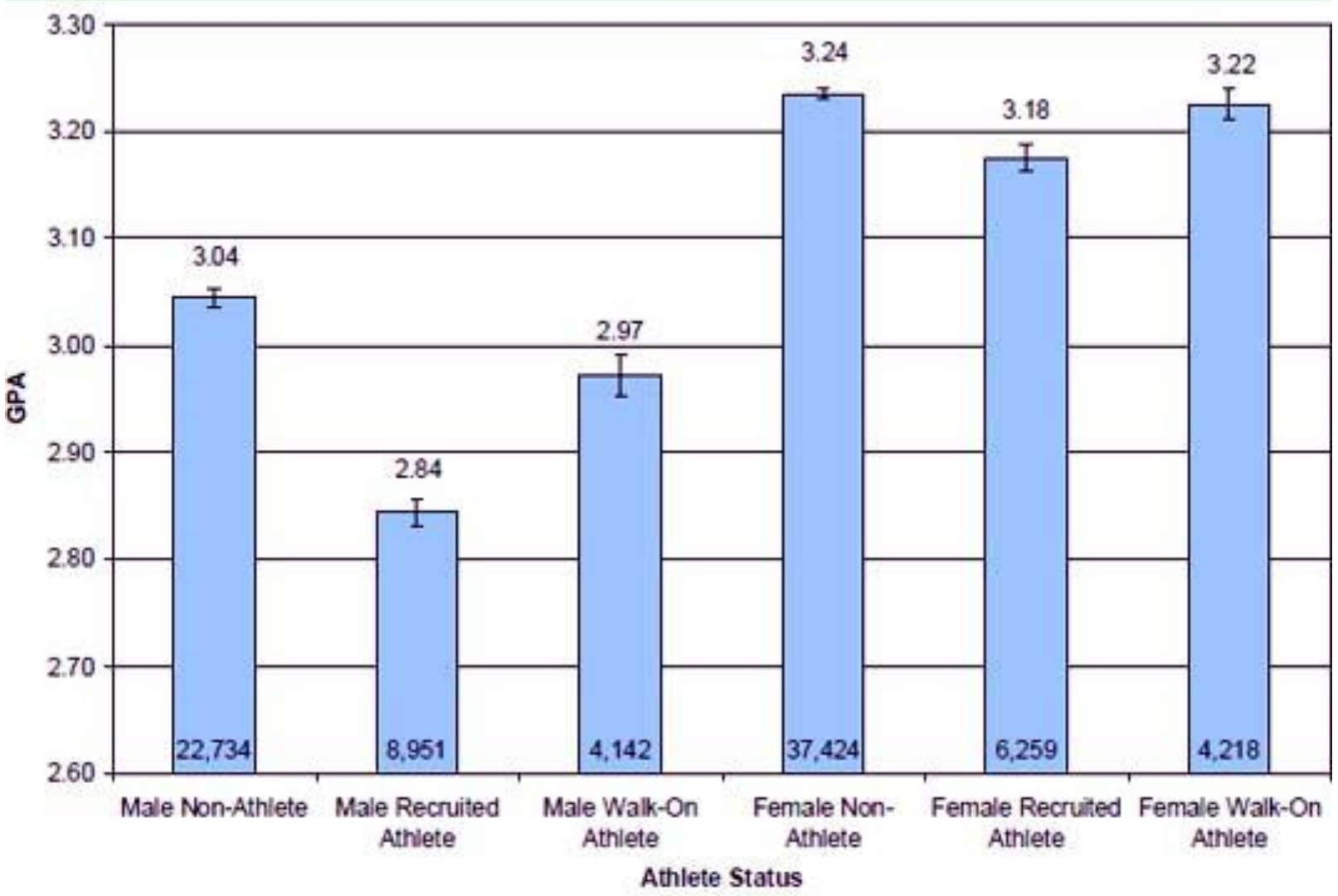
*Type II: Believing the untruth



The Confidence Interval



GPA by Athlete Status and Gender



College Sports Project

Error bars show 95% confidence interval. Numbers at base show counts.