Wake Forest®
School of Medicine

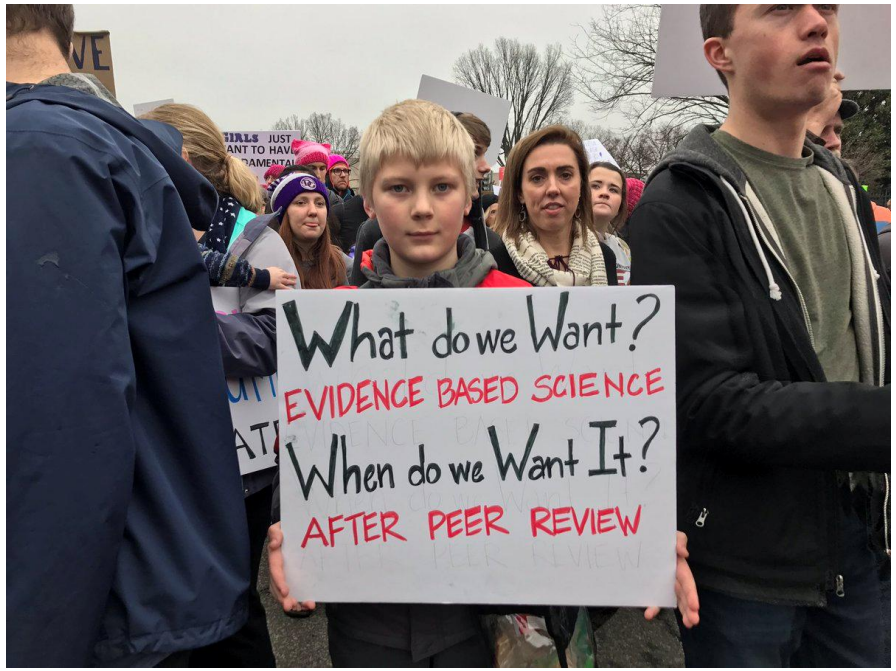
Rigor and Reproducibility

Ellen Quillen, PhD

Internal Medicine – Molecular Medicine

Science in Crisis

- We assume peer-review ensures accurate science
- Large scale attempts to replicate published results mostly fail
  - <1/3 of 100 psychology studies could be replicated
  - 6/53 major studies in oncology
  - 2/18 microarray gene expression studies



@LDMay, Twitter

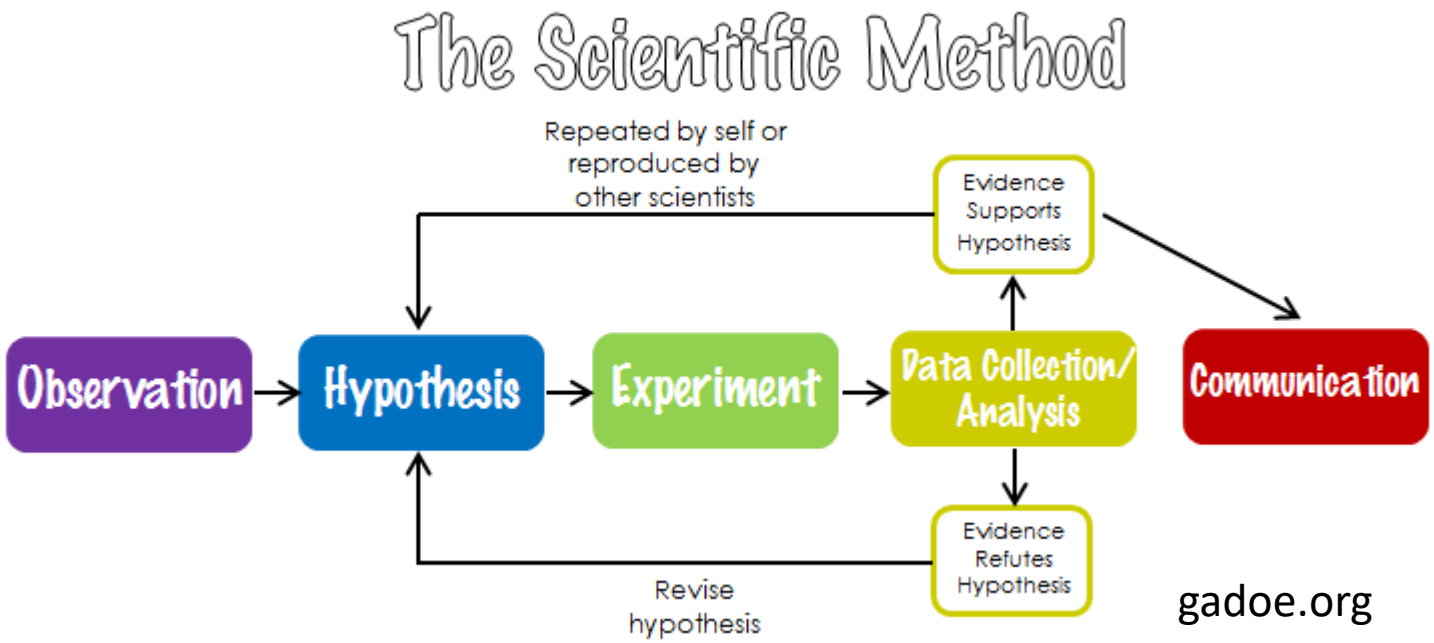
Open Science Collaboration 2015; Begley et al 2012 *Nature*; Ioannidis et al *Nat Gen* 2009

What is ultimately at stake if research data cannot be trusted?

Nobody has responded yet.

Hang tight! Responses are coming in.

Is Science Self-Correcting?



gadoe.org

- Over decades, probably
- In the short term…only if we focus on improving our rigor and reproducibility

How does NIH Define R&R?

- The application of **rigor** ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results.
- When a result can be **reproduced** by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Failures of Rigor and Reproducibility…

- Waste resources
- Delay cures
- Harm patients
- Lead down spurious paths
- Reduces public confidence

Theoretical Levels of Rigor

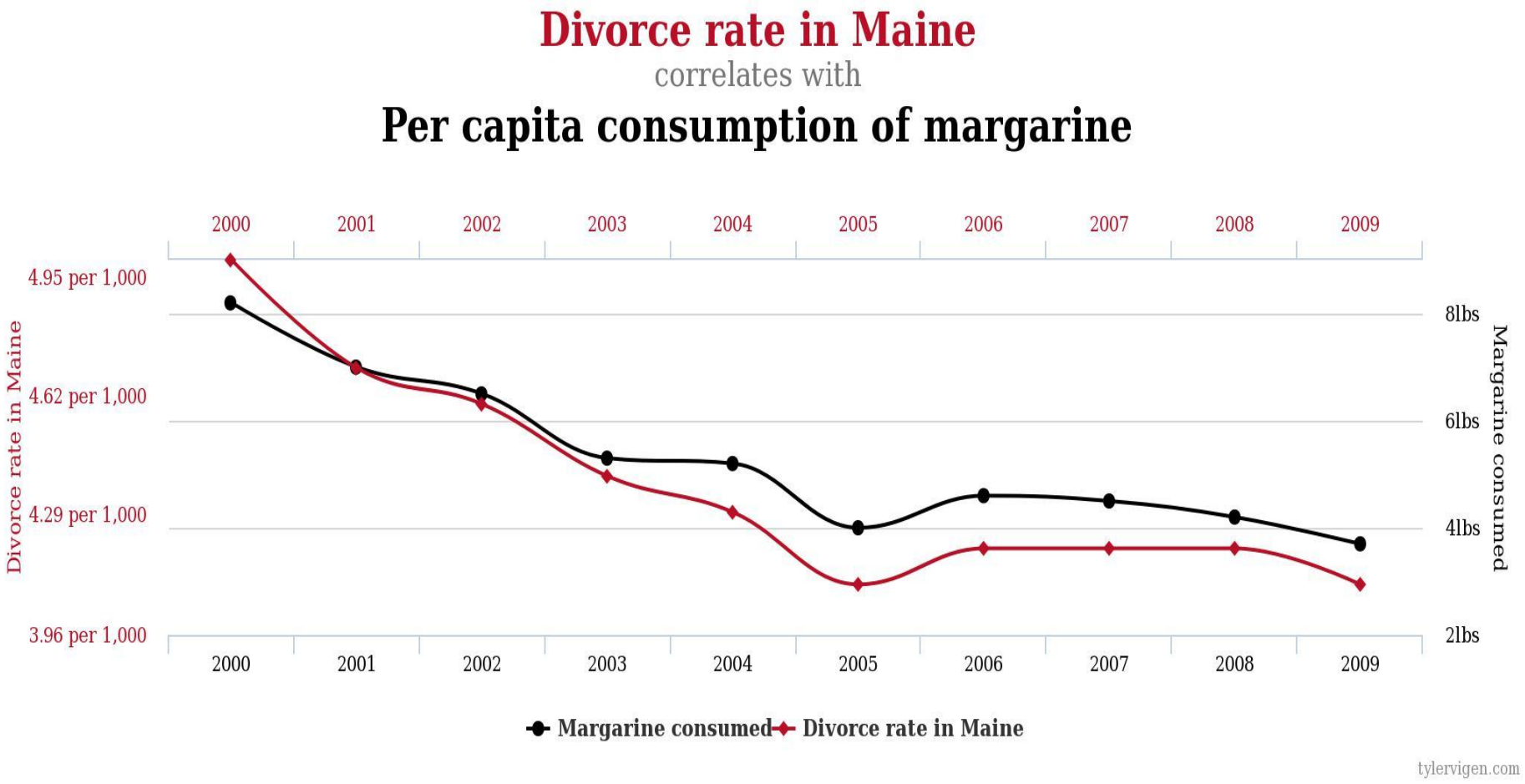
Name	Description	Outcome
Insidious Rigor	Scientist purposely engages in falsifying data from initial grant review to publication	Misleading Misconduct Possibly Criminal
Creative Rigor	Scientist deliberately targets or avoids targets where rigor needs to be applied; shows best results to support hypothesis; cherry-picks data	Misleading Low Chance of reproducibility
Careless Rigor	Scientist randomly applies rigor only when necessary or if asked to (e.g. verify cell lines)	Modest chance of reproducibility
Selective Rigor	Scientist applies rigor where their experience dictates it necessary	Good chance of reproducibility
Careful Rigor	Scientist carefully applies rigor to each aspect of the study	High chance of reproducibility
Enduring Rigor	Results are independently repeated	Reproducible

Innate Bias Influences Science

- Self-deception is surprisingly easy
- Almost no scientists are setting out to deceive
- Confirmation bias is strong
- Publish or perish is a real fear

Nuzzo 2015 *Nature News*

False Correlation is rampant



Rigor: Good Statistical Analysis

- Study Design: correct collection of data, pre-plan sample size, stopping points
- Post-experimental data analysis: pool data properly; address missing data points; exclude data properly
- Statistical design: proper statistical tests; distinguish between hypothesis-driven experiments and other types

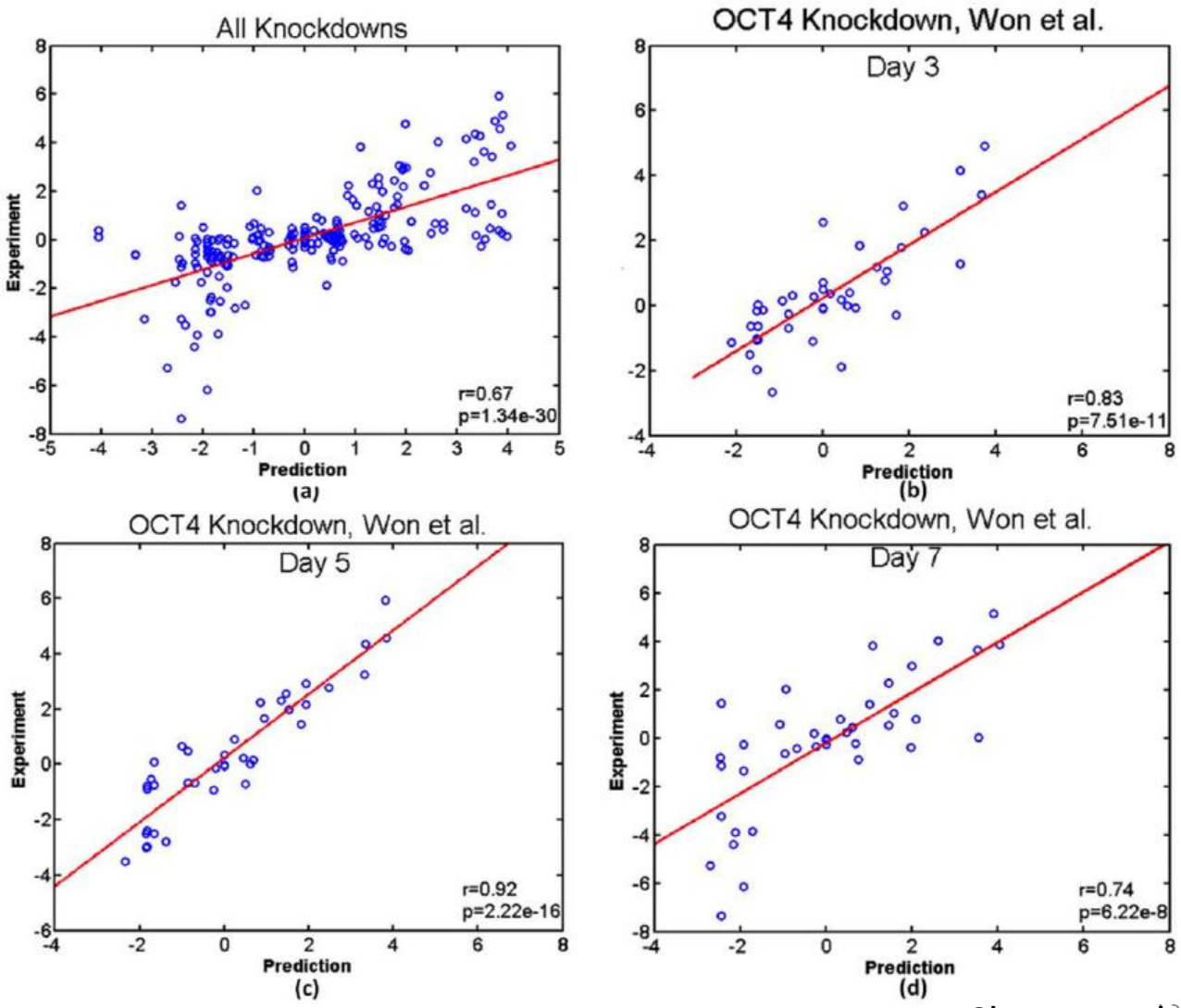
Rigor: Good Statistical Analysis

- Statistical methods should be planned before data collection begins
- Select the right statistical tests for type of data and study design
  - Categorical/quantitative; Case/control or observational; Number of groups; Type of covariates
- Understand the limitations and assumptions of the tests you are performing
  - Sample size, Distribution of Data, Mean vs. Median, Ordinal or Continuous Data

Common Errors in Statistical Analysis

What is a *p*-value?

- Test statistics describe the magnitude of the results
  - Difference in mean
  - Correlation
- *p*-values describe the strength of the evidence



Chang et al. 2011

Which of the following is a correct interpretation of a p-value or 0.05?

The chance that the studied (alternative) hypothesis is true is 95%.

0%

The effect of our independent variable on our dependent variable is strong.

0%

Assuming the null hypothesis is true, we would have gotten the same result 5% of the time.

0%

There is a 5% chance that our results are wrong.

0%

Which of the following is a correct interpretation of a p-value or 0.05?

The chance that the studied (alternative) hypothesis is true is 95%.

0%

The effect of our independent variable on our dependent variable is strong.

0%

Assuming the null hypothesis is true, we would have gotten the same result 5% of the time.

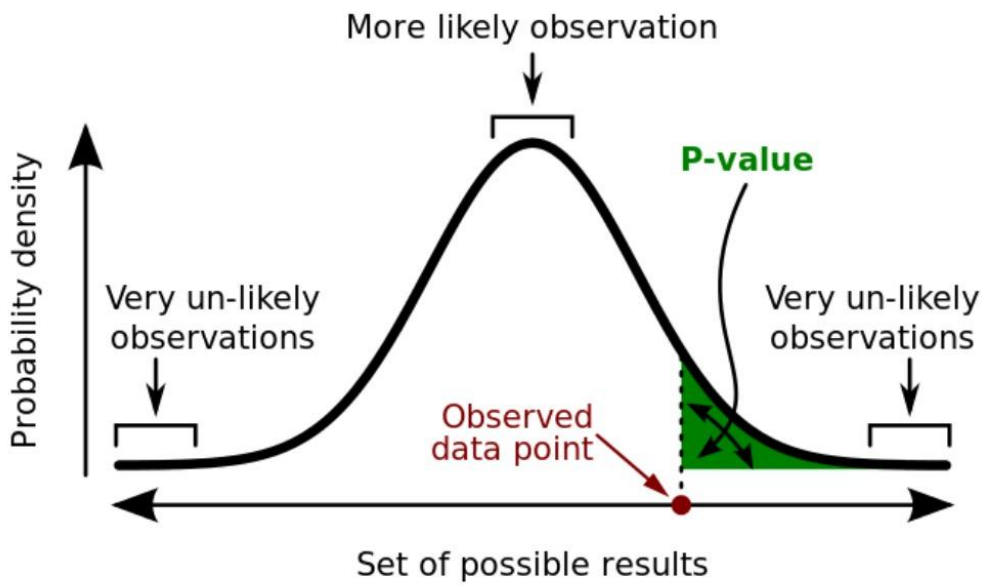
0%

There is a 5% chance that our results are wrong.

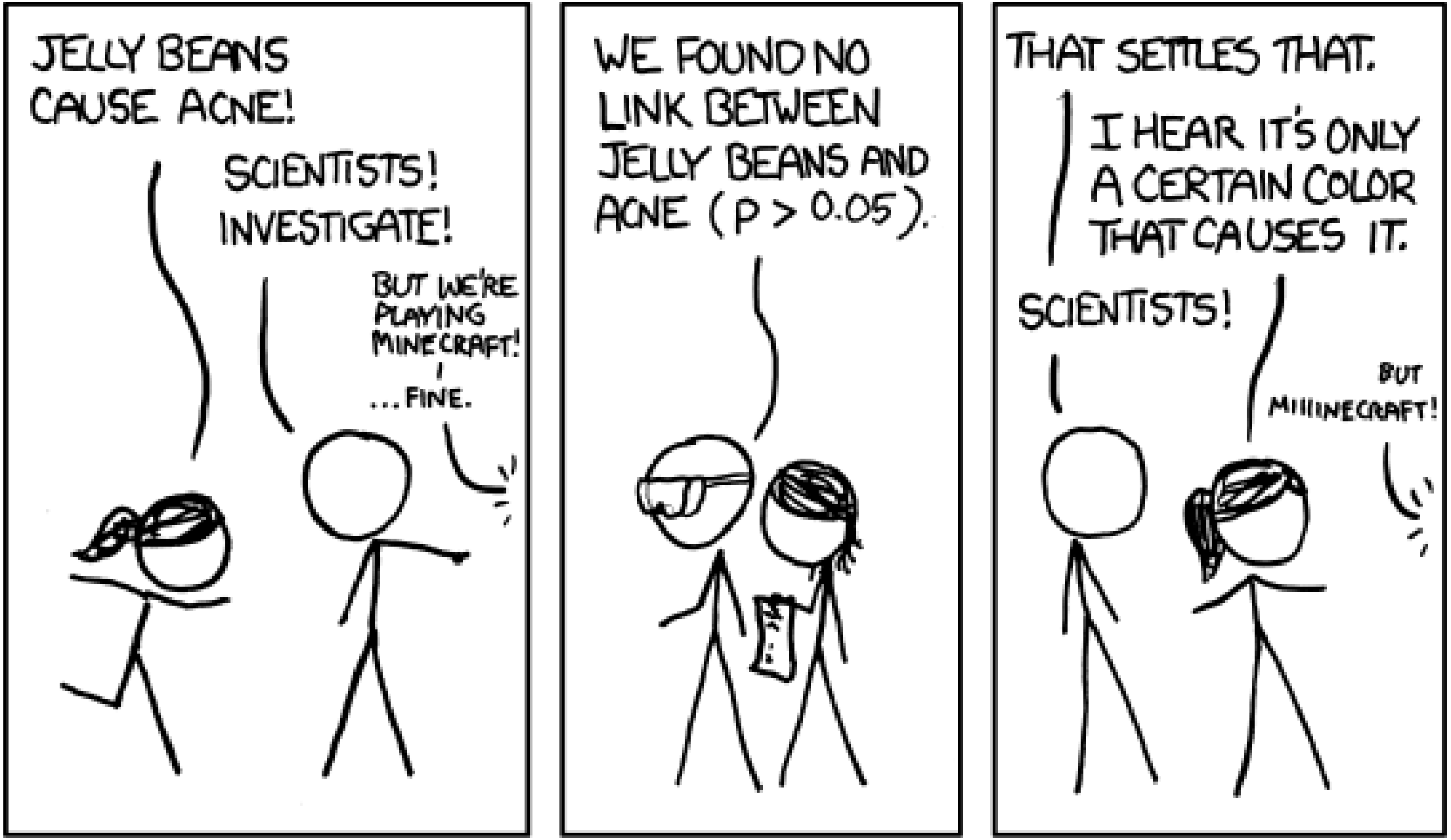
0%

Significance Testing

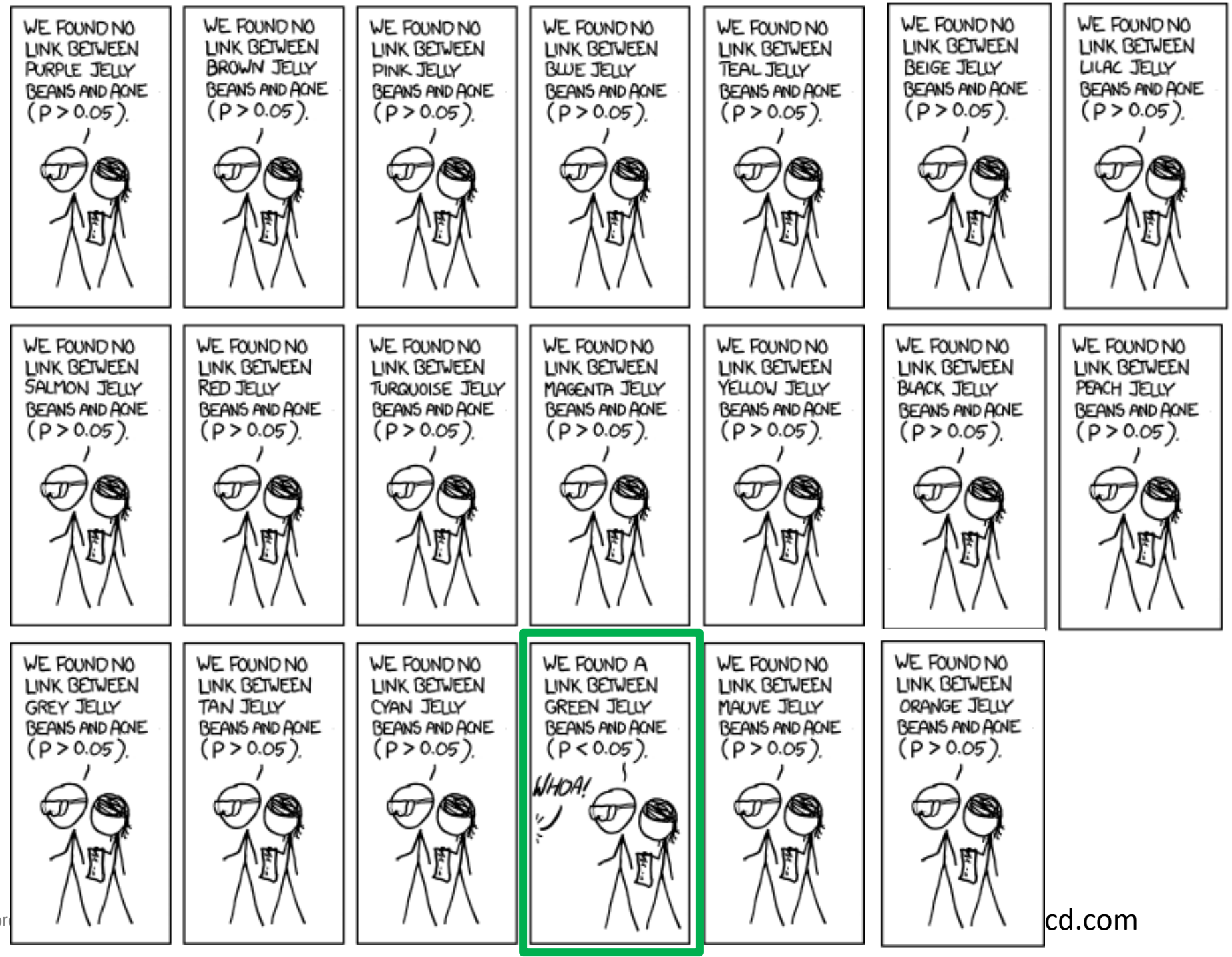
- If p = 0.05, there is a 5% chance of getting the same result by chance
- 95% chance random sample gives non-significant result

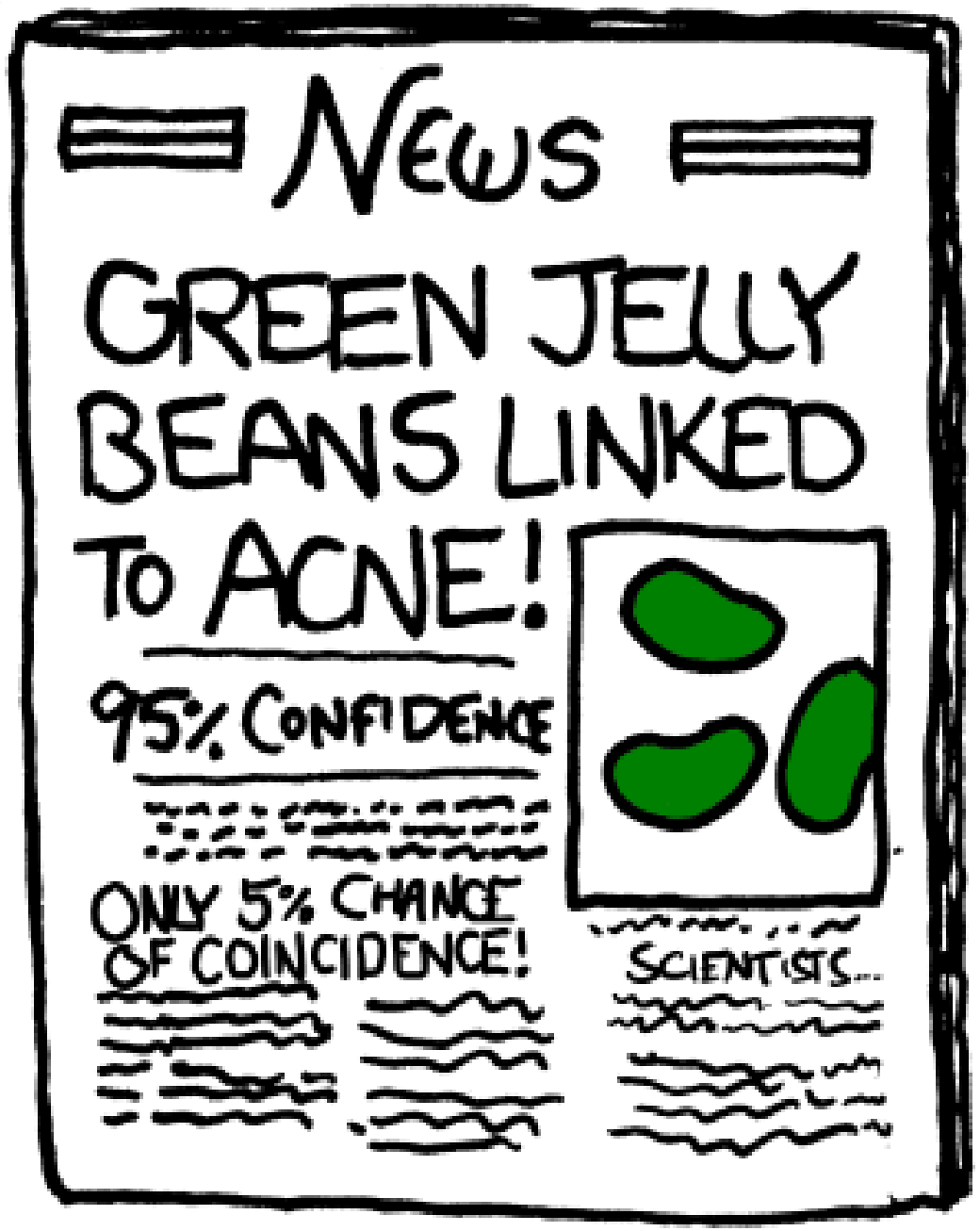


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

xkcd.com


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com

Assuming no correlation between any color of jellybean and acne (Ho)

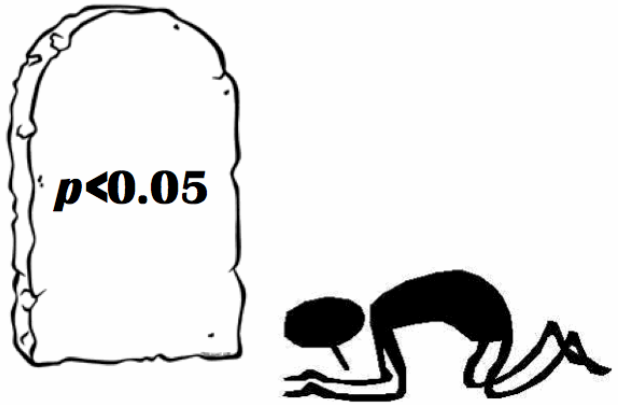
- Chance of (correctly) drawing from “non-significant” part of the graph 20 times is:
  - 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95 x 0.95
  - = (0.95)^20 = 0.35
- Chance of randomly getting at least one $p < 0.05$ = 0.65

Correction for Multiple Testing

- Divide your α by the number of tests you are doing (Bonferroni Correction):
  - α = 0.05/20 = 0.0025
  - Chance of (correctly) drawing from “non-significant” part of the graph 20 times is:
    - 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975 x 0.9975
  - = (0.9975)^20 = 0.951
- (There are other ways to correct for this like False Discovery Rate)

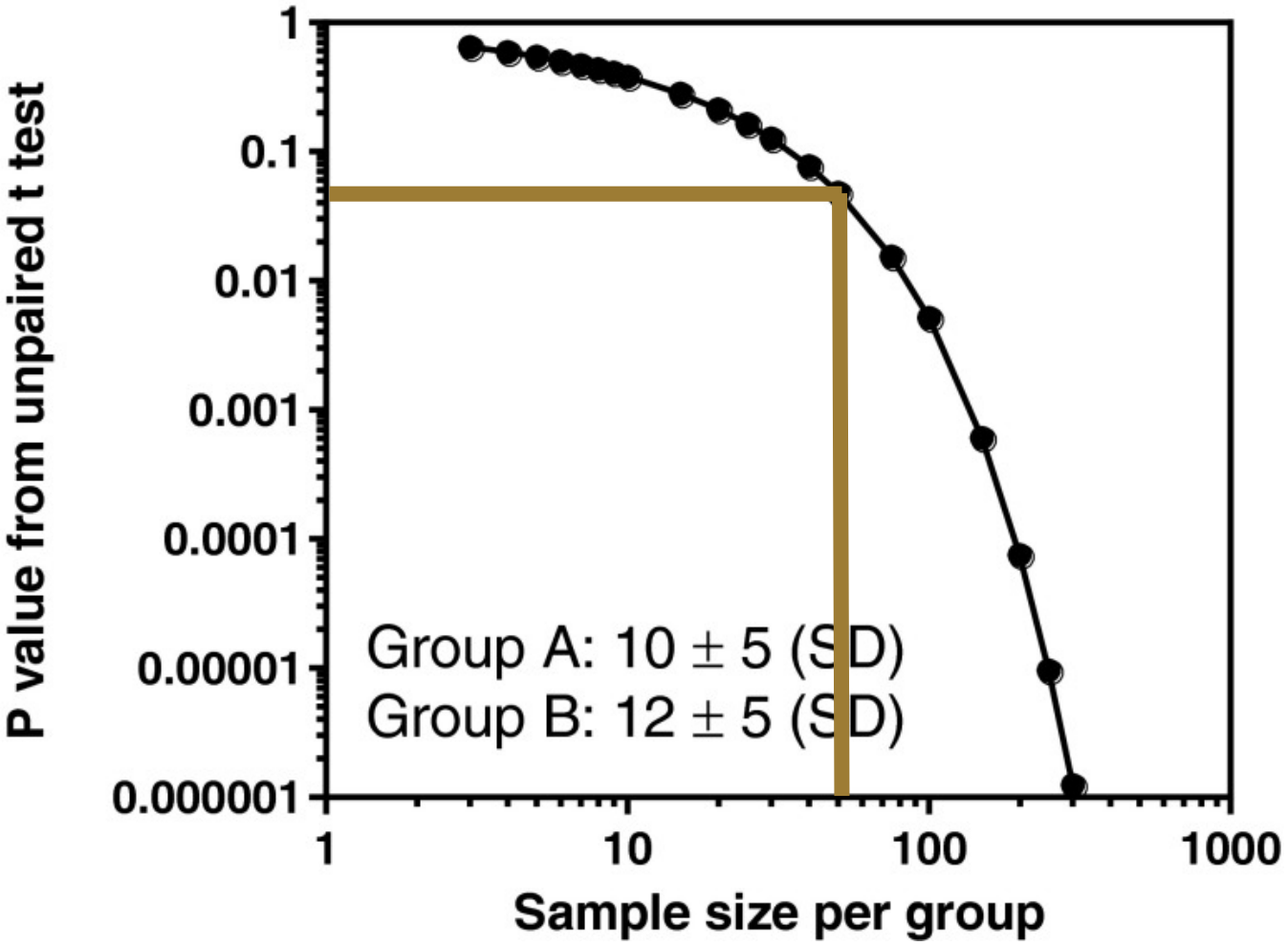
What does a p-value really mean?

- Probability of given result, under null hypothesis. It is not the probability that either hypothesis is true.
- $\alpha = 0.05$ (aka $p < 0.05$) is an arbitrary convention
- Having a p-value less than 0.05 does not mean you have done rigorous research!



Statistical Power is Driven by Sample Size, Effect Size, and Variance

- Statistical power is the probability you will correctly reject null hypothesis (e.g. detect true difference)
- Effect Size is the magnitude of difference between groups



Motulsky 2014

Summer Project Details

0 surveys completed

0 surveys underway

How many individuals are in your total study population?

<10

10-20

20-50

50-100

>100

How many dependent variables (outcomes) are you looking at? (Note each omic feature is an independent outcome.)

1

2-5

5-7

7-10

>10

Are there additional variables you are including in your analysis?

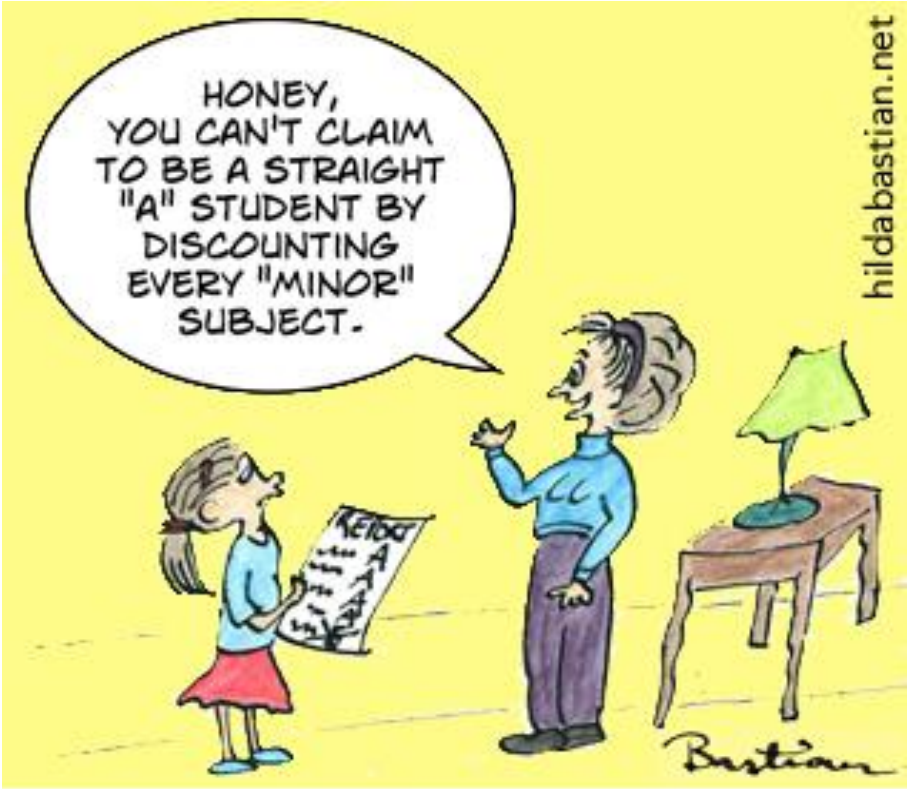
Join by Web

PollEv.com/ellenequillen

Join by QR code

Scan with your camera app

p-Hacking and “Researcher Degrees of Freedom”



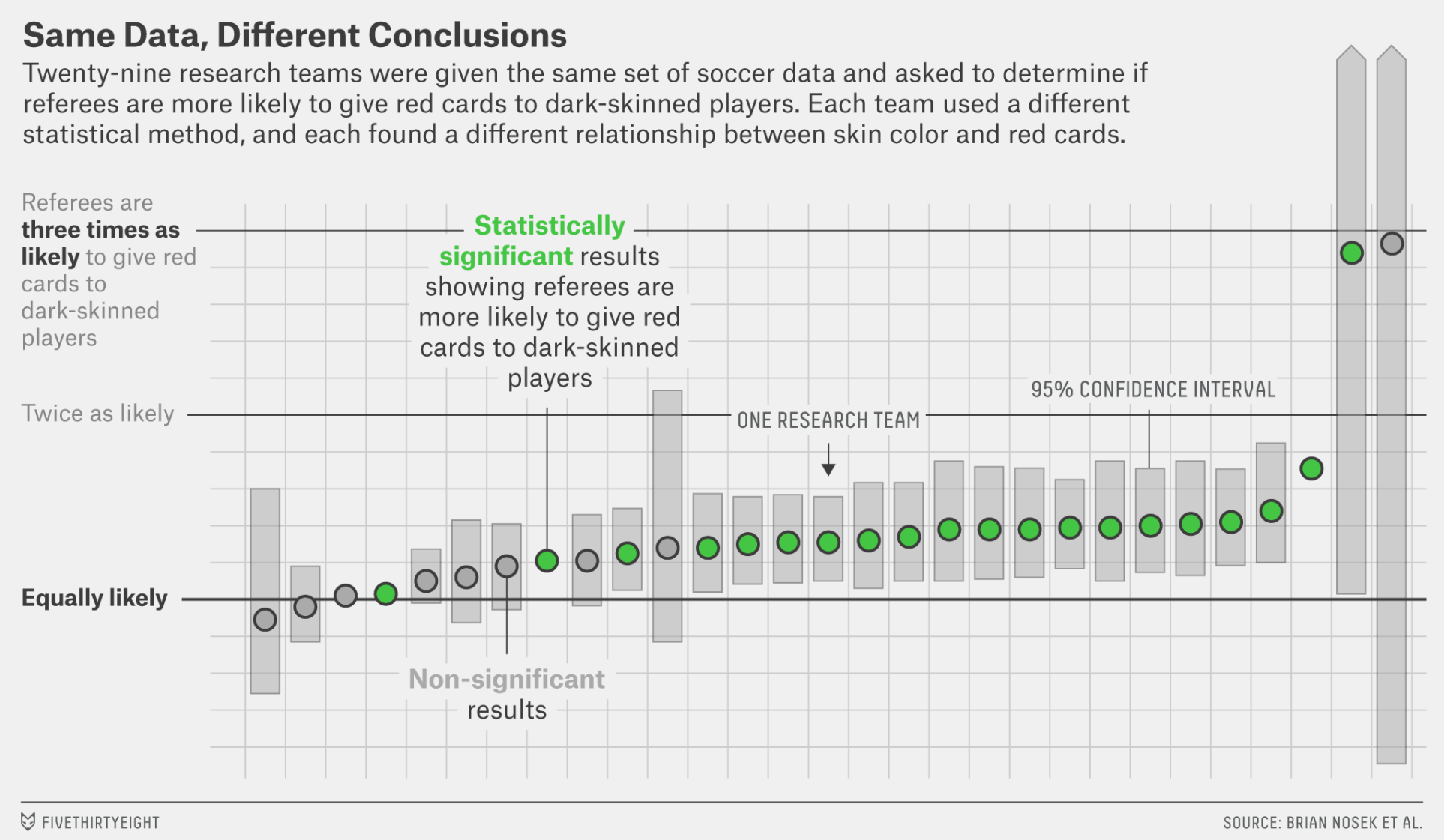
JULIE WAS EXCITED WHEN HER DAUGHTER FAILED HISTORY. AT LAST A TEACHABLE MOMENT ON THE NEED FOR UNBIASED CONSIDERATION OF ***ALL*** THE EVIDENCE!

- Analyzing data in different ways until you get a statistically significant result
  - Each test you do – including each dependent variable – should be treated like an independent test and corrected for
    - https://shinyapps.org/apps/p-hacker/

Testing New Hypotheses

- Data exploration can lead to new hypotheses but they must be tested in a SEPARATE sample set
- A new, larger sample should be recruited and fed green jelly beans
  - Sample size estimate should account for “winner’s curse” – effect size will be smaller in replication cohort
- Replication in separate cohort now mandatory for most major genetics journals

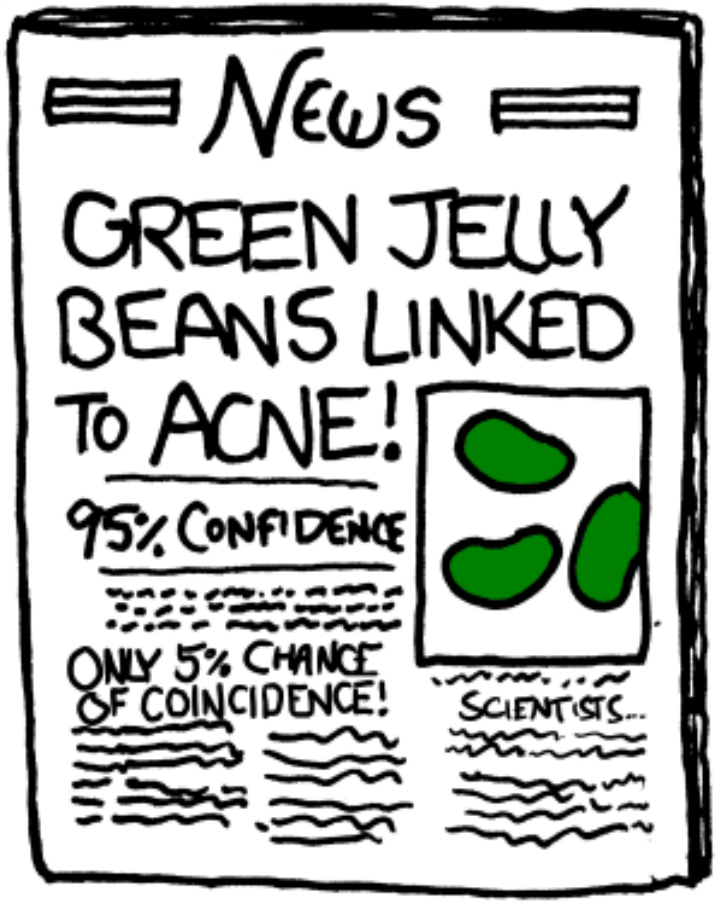
A Caveat


Same Data, Different Conclusions
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.
Referees are three times as likely to give red cards to dark-skinned players
Twice as likely
Equally likely
Statistically significant results showing referees are more likely to give red cards to dark-skinned players
ONE RESEARCH TEAM
95% CONFIDENCE INTERVAL
Non-significant results
FIVETHIRTYEIGHT
SOURCE: BRIAN NOSEK ET AL.

Traps in Interpreting Data

HARKing: hypothesizing after results are known

- Scientists frequently report only significant results
- All tested hypotheses should be reported (and adjusted for)
- Unfortunately, there is little room for negative results in most journals



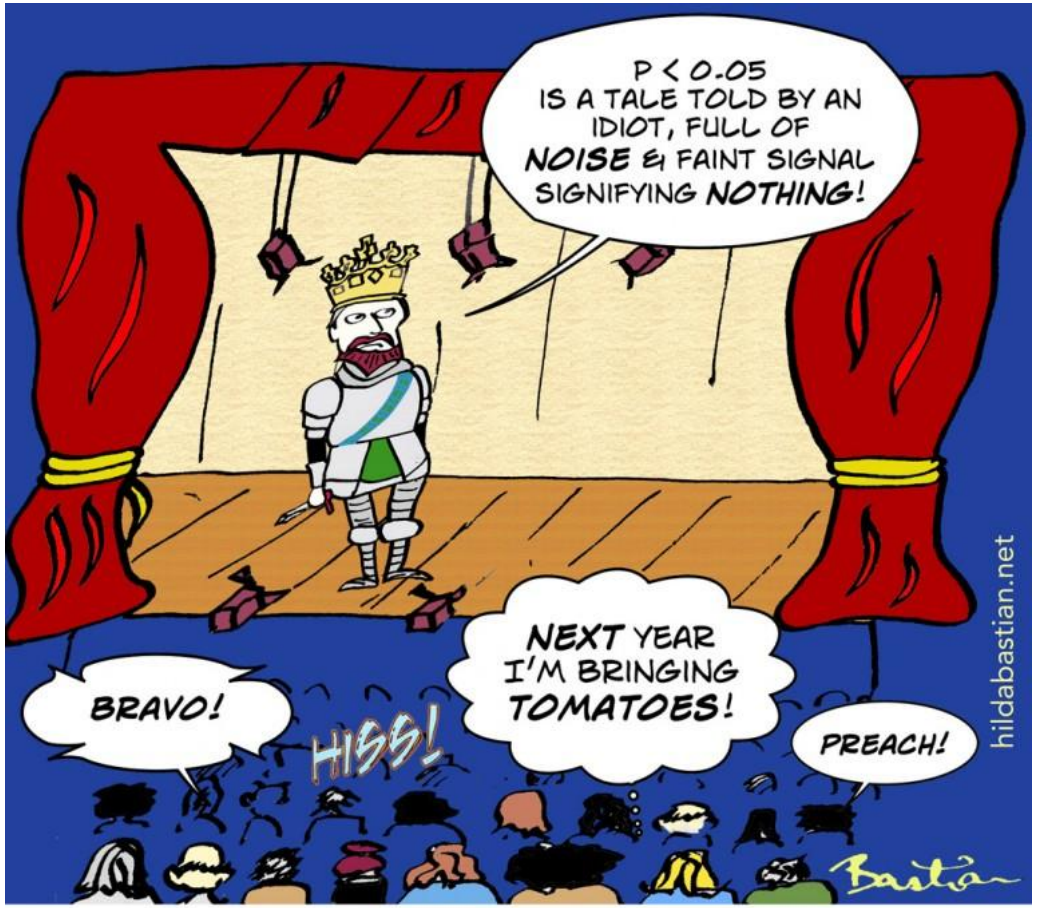
xkcd.com

Testing New Hypotheses

- Data exploration can lead to new hypotheses but they must be tested in a SEPARATE sample set
- A new, larger sample should be recruited and fed green jelly beans
  - Sample size estimate should account for “winner’s curse” – effect size will be smaller in replication cohort
- Replication in separate cohort now mandatory for most major genetics journals

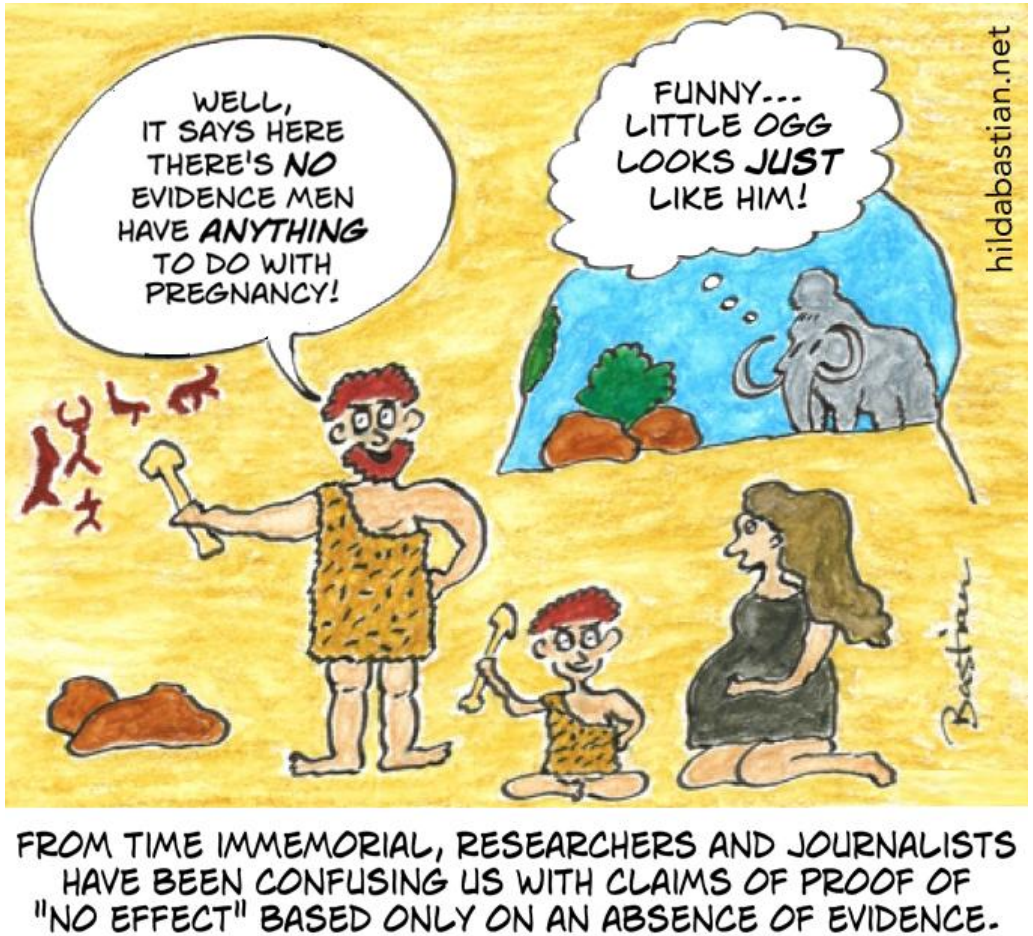
Significant does not mean important

- Results can be significant but effect very small
- Difference in mean (effect size), standard deviations, and confidence intervals give more context



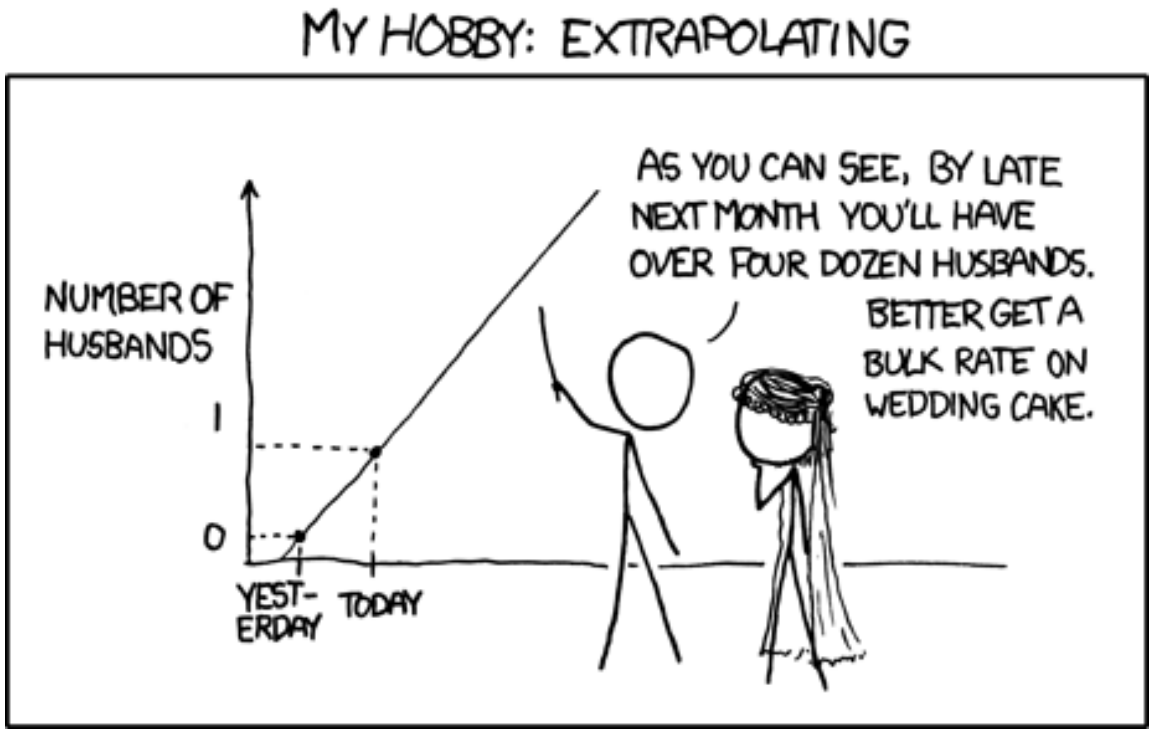
THE PROBABILITY OF AUDIENCE UPROAR IS ALWAYS HIGH FOR SHAKESPEARE NIGHT AT THE STATISTICAL SOCIETY.

Absence of Evidence is not Evidence of Absence



- $p > 0.05$ doesn't support your null hypothesis, it fails to reject it
- You can't simply reverse analysis

Extrapolating from Limited Data



- Generalizability/external validity
- The goal of biomedical research is to understand biology well enough to predict future events
- Relationship between variables may change out of tested range

xkcd.com

Which of the following is a correct interpretation of a p-value or 0.05?

The chance that the studied (alternative) hypothesis is true is 95%.

0%

The effect of our independent variable on our dependent variable is strong.

0%

Assuming the null hypothesis is true, we would have gotten the same result 5% of the time.

0%

There is a 5% chance that our results are wrong.

0%

Which of the following is a correct interpretation of a p-value or 0.05?

The chance that the studied (alternative) hypothesis is true is 95%.

0%

The effect of our independent variable on our dependent variable is strong.

0%

Assuming the null hypothesis is true, we would have gotten the same result 5% of the time.

0%

There is a 5% chance that our results are wrong.

0%

Transparency and Reproducibility

Pre-Registration is Increasingly Common

- Only reporting positive results biases the data as a whole
- Pre-registration prevents post-hoc changes in methods & allows reporting of negative results

Forbes.com

Transparency is the Bedrock of Reproducibility

- Reproducibility means an experiment will achieve same results when independently repeated
- Three types of reproducibility:
  - Methods
  - Results
  - Inferential

The Economist

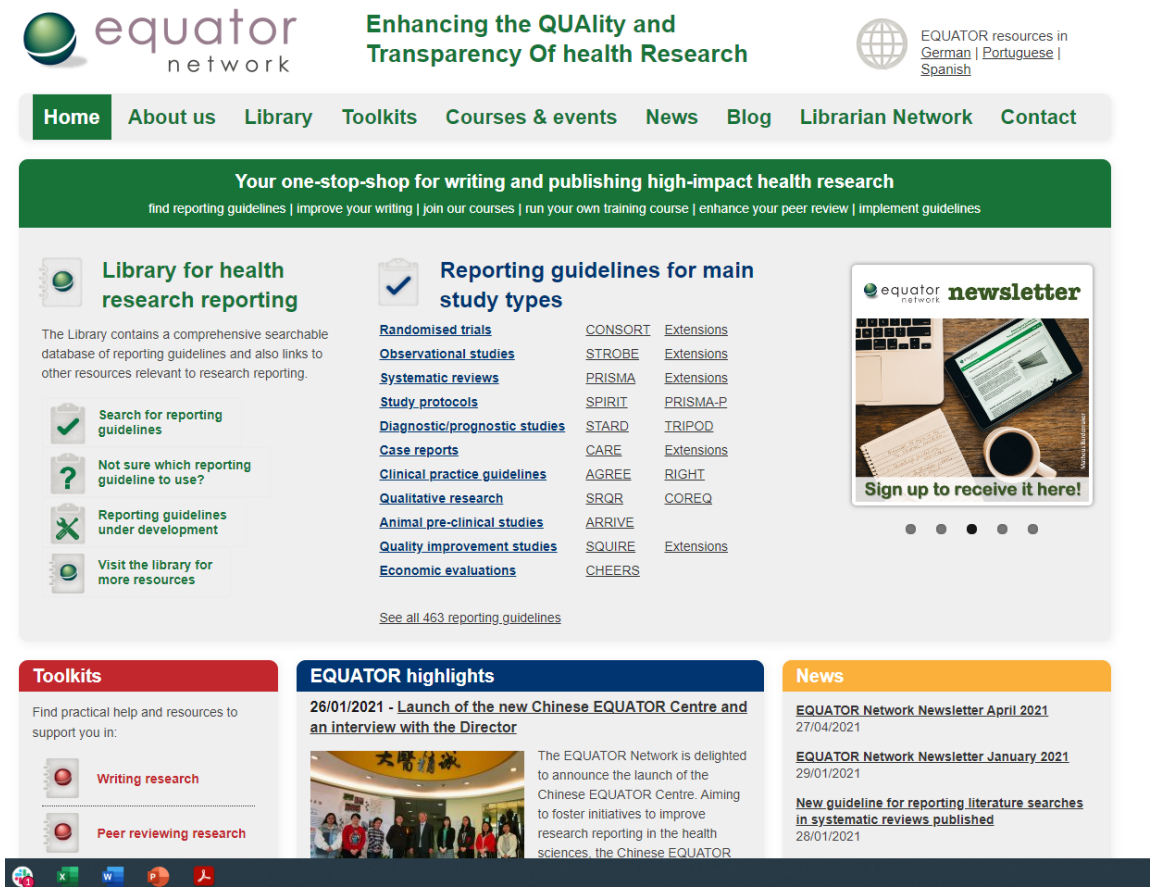
Goodman et al Science Trans Med 2016

Methods Reproducibility

- Lab protocols are often poorly documented in assumption everyone uses same methods
- Specification of any non-standard analytical methods and sharing of custom scripts
- Raw datasets must be shared, preferably in public repositories
- Specification of unique biological material (antibodies, cell lines, animals) to allow replication

Methods Reproducibility - Reporting

- Standards
- Replicates
- Statistics
- Randomization
- Blinding
- Sample-Size/Power Calculation
- Inclusion/Exclusion Criteria

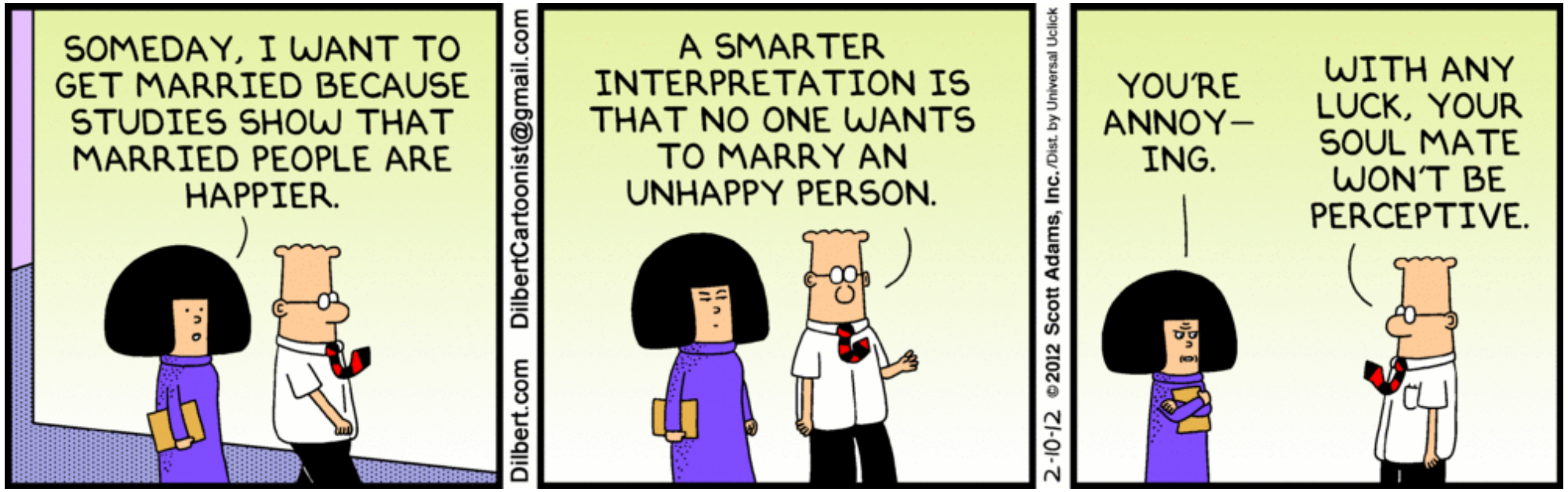


Results Reproducibility

- “Minor” environmental factors can radically alter results
- Age, sex, ancestry, etc. have major impacts on generalizability of results
  - New NIH requirements for studying sex as a biological variable
- Sample heterogeneity can mask results
- Sample homogeneity can lead to missing important effects

Inferential Reproducibility

- Scientists may draw different conclusions from the same data
- How much evidence is needed to support a hypothesis depends in part on how strong the priors are for that hypothesis



Science is People.

People are Fallible, but Improvable