When do Scientists Change their Minds?

Week 6/7 – Science, statistics, and reproducibility

EGMT-1520 Monday, Feb 21, 2022
Bill Pearson wrp@virginia.edu

Overview of this session:

- Statistics
  - p()-values ($p < 0.05$???)
  - false negatives and false positives
  - Effect size
  - Correlation and causation
  - Multiple tests

- Are most scientific papers wrong?

Final project products (Preview due Wednesday, Feb 23):

1. A 5 minute presentation (Powerpoint/Google slides) with 5 – 6 slides
   1. 2 slides explaining why the incorrect explanation is correct – please try to make a convincing case for the wrong explanation that a 10 year old would believe
   2. 1-2 slides describing the change of perspective – focus on the perspective – what is being "seen" differently (not just equations)
   3. 2 slides explaining how the change of perspective explains the phenomena, highlighting the contrast between the "intuitive" perspective and the "correct" perspective
2. (for March 2) A 750–1000 word paper making the arguments in text. Arguments should be developed in paragraphs with topic sentences and complete sentences.
3. Each slide in the presentation or section of the paper should be attributed to at least one member of the group. Each member of the group should have an attributed contribution. Slides should not overlap with other slides; like wise paragraphs in the paper should have minimal overlap.

For Wednesday (Feb 23)

Statistics in groups (15 min):

1. propose an hypothesis to be tested by measuring something (e.g. winning the NCAA basketball tournament is correlated with team height)
2. Describe a measurement result that might be a false positive, and a measurement result that would be a false-negative

Presentation pre-review (60 min):

- Review, comments on presentation paragraphs (quick look at presentation slides)

Statistics in a Nutshell

- Scientists like “reproducible” results. If only Avery can transform *Pneumococci*, why should we believe it?
- Random results are not “reproducible”, they happened by chance
- We seek results that are “not random” – so they are more likely to be “reproducible”
- p()-values attempt to establish “not random”
  - p()<0.05 says the probability of occurring ”by chance” (randomly) is < 0.05
  - But is p() < 0.049 really different from p() < 0.051?
- “significant” results can occur because of very small (but reproducible) effects measured many times (effect size)
- “significant” results can occur because of repeated tests

Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1-\beta)R/(R-\beta R+\alpha)$. A research finding is thus

5

Statistics (reproducibility) in a nutshell

- Why do we care?
  - Statistics/Reproducibility – If I make a measurement today, will I get a consistent result next week? Next year?
  - If someone claims that vaccines work, or cause serious side effects, should I believe them?
  - If we are supposed to "trust the data", is it the data we should trust, or the conclusions drawn from the data
- Scientists tend to trust data that is "statistically significant" and has a sensible mechanism
  - Double stranded DNA for replication

6

Statistical significance in a nutshell p()-values

- Scientific results are more compelling if they measure an effect that is unlikely to occur by chance
  - Vaccine adverse effects – after 220 million vaccinations, are there more heart problems than expected without vaccination?
    - How many expected
    - How many more to raise concerns?
  - If I follow Bradley Richard's investment suggestions, will I be better off than simply buying "the market"
  - If I receive a positive test for ??? (Covid19, HIV, pregnancy), is the test correct?

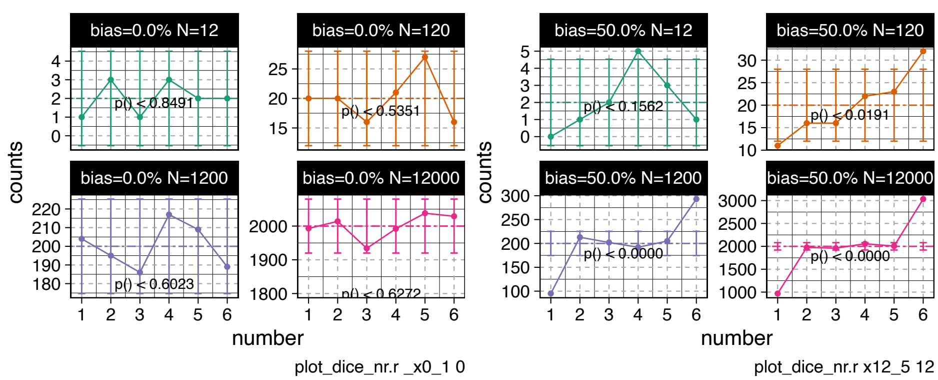
When do scientists change their minds? A quick overview of statistics

- We are more persuaded by results that are:
  - Statistically significant ($p < 0.05$?)
  - Biologically/physically/physiologically significant (effect size)
- p()-values estimate how often results would occur if a null-model is correct (?by chance?)
  - What if the null-model is wrong?
  - P()-values do not indicate the strength of the relationship
- p()<0.05 indicates the null-model would produce the results one time in 20
  - How many experiments were actually done?
- All experimental methods can produce false-positives and false-negatives
  - Statistical corrections can reduce false-positives (by increasing false negatives), and vice-versa
- Tiny effects can be statistically significant in large datasets

Statistical significance in a nutshell p()-values and the null hypothesis

- Traditionally, statistical significance is measured using “Null Hypothesis Significance Testing (NHST)”
  - Null-hypothesis testing is *backwards*
  - it does not estimate the probability that a hypothesis is true
  - It estimates the probability that the NOT-True (null) hypothesis is correct.
  - If the null-hypothesis significance test gives a probability p() < 0.05, the *hypothesis is accepted*, because the null hypothesis is likely to be wrong (how likely?)

testing statistical models: random? dice



P(1,2,3,4,5,6)=1/6
"fair"

P(1)=1/12
P(2,3,4,5)=1/6
P(6)=3/12
"loaded"

Is a die "fair" or "loaded" (un-fair)?

The debate about p()-values

Statistical Thinking Home Posts Talks Projects Datamethods Links Publications Teaching

A Litany of Problems With p-values

www.fharrell.com/post/pval-litany/

Last updated on 2020-09-15 · 10 min read · 82 Comments

In my opinion, null hypothesis testing and p-values have done significant harm to science. The purpose of this note is to catalog the many problems caused by p-values. …

Psychology journal bans *P* values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values, because the values were too often used to support lower-quality research.

Home / News & Opinion

Drop Statistical Significance, Scientists Say

In service of an arbitrary threshold, p-values often lead researchers to make poorly supported claims and ignore interesting but insignificant results, scientists argue.

Carolyn Wilke
Mar 21, 2019

346

The current way many researchers apply p-values to draw conclusions on statistical significance is incorrect and unhelpful, three scientists argue in a *Nature* commentary published yesterday (March 20). The authors urge the research community to drop the concept of statistical significance altogether, and more than 800 statisticians and scientists have signed on to the idea.

ABOVE: © ISTOCK.COM, SANJERI

Scientists have often misused p-values to make claims about what hypotheses their statistically significant or insignificant results "prove," leading to hyped claims or artificial conflict between studies, the authors say. Because of the bias in journals towards publishing findings with p-values below 0.05, scientists may ignore interesting results that don't meet the bar and may pick data or methods to try to surpass the threshold.

The authors say they are not trying to ban p-values. Rather, "we are calling for a stop to the use of P values in the conventional, dichotomous way—to decide whether a result refutes or supports a scientific hypothesis," they write.

www.nature.com/news/psychology-journal-bans-p-values-1.17001

www.the-scientist.com/news-opinion/drop-statistical-significance--scientists-say-65635

p()-values and reproducibility What a *p*()-value is not?

1. p()-values can indicate how incompatible the data are with a specified statistical model.
   – what if the model is wrong?
2. p()-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A *p()*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p()*-value does not provide a good measure of evidence regarding a model or hypothesis.

Wasserstein & Lazar (2016) *The American Statistician* **70:**129–133.

p()-values and reproducibility
What is a p-Value?

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

p-values need:
1. a statistical model (how often do we expect the result by chance)
2. a "Null" hypothesis – the result by chance would be: XYZ
3. a measurement that would reflect the effect

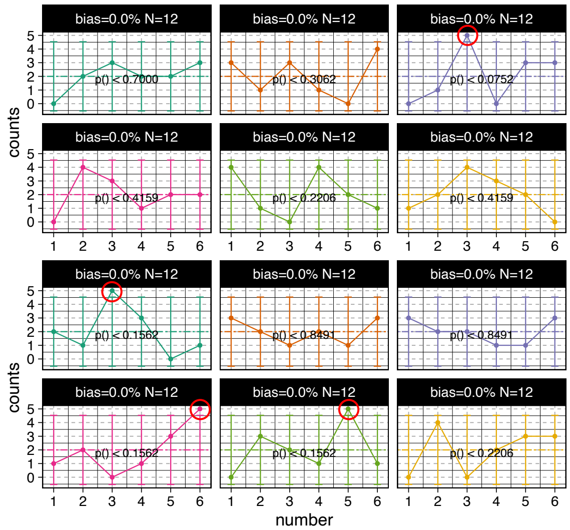
A random or loaded "die":
1. statistical model: uniform distribution p(1,2,3,4,5,6)=1/6
2. null-hypothesis: all sides equally likely
3. measurement: count how often each side appears

Wasserstein & Lazar (2016) *The American Statistician* **70:**129–133.

p()-values
What are we looking for?
"statistical" significance vs “biological” significance

- All statistical tests have two types of errors:
  - False-positives: reporting something is true when it is not
  - False-negative: reporting something is not-true when it is
- Statistical testing is more challenging when multiple tests are done
  - data-dredging, p()-hacking
- Very large datasets can generate "statistically significant" results that are very small
  - effect size

testing statistical models: random? dice



plot_dice_nr.r _x2 0

P(1,2,3,4,5,6)=1/6
"fair"

◯ False positive

How many tests?
72

With p()<0.05, how
Many false positives
Expected?
0.05 * 72 = 3.6

4 false positives in
72 "tests"

Why are all the false positives "high"? Shouldn't half be "high" and half "low"?

fasta.bioch.virginia.edu/egmt1520 15

15

Sensitivity, specificity
true-positives, true-negatives

real/measured state	Meaure True	Measure False
real True infected / loaded	True Positive TP	False Negative FN Type II error
real False healthy / fair	False Positive FP Type I error	True Negative TN

Sensitivity: TP / (TP + FN)
Specificity: TN / (TN + FP)

False Discovery Rate (FDR): FP / (TP + FP)
Positive predictive value: TP / (TP + FP)

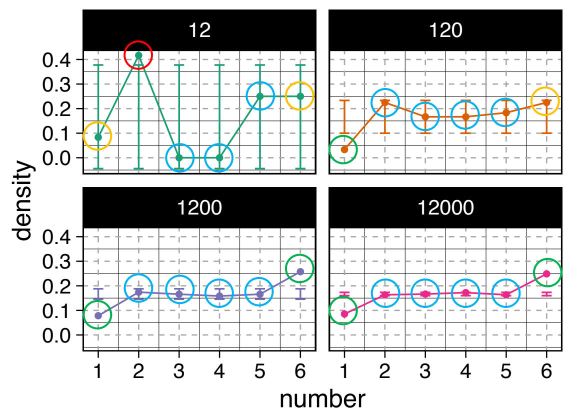
In general, false positives are considered more harmful than false-negatives (except for infectious diseases)

fasta.bioch.virginia.edu/egmt1520 16

16

sensitivity and specificity: non-random die

P(1)=1/12
P(2,3,4,5)=1/6
P(6)=3/12

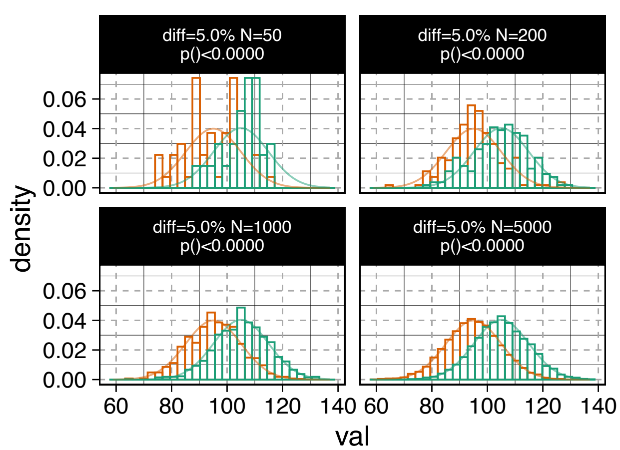


False Negative
False Positive
True positive
True negative

Sens: TP/(TP+FN)	Spec: TN/(TN+FP)
12: sens: 0/2	spec: 3/4
120: sens: 1/2	spec: 4/5
1200: sens: 2/2	spec: 4/4

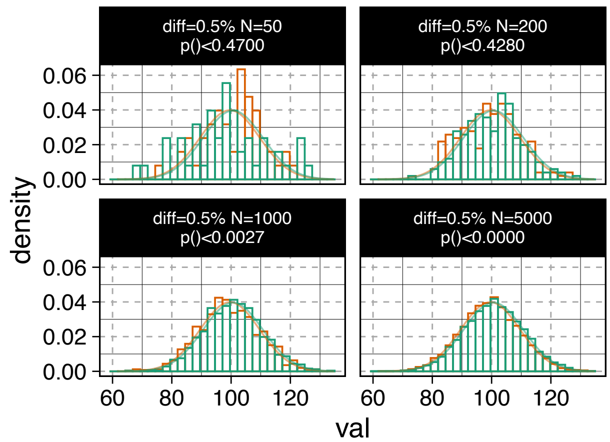
ice_nr.r 1

Statistics in a nutshell: **effect size**
statistical significance (p()< 0.05)
may not be very significant)



plot_norm_nrc.r _n5r 5

Statistical significance (p()< 0.05) may not be very "significant"



plot_norm_nrc.r _n05r 0.5

Which results are "statistically significant"?

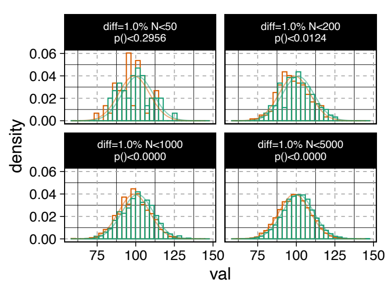
Is a 0.5% difference "significant"?

Note that the "statistical significance" does not correlate well with sample size

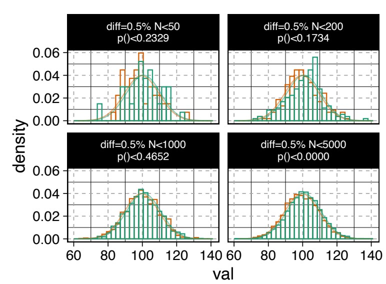
fasta.bioch.virginia.edu/egmt1520 19

19

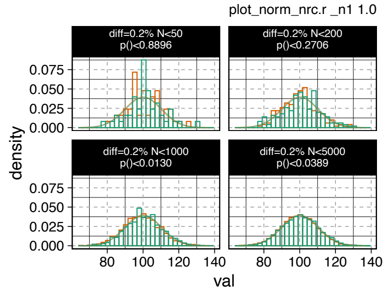
Tiny effects can be (statistically) "significant"



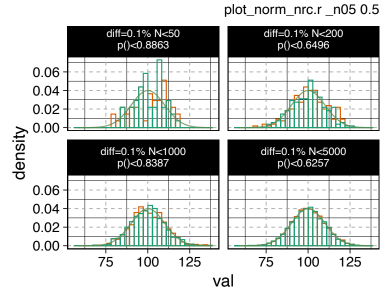
plot_norm_nrc.r _n1 1.0



plot_norm_nrc.r _n05 0.5



plot_norm_nrc.r _n02 0.2

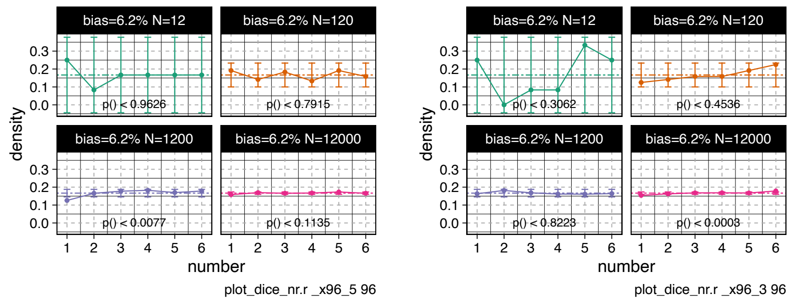


plot_norm_nrc.r _n01 0.1

fasta.bioch.virginia.edu/egmt1520 20

20

Tiny effects can be (statistically) "significant"

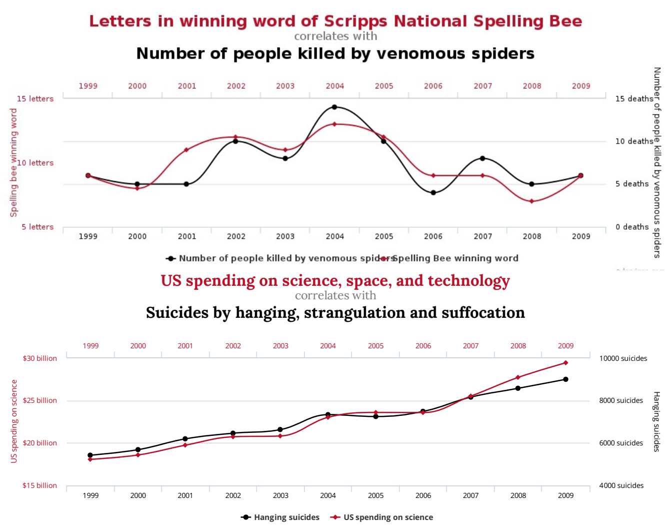


Which are false positives?
Which are false negatives?

fasta.bioch.virginia.edu/egmt1520 21

21

Data dredging and p-hacking / Correlation and causation

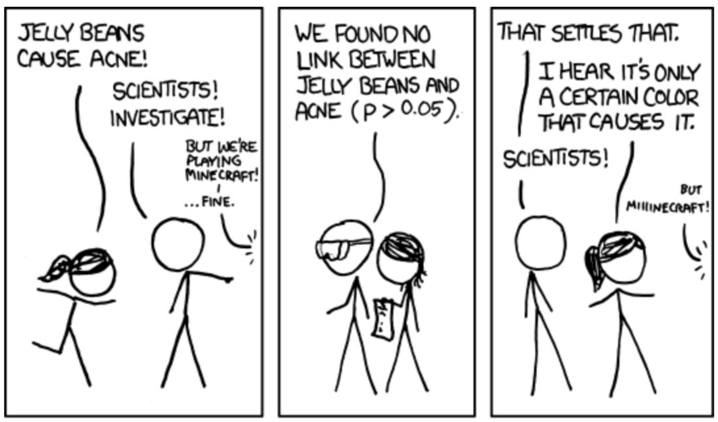


tylervigen.com/spurious-correlations

fasta.bioch.virginia.edu/egmt1520 22

22

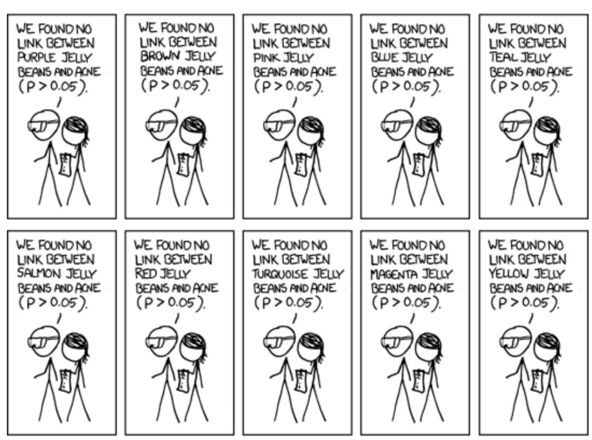
Multiple testing: So many tests, what is significant?



fasta.bioch.virginia.edu/egmt1520 23

23

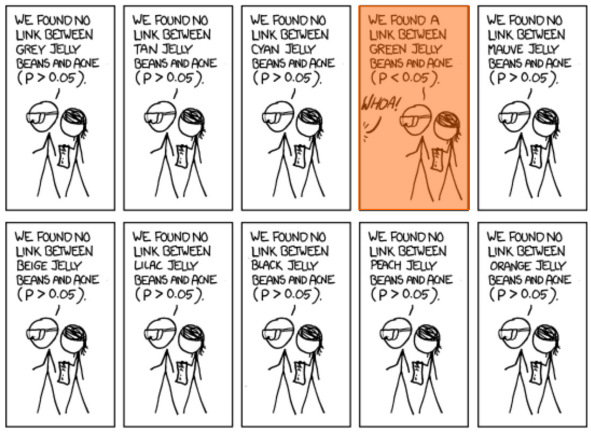
So many tests, what is significant?



fasta.bioch.virginia.edu/egmt1520 24

24

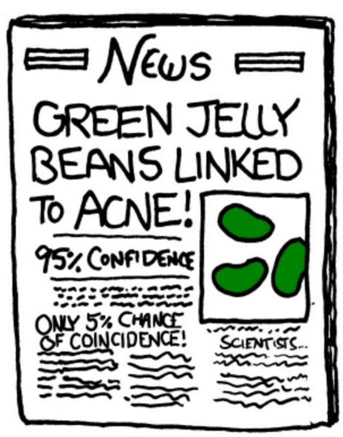
So many tests, what is significant?



fasta.bioch.virginia.edu/egmt1520 25

25

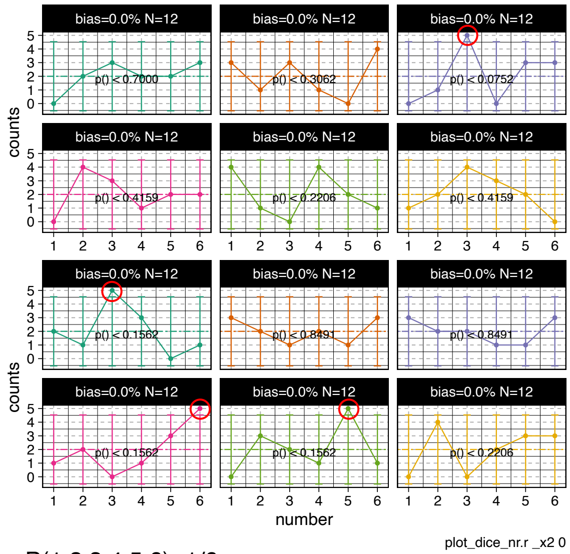
So many tests, what is significant?



fasta.bioch.virginia.edu/egmt1520 26

26

testing statistical models: random? dice



False positive

How many tests?
72

With p()<0.05, how Many false positives Expected?
0.05 * 72 = 3.6

4 false positives in 72 "tests"

P(1,2,3,4,5,6)=1/6
"fair"

fasta.bioch.virginia.edu/egmt1520 27

27

Statistics in a Nutshell

- Scientists like "reproducible" results. If only Avery can transform *Pneumococci*, why should we believe it?
- Random results are not "reproducible", they happened by chance
- We seek results that are "not random" – so they are more likely to be "reproducible"
- p()-values attempt to establish "not random"
  - p()<0.05 says the probability of occurring "by chance" (randomly) is < 0.05
  - But is p() < 0.049 really different from p() < 0.051?
- "significant" results can occur because of very small (but reproducible) effects measured many times (effect size)
- "significant" results can occur because of repeated tests

fasta.bioch.virginia.edu/egmt1520 28

28

Final project products (prelim due Feb 23):

1. A 5 minute presentation (Powerpoint/Google slides) with 5 – 6 slides
   1. 2 slides explaining why the incorrect explanation is correct – please try to make a convincing case for the wrong explanation that a 9 year old would believe
   2. 1-2 slides describing the change of perspective – focus on the perspective – what is being "seen" differently (not just equations)
   3. 2 slides explaining how the change of perspective explains the phenomena, highlighting the contrast between the "intuitive" perspective and the "correct" perspective
2. (Mar. 2) A 750–1000 word paper making the arguments in text. Arguments should be developed in paragraphs with topic sentences and complete sentences.
3. Each slide in the presentation or section of the paper should be attributed to at least one member of the group. Each member of the group should have an attributed contribution. Slides should not overlap with other slides; likewise paragraphs in the paper should have minimal overlap.