Statistical mistakes and misunderstandings: More common than you might think

David Schlesinger, Ph.D.
Lars Leksell Gamma Knife Center
University of Virginia

Conflicts of Interest
None

The “truth” is trendy at the moment....
The truth is hard to come by

One study's result is not necessarily the truth

Attempted replication of 100 studies published in 3 psychology journals. 97% of original but only 36% of replicated studies had statistically significant results. Only 47% of original effect sizes were within the 95% confidence interval of replication effect size.

Including the study about non-reproducibility

Reanalyzes the previous paper's data using a model that accounts for publication bias toward significant p-values and estimates distribution of effect sizes. Model suggests that 90% of the studies tested negligible effects. Publication bias towards p<0.05 is a main cause of lack of reproducibility.
The number of retractions is sharply rising

Number of retraction notices has increased by factor of 10
Literature has only increased by 44%
30-40% of retractions are for technical reasons

A lack of statistical fluency may be part of the problem

Many medical physicists receive little training in practical statistics as applied to clinical outcomes studies. However...these studies are at the heart of our profession. How to recognize when the statistics don't quite add up?

https://xkcd.com/925/
Session educational objectives

1. Learn about the presence of statistical problems in published studies
2. Identify common signs and symptoms of potential problems in various types of statistical tests
3. Learn methods for correctly implementing statistical analyses of the type commonly found in medical physics publications and in routine clinical activities

How we will spend the morning...

David Schlesinger  
University of Virginia  
Study irreproducibility, philosophy of inferential statistics, how to use and misuse p-values

Jenghwa Chang  
Hofstra/Northwell Health  
Normal and non-normal distributions: Why understanding distributions are important when designing experiments and analyzing data

William Sensakovic  
Florida Hospital/UCF/FSU/Adventist  
Linear and logistic regressions, what they try to explain and how to interpret the results

Mike Altman  
Washington University  
Statistical significance, statistical power, and clinical significance. How to explain your results in context.

A little quiz (not a SAMS question!)

You have a treatment you expect might alter performance on certain task:

Test with control and experimental groups of 20 subjects each  
Compare the means of each group's performance  
Result is significant per independent means t-test  
(t=2.7, df=18, p=0.01)

Which statements are true? (several or none may be correct)

Remember: (t=2.7, df=18, p=0.01)

1) You have absolutely disproved the null hypothesis
2) You have found the probability of the null hypothesis being true
3) You have absolutely proved your experimental hypothesis
4) You can deduce the probability of the experimental hypothesis being true
5) If you decide to reject the null hypothesis, you know the probability that you are making the wrong decision
6) You have a reliable experimental finding – if you repeat the experiment a large number of times you would obtain a significant result 99% of the time.


---

**Cellphone use causes cancer**

**Major Cell Phone Radiation Study Reignites Cancer Questions**

Exposure to radio frequency radiation linked to tumor formation in rats.

**Scientific American**

Led by the National Toxicology Program (NTP) under the NIH
Rats exposed to calibrated RF (GSM and CDMA) radiation for 9 hours/day over 2 years
Division into groups by SAR exposure
Association between exposure and cardiac schwannoma in male rodents (no association in female rodents)


---

**Cellphone use causes cancer – maybe?**

Study that found cell phones cause cancer in rats is riddled with red flags

Researchers strongly release partial results without external review, cause alarm.

Study was released before complete peer-review on a pre-publication website
Control rats showed less than expected natural rate of tumor incidence and died early
Incidence of tumor development correlates with age, so the early control death may have magnified the statistical findings

http://arstechnica.com/science/2016/05/study-that-found-cell-phones-cancer-in-rats-is-riddled-with-red-flags/
Human studies are mostly one-sided

<table>
<thead>
<tr>
<th>Publication Year</th>
<th>Study</th>
<th>Type</th>
<th># participants</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Interphone Study Group</td>
<td>Case-control study</td>
<td>~5000 cases; ~5000 matched controls, 13 countries</td>
<td>No overall risk*</td>
</tr>
<tr>
<td>2001 (updated 2007, 2011)</td>
<td>Danish cohort study</td>
<td>Cohort study</td>
<td>~500,000</td>
<td>No association</td>
</tr>
<tr>
<td>2013 (updated 2014)</td>
<td>Million Women Study</td>
<td>Prospective cohort study</td>
<td>791,710</td>
<td>No association</td>
</tr>
<tr>
<td>2014</td>
<td>CERENAT</td>
<td>Multicenter case control</td>
<td>447 cases, 892 matched controls</td>
<td>No association with regular use; yes association with heaviest use</td>
</tr>
<tr>
<td>2011</td>
<td>Swedish pooled analysis</td>
<td>Pooled analysis of 2 case control studies</td>
<td>1251 cases, 2438 controls</td>
<td>Increased risk of glioma</td>
</tr>
</tbody>
</table>

Cellphone use causes cancer?

We still don't know!

What question are we trying to ask?
Does cellphone use cause cancer?

Or more specifically since we are scientists running a study:

What is the probability that cellphone use causes cancer given the data from our study?

Two ways to consider the problem

<table>
<thead>
<tr>
<th>Bayesian methods</th>
<th>Frequentist methods - Null-hypothesis significance testing (NHST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directly calculates conditional probability of a hypothesis</td>
<td>Determines how extreme the observed data is</td>
</tr>
<tr>
<td>Requires an estimate of the prior probability of a hypothesis</td>
<td>Never gives the probability of a hypothesis</td>
</tr>
<tr>
<td>Data can be used as it comes in</td>
<td>Does not require an estimate of prior probability</td>
</tr>
<tr>
<td>Assumes data is fixed and hypotheses vary</td>
<td>Requires the exact specification of the experiment in advance</td>
</tr>
<tr>
<td>Assumes hypothesis is fixed and data varies</td>
<td></td>
</tr>
</tbody>
</table>

Bayes’ Theorem

\[
P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}
\]
### Bayes' Theorem

\[ P(\text{hypothesis} | \text{data}) = \frac{P(\text{data} | \text{hypothesis}) \cdot P(\text{hypothesis})}{P(\text{data})} \]

**Probability of our hypothesis given the data (posterior probability)**

- Probability we would see our data if our hypothesis is correct (likelihood)
- Our prior estimate of the probability (prior probability)

Collect data that you then use to update your previous belief

**When there are many possible hypotheses**

\[ P(\text{hypothesis} | \text{data}) = \frac{P(\text{data} | \text{hypothesis}) \cdot P(\text{hypothesis})}{P(\text{data} | \text{hypothesis}) + \sum P(\text{data} | \text{alt hypotheses})} \]

Denominator is the total probability of the data (mostly treated as a scaling constant)
What is the probability that cellphones cause cancer given the data from our study?

Develop a probability distributions for our prior beliefs (from prior studies, population incidence, etc.) – Prior probability.

Collect some sample data from a population. This can also be modeled as a distribution - likelihood.

Calculate the probability of the hypothesis given the data using Bayes’ theorem – posterior probability.

What is the probability that cellphones cause cancer given the data from our study?

Develop a probability distribution for our prior beliefs (from prior studies, population incidence, etc.)—prior probability.

Collect some sample data from a population. This can also be modelled as a distribution—likelihood.

Calculate the probability of the hypothesis given the data using Bayes’ theorem—posterior probability.

Adapted from J. Bland, D. Altman, BMJ (317), 1998.
Plotting code adapted from A. Coghlan, “a little book of r for bayesian statistics.”

Another (implicit) Bayesian example—linac QA

QA tests are a way to update prior probability of machine function to be beyond a decision threshold.

Exclusion threshold

Action threshold

Prior probability

TG-142 monthly QA mechanical

P. Chang, AJR Am J Roentgenol (152), 1989

But...we mostly don't do this....
Bayes is sometimes impractical

\[ P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis}) \cdot P(\text{hypothesis})}{P(\text{data})} \]

Real-world problems involve distributions. Often must be solved numerically. (Markov-chain Monte-Carlo)

Prior can be subjective or unknown

Sometimes requires generating all of the possible probability combinations (although we generally can avoid this)

Sometimes requires generating all of the possible probability combinations (although we generally can avoid this)

Instead, here is what we do...

Instead, here is what we do...

We get rid of what we don't know...

\[ P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis}) \cdot P(\text{hypothesis})}{P(\text{data})} \]

In real life, we often don't have data that can help us describe a prior probability. So, Bayes' theorem won't work directly....
...and we turn the problem around

We create a reference hypothesis called the Null Hypothesis (H₀).
We calculate how likely our data is assuming that this reference hypothesis is true.

Philosophically, we measure how unexpected our experimental data is. The more unexpected, the less likely the null hypothesis is true.

The null hypothesis (H₀)

Serves as a reference hypothesis
Is usually the opposite of the hypothesis you hope to be true
Is frequently stated as "no difference", but doesn't have to be
Any statistically significant result assumes the null hypothesis is true.

What is the probability I would get data this (or more) extreme, assuming cellphones do not cause cancer?

1. Identify null and experimental hypotheses
2. Determine the appropriate test statistic and its distribution
3. Specify the significance level you are going to use and get critical value
4. Calculate value of test statistic from your data
5. See if this is more extreme than critical value (i.e. calculate a p-value)

What is the probability I would get data this (or more) extreme, assuming cellphones do not cause cancer?

1. Identify null and experimental hypotheses.
   
2. Determine the appropriate test statistic and its distribution.

3. Specify the significance level you are going to use and get critical value.

4. Calculate value of test statistic from your data.

5. See if this is more extreme than critical value (i.e. calculate a p-value).

What is the probability I would get data this (or more) extreme, assuming cellphones do not cause cancer?

1. Identify null and experimental hypotheses.
2. Determine the appropriate test statistic and its distribution.
3. Specify the significance level you are going to use and get critical value.
4. Calculate value of test statistic from your data.
5. See if this is more extreme than critical value (i.e. calculate a p-value).


What should I do with the p-value?

Option 1 (Fisherian method)
Report the p-value without any statement of "rejecting the null hypothesis".

Option 2 (Neyman-Pearson method)
"Reject" the null hypothesis if p-value is below significance threshold.
Assumes you specified alternate hypothesis.
(Balance type I and type II errors).

In both cases, report the statistical test, test statistic, degrees of freedom, etc.
Not just the p-value.

Ok…but what is a p-value?

"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."

R. Wasserstein, et al., ASA Statement on Statistical Significance and P-Values, American Statistician (70), 2016.

Not clear?

R. Wasserstein, et al., ASA Statement on Statistical Significance and P-Values, American Statistician (70), 2016.
Some facts about p-values and NHST

A p-value is the probability of obtaining data equal to or more extreme than the data you actually collected if you assume the null hypothesis is true.

A p-value is always between 0.0 and 1.0

Scientific conclusions should not be solely based on whether a p-value passes a threshold

Using a threshold of p<0.05 is completely arbitrary

p-values depend on exact experimental setup (some of which can be implicit)

p-values depend on the sample size and spread of the data

Scientific conclusions should not be solely based on whether a p-value passes a threshold

p-values depend on the sample size and spread of the data


What a p-value is NOT

A p-value is not the probability that the null hypothesis is true

1.0 minus the p-value is not the probability the alternative hypothesis is true

1.0 minus the p-value is not the probability the results will hold up under repeated experiments

A high p-value does not mean the null hypothesis is true

A p-value is not the probability of your results being "a random coincidence"

p-values are not a measure of the effect size or importance of a result


The ASA's statement on p-values: context, process, and purpose

Ronald L. Wasserstein & Nicole A. Lazar

1. P-values can indicate how incompatible the data are with a specified statistical model.
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.

So why do we use NHST?

- It is objective in that everyone will agree on the p-value for given data/statistical test/experimental design
- Doesn't require prior probabilities
- Requires careful description of the experiment and choice of p-value thresholds ahead of time
- Computationally simple (and widely available)
- Used for a long time (over 100 years)

The key is to understand the limits of each method

Example - Two approaches, two different answers

|              | Using Bayes' theorem | Using NHST
|--------------|----------------------|------------
| P(fault)     | 0.01                 | P(test+ | no fault) = 0.01
| P(test+ | fault)     | 0.99                 | so p=0.01 (in this case)
| P(test+ | no fault) | 0.01                 |

a-priori probability of machine fault
probability QA test (T) is positive故障 exists
probability QA test (T) is positive if no fault

What is the probability P(fault | test+) that a positive test means machine really has the fault?

Some of our QA tests show this difference in action!
So now...back to our quiz

Which statements are true? (several or none may be correct)

Remember: (t=2.7, df=18, p=0.01)

1) You have absolutely disproved the null hypothesis
2) You have found the probability of the null hypothesis being true
3) You have absolutely proved your experimental hypothesis
4) You can deduce the probability of the experimental hypothesis being true
5) If you decide to reject the null hypothesis, you know the probability that you are making the wrong decision
6) You have a reliable experimental finding – if you repeat the experiment a large number of times you would obtain a significant result 99% of the time.

G. Gigerenzer et al., pub in The Sage handbook of quantitative methodology for social sciences, 2004
**Table 1**

<table>
<thead>
<tr>
<th>Statement (altered)</th>
<th>Psychology module</th>
<th>Professors and lecturers (teaching statistics)</th>
<th>Professors and lecturers (non-teaching statistics)</th>
<th>Professors and lecturers (non-teaching statistics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It is absolutely true</td>
<td>34</td>
<td>15</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2. Probability of it being true</td>
<td>32</td>
<td>26</td>
<td>17</td>
<td>36</td>
</tr>
<tr>
<td>3. It is absolutely false</td>
<td>20</td>
<td>13</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>4. Probability of it being false</td>
<td>59</td>
<td>33</td>
<td>34</td>
<td>60</td>
</tr>
<tr>
<td>5. Probability of wrong decision</td>
<td>64</td>
<td>67</td>
<td>72</td>
<td>86</td>
</tr>
<tr>
<td>6. Probability of right decision</td>
<td>67</td>
<td>43</td>
<td>42</td>
<td>80</td>
</tr>
</tbody>
</table>

Note: For comparison, the results of Vickers (1996) study with academic psychologists in the United Kingdom are shown in the right column.

---

**Reporting statistical tests with p-values has become a de-facto requirement**

**Medicine increasingly relies on p-values**

---


Chavalarias et al., JAMA, 3/15/2016.

---

The Journal of the American Medical Association (JAMA)
But there are problems.....

p-values just below p=0.05 are over-represented

Conclusions:
- p-values immediately below 0.05 appear to be over-represented in the literature relative to their expected frequency
- Shows evidence of systematic error including publication bias, selective reporting, methodological errors, or fraud.

Authors love to invent ways of getting around the "rules"

- "a trend towards significance (p=0.06)"
- "not absolutely significant but very probably so (p>0.05)"
- "not significant in the narrow sense of the word (p=0.29)"
- "teetering on the brink of significance (P=0.06)"
- "tantalizingly close to significance (p=0.104)"
- "possibly statistically significant (p=0.10)"
- "a nonsignificant trend toward significance (p=0.1)"

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/
The search for a p-value you like

P-hacking
Data dredging
Cherry-picking etc....
The search for a p-value you like

P-hacking
Data dredging
Cherry-picking
etc....

The search for a p-value you like

P-hacking
Data dredging
Cherry-picking
etc....

Try it yourself - there are many ways to achieve a desired story

https://xkcd.com/882/

https://xkcd.com/882/

http://fivethirtyeight.com/features/science-isn’t-broken/#part2

http://fivethirtyeight.com/features/science-isn’t-broken/#part2

http://fivethirtyeight.com/features/science-isn’t-broken/#part2

http://fivethirtyeight.com/features/science-isn’t-broken/#part2
Many papers are published with statistical shortcomings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of a wrong or suboptimal statistical test</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>No sample size / power calculation</td>
<td>13</td>
<td>41.9</td>
</tr>
<tr>
<td>Failure to prove test assumptions are not violated</td>
<td>16</td>
<td>51.6</td>
</tr>
<tr>
<td>Failure to define all statistical tests clearly and correctly</td>
<td>20</td>
<td>64.5</td>
</tr>
<tr>
<td>Failure to state which values of p indicate statistical significance</td>
<td>14</td>
<td>45.2</td>
</tr>
</tbody>
</table>

Excerpted from A. Strasak et al., American Statistician (61), 2007

(Several) Concluding thoughts....

Statistical literacy is only becoming more important

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>1990 (n=133)</th>
<th>2000 (n=122)</th>
<th>2010 (n=106)</th>
<th>p-value (X² test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>5.2%</td>
<td>26.4%</td>
<td>59.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>STATA</td>
<td>0.0%</td>
<td>5.7%</td>
<td>12.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SPSS</td>
<td>2.3%</td>
<td>4.9%</td>
<td>13.2%</td>
<td>0.002</td>
</tr>
<tr>
<td>t-test</td>
<td>21.1%</td>
<td>25.4%</td>
<td>26.4%</td>
<td>0.577</td>
</tr>
<tr>
<td>Chi-square</td>
<td>40.6%</td>
<td>41.8%</td>
<td>41.8%</td>
<td>0.471</td>
</tr>
<tr>
<td>Multiple regression</td>
<td>24.1%</td>
<td>42.6%</td>
<td>48.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>14.3%</td>
<td>22.1%</td>
<td>42.4%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

There is no substitute for scientific reasoning

Let's say we have 2 positioning devices and we are comparing position error

Device 1: -0.029 (0.067) mm
Device 2: -0.097 (0.436) mm

Welch t-test: t=1.097, df=51.325, p=0.2778
Levene's Test: F value=46.374, p<0.001

Some (I think) great resources


Special thanks to:

Jeff Sloan, Ph.D., Mayo Clinic
Michael Altman, Ph.D., Washington University
Questions?