SPER Advanced Methods Workshop

June 18th, 2018

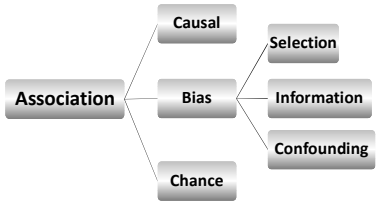
Epidemiologic evidence without p-values

Sonia Hernández-Díaz, MD, DrPH

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

HARVARD PROGRAM ON PERINATAL AND PEDIATRIC PHARMACOEPIDEMIOLOGY

Alternative explanations



Presentation follows mainly:

- Eur J Epidemiol. 2016;31:337-50. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. **Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG.**
- Am J Epidemiol. 2017 186(6):639-645 The Need for Cognitive Science in Methodology. **Greenland S**.

Warning: Am going to ruin your day

- There is no interpretation regarding p-values and related concepts that is both intuitive and totally correct.
- There are many ways to get it wrong, though some are worse than others.
- Aspects to discuss range from statistical (e.g., what is a p-value) to philosophical (e.g., do epidemiologist need to decide between true or false?).
- In my examples am not trying to point fingers. I have been wrong too.

5

Warm up example: What is wrong in this picture?

The incidence of autism spectrum disorder was 4.51 per 1000 person-years among children exposed to antidepressants vs 2.03 per 1000 person-years among unexposed children (hazard ratio [HR], 2.16 [95% CI, 1.64-2.86]; adjusted HR, 1.59 [95% CI, 1.17-2.17]). After inverse probability of treatment weighting based on the high-dimensional propensity score, the association was not significant (HR, 1.61 [95% CI, 0.997-2.59]). The association was also not significant when exposed children were compared with unexposed siblings (incidence of autism spectrum disorder was 3.40 per 1000 person-years vs 2.05 per 1000 person-years, respectively; adjusted HR, 1.60 [95% CI, 0.69-3.74]).

Conclusions: … in utero antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child. …the previously observed association may be explained by other factors.

6

No Significant -> No association

- Many studies had found that first-trimester exposure to selective serotonin reuptake inhibitors (SSRIs) was associated with increased risk of cardiovascular malformations (pooled RR=1.7, narrow CI)
- New study: "Exposure to SSRIs during the first trimester was not associated with increased risk of cardiovascular malformations (adjusted OR, 1.51; 95% CI, 0.67-3.43). …This study does not suggest a strongly increased risk of malformations, following prenatal exposure to antidepressants. "

Statistically significant -> Significant effect

Baseline characteristics	Exposed (n=6132)	Control (n=6132)	p value
Male sex	2364 (39%)	2364 (39%)	1
Age (years)	48·5 (9·8)	50·5 (12·7)	<0·0001
BMI (kg/m²)	42·0 (5·7)	41·4 (5·7)	<0·0001
Previous congestive heart failure	172 (3%)	254 (4%)	<0·0001
Previous stroke	131 (2%)	179 (3%)	0·0057
Smoking	540 (9%)	1049 (17%)	<0·0001

Public health

Is screening for breast cancer with mammography justifiable?

Peter C Gøtzsche, Ole Olsen

	Number randomised		Number of deaths from breast cancer		Relative risk (95% CI)
	Screening	Control	Screening	Control	
Randomisation adequate					
Malmö[6]	21 088	21 195	63	66	0·96 (0·68–1·35)
Canada[3,27]	44 925	44 910	120	111	1·08 (0·84–1·40)
Total	66 013	66 105	183	177	1·04 (0·84–1·27)
Randomisation not adequate					
Göteborg[4]	11 724	14 217	18	40	0·55 (0·31–0·95)
Stockholm[4]	40 318	19 943	66	45	0·73 (0·50–1·06)
Kopparberg[5]	38 589	18 582	126	104	0·58 (0·45–0·76)
Östergötland[7]	38 491	37 403	135	173	0·76 (0·61–0·95)
New York[2]	30 131	30 565	153	196	0·79 (0·64–0·98)
Edinburgh[3]	22 926	21 342	156	167	0·87 (0·70–1·08)
Total	182 179	142 052	654	725	0·75 (0·67–0·83)

Table 2: **Relative risk of death from breast cancer in screened versus control groups**

10

Major criticism

- For Göteborg Trial:
  - Women randomized to mammography on average were younger
    - by 0.09 years!
  - Suggestive of baseline differences
- Kopparberg & Östergötland:
  - Imbalance by age—women assigned to mammography older than control groups
    - 0.45 years in one
    - 0.27 years in the other

Randomization may have been inadequate

This created a great deal of controversy, misuse of p-values is no joke

"…Screening for breast cancer with mammography causes more deaths than it [prevents]."
— *Peter Gøtzsche, MD, MSc*

Eur J Epidemiol. 2017;32(1):21-29. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. Stang A, Deckert M, Poole C, Rothman KJ

- Null hypothesis significance testing (NHST) is a hybrid of significance testing advocated by Fisher and null hypothesis testing developed by Neyman and Pearson.
- NHST has become widely adopted, and widely debated.
- The principal alternative is estimation with point estimates and confidence intervals (CI).
- In Epidemiology abstracts, the CI-only approach has always been the most common approach (50%). NHST is becoming less popular.
- In JAMA, NEJM and Lancet abstracts, the predominance of the NHST approach prevailed over time. P values are reported numerically along with declarations of the presence or absence of statistical significance.

Why do p-value and NHST persist ?

- The editor of the intended journal requests statements of statistically significant.
- Researchers do not want to overinterpret findings that are non-significant.
- A "significant" result seems stronger; it sounds decisive.
- Statistical significance language provides standardized phrases. Describe the findings with own words takes longer.
- Statistical courses teach significance testing.
- …because they are intellectual shortcuts that avoid more thoughtful approaches. Dichotomization is appealing.

Why do we need to discuss p-values ?

- Epidemiologist have criticized the misuse and abuse of p-values for decades.
- Significant testing remains at the core of scientific communication.
- The abuse of statistical tests has been so uncontrollable, that some journals discourage use of ''statistical significance'' based on a P value, or even ban all statistical tests.

Enough harm done… you may not play with p-values anymore

Why do we need to discuss p-values ?

- It is critical for epidemiologists to understand the basic statistics, definition and interpretation of p-values because we participate in the consumption, production, and translation of research.
- The ultimate goal of the workshop is to reduce misreporting of results in our field (e.g., deciding which factors have an effect based on the p-values).

Agenda

- Review basic concepts
  - meaning of significance tests
  - confidence intervals, and
  - statistical power
- Review common misunderstandings:
  - Misconceptions
  - Distortions from statistical testing (dichotomania, nullisim, reification)
- Provide specific recommendations for improving interpretation and communication of results while acknowledging random variability

Models and testing

- All statistical methods are premised on the assumption that the model provides a valid representation of the variation.
- One assumption in the model is a hypothesis that a particular effect has a specific size (the test hypothesis), and the statistical methods used to evaluate it are called statistical hypothesis tests.
- Most often, the targeted effect size is a ''null'' value (zero effect), in which case the test hypothesis is called the null hypothesis.
- Nonetheless, it is also possible to test other effect sizes.
- We may also test hypotheses that the effect is no greater than a particular amount, in which case the hypothesis is said to be a one-sided.

Uncertainty and probability

- A goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect.
- It is natural to express such certainty in terms of ''probabilities'' of hypotheses.
- In conventional statistical methods, however, ''probability'' refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model.
- These methods are thus called frequentist methods, and the hypothetical frequencies they predict are called ''frequency probabilities.''
- We tend to misinterpret these frequency probabilities as hypothesis probabilities (as in Bayesian hypothesis testing)

Uncertainty and probability

- It is not the probability of the hypothesis given the observed data but the probability of the data given the hypothesis.

Pr (observation| hypothesis) ≠ Pr (hypothesis | observation)

P value tests all model assumptions

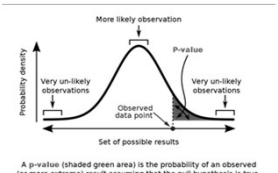
- **P value**: a statistical summary of the compatibility between the observed data and what we would predict or expect to see if the entire statistical model were correct.
- The P value tests all the assumptions about how the data were generated, not just the targeted hypothesis it is supposed to test (e.g., a null hypothesis).
- Yet, the focus of definitions of P values and statistical significance has been on null hypotheses, treating all other assumptions used to compute the P value as if they were known to be correct.
- These other assumptions, often questionable, include uncontrolled nonlinearity, randomness in sampling, treatment assignment, loss, and missingness; as well as that results were not selected for presentation based on the p-value size (P-hacking) or some other aspect.

P value tests all model assumptions

- The distance between the data and the model prediction is measured using a test statistic (such as Chi squared statistic).
- The P value is then the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis.
- The smaller the P value, the more unusual the data would be if every single assumption were correct; but a very small P value does not tell us which assumption is incorrect. It says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct.

P value distribution

- The P value can be viewed as a continuous measure of the compatibility between the data and the model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility. It measures the fit of the model to the data.



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

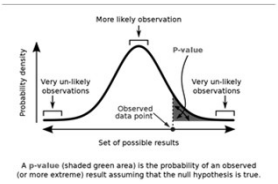
By User:Repapetilto @ Wikipedia & User:Chen-Pan Liao @ Wikipedia - File:P value.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=36661887

Statistical significance

- The P value is often degraded into a dichotomy in which results are declared ''statistically significant'' if P falls on or below a cut-off (usually 0.05) and declared ''nonsignificant'' otherwise.
- The null hypothesis is rejected if P is less than a fixed but arbitrarily pre-defined threshold value α, which is referred to as the level of significance.
- Of note, the term ''significance level'' or "alpha level" is a cut-off and should not be confused the P value itself: the cut-off value a is fixed in advance and is thus part of the study design, the data P value is a number computed from the data and thus an analysis result.

Statistical significance

- The statistically significant result should be highly improbable if the null hypothesis is true.
- However, unless there is a single alternative to the null hypothesis, the rejection of null hypothesis does not tell us which of the alternatives might be the correct one.

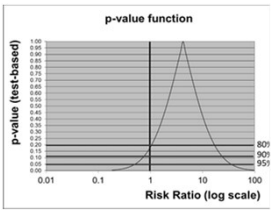


A p-value (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

By User:Repapetilto @ Wikipedia & User:Chen-Pan Liao @ Wikipedia - File:P value.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=36661887

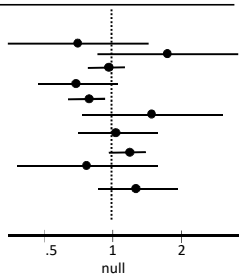
From tests to estimates

- We can estimate the P value across competing test hypotheses. For example, we may test the hypothesis that the risk ratio is 1 (the null hypothesis), or that it is 2 or 0.5.
- The effect size whose test produced P = 1 is the size most compatible with the data (in the sense of predicting what was in fact observed), and provides a point estimate of the effect under the assumption that the model is correct.
- The effect sizes whose test produced P=0.05 will typically define a range of sizes that would be considered more compatible with the data than sizes outside the range. This range corresponds to a 1 - 0.05 = 0.95 or 95 % confidence interval, and provides a convenient way of summarizing the results of hypothesis tests for many effect sizes.



From tests to estimates

- **Property**: If one calculates 95 % confidence intervals repeatedly in valid applications, 95 % of them, on average, will contain (i.e., include or cover) the true effect size.
- This coverage probability is a property of a long sequence of confidence intervals rather than a property of any single confidence interval.



From tests to estimates

- Therefore, there is a close relationship between P values and confidence intervals. However,
- The P value blends precision with effect size
- Estimates of effects + confidence intervals separate these two essential aspects: magnitude of effect and degree of precision
- For causal inference, estimating the magnitude of the effect is preferable to statistical testing. And confidence intervals represent the random error, giving a range of parameters that are consistent with the data.

From tests to estimates

- If confidence intervals are used to judge whether they contain the null value or not, they are converted to significance testing ☹
  - Lack of statistical significance ≠ Lack of effect
  - Statistical significance ≠ Important effect

Interval estimation -> dichotomization -> nullisim

Agenda

- Review common misunderstandings:
  - Misconceptions (for p-value, confidence intervals and power)
  - Distortions from statistical testing (dichotomania, nullisim, reification)

Misconceptions in the interpretation of P-values

- **Summary**:
  - The P value is the probability that the statistic would have been at least as large as its observed value if the hypothesis were correct.
  - Consequently, the P value measures the compatibility of the data with the (null) hypothesis, not the probability that the (null) hypothesis is correct.
  - The p-value blends precision with size effect.
  - A significance test is a degraded version of a p-value.

Misunderstanding #1

- **The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1 % chance of being true; if instead it gave P = 0.40, the null hypothesis has a 40 % chance of being true.**
- No. The P value assumes the test hypothesis is true. The P value indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the underlying model.
- Thus P = 0.01 would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while P = 0.40 would indicate that the data are much closer to the model prediction.

Misunderstanding #2

- **The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8 % probability that chance alone produced the association.**
- No. To say that chance alone produced the observed association is logically equivalent to asserting that every assumption used to compute the P value is correct, including the null hypothesis.
- The P value is a probability computed assuming chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the P value, which is a probability deduced from a set of assumptions (the statistical model), can possibly refer to the probability of those assumptions.

Misunderstanding #9

- **The P value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5 % of the time under the test hypothesis.**
- No. The P value refers not only to what we observed, but also observations more extreme than what we observed (where ''extremity'' is measured in a particular way).

Misunderstanding #10

- **If you reject the test hypothesis because P < 0.05, the chance you are in error (the chance your ''significant finding'' is a false positive) is 5 %.**
- No. To see why this description is false, suppose the test hypothesis is in fact true. Then, if you reject it, the chance you are in error is 100 %, not 5%. The 5% refers only to how often you would reject it, and therefore be in error, over very many uses of the test across different studies when the test hypothesis is true. It <u>does not refer to your single use of the test</u>.

Misunderstanding #11

- **P = 0.05 and P < 0.05 mean the same thing.**
- No. This is like saying reported height = 2 m and reported height <2 m are the same thing: ''height = 2 m'' would include few people and those people would be considered tall, whereas ''height <2 m'' would include most people.
- Similarly, P = 0.05 would be considered a borderline result in terms of statistical significance, whereas P < 0.05 lumps borderline results together with results very incompatible with the model (e.g., P = 0.0001) thus rendering its meaning vague, for no good purpose.

Misunderstanding #12

- **P values are properly reported as inequalities (e.g., report ''P < 0.02'' when P = 0.015 or report ''P > 0.05'' when P = 0.06 or P = 0.70).**
- No. This is bad practice because it makes it difficult or impossible for the reader to accurately interpret the statistical result.
- Only when the P value is very small (e.g., under 0.001) does an inequality become justifiable: There is little practical difference among very small P values.

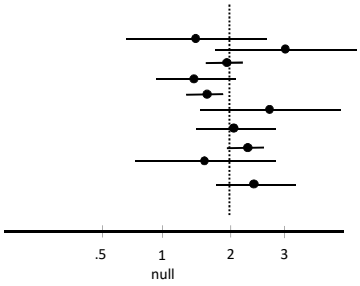
Misconceptions in the interpretation of CIs

- **Summary:**
  - Measuring effect size and its precision separately is an approach preferable to statistical testing. Random error can be expressed through confidence intervals.
  - Confidence intervals are quantitative measures that indicate the magnitude of effect and degree of precision.
  - If the confidence intervals are used to merely decide whether they contain the null value or not, they are converted into a significance test. Dichotomous interpretations are an unfortunate consequence of significance testing.

Misunderstanding #19

- **The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size.**
- No. A reported confidence interval is a range between two numbers. The frequency with which an observed interval contains the true effect is either 100 % if the true effect is within the interval or 0 % if not; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true size if all the assumptions used to compute the intervals were correct.
- It is possible to compute an interval that can be interpreted as having 95 % probability of containing the true value; nonetheless, such computations require further assumptions about the size of effects in the model (a prior distribution), and the resulting intervals are usually called Bayesian posterior (or credible) intervals.

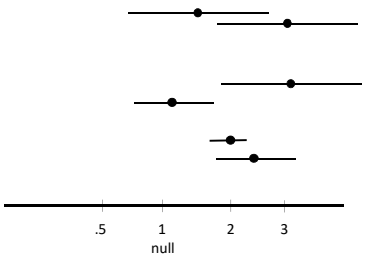
Interpretation of CIs



Misunderstanding #21

- **If two confidence intervals overlap, the difference between two estimates or studies is not significant.**
- No. The 95 % confidence intervals from two subgroups or studies may overlap substantially and yet the test for difference between them may still produce P<0.05.
- Comparison between groups requires statistics that directly test and estimate the differences across groups.
- However, if the two 95% confidence intervals fail to overlap, then when using the same assumptions used to compute the confidence intervals we will find P<0.05 for the difference; and if one of the 95 %intervals contains the point estimate from the other group or study, we will find P>0.05 for the difference.

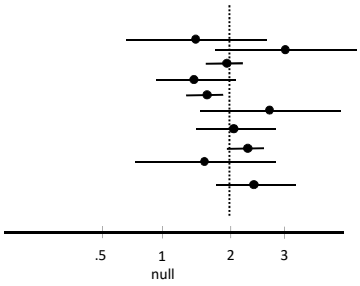
Interpretation of CIs



Misunderstanding #22

- **An observed 95 % confidence interval predicts that 95 % of the estimates from future studies will fall inside the observed interval.**
- First, 95% is the frequency with which other unobserved intervals will contain the true effect, not how frequently the one interval being presented will contain future estimates. The chance that a future estimate will fall within the current interval will usually be much less than 95%.
- Again, an observed interval either does or does not contain the true effect; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

Interpretation of CIs

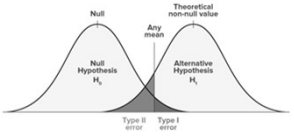


Misunderstanding #23

- **If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one.**
- No. When the model is correct, precision of statistical estimation is measured directly by confidence interval width (measured on the appropriate scale). It is not a matter of inclusion or exclusion of the null or any other value.

Interpretation of Power

- The power of a test to detect a correct alternative hypothesis is the pre-study probability that the test will reject the test hypothesis (e.g., the probability that P will not exceed a pre-specified cut-off such as 0.05). The corresponding pre-study probability of failing to reject the test hypothesis when the alternative is correct is one minus the power, also known as the Type-II or beta error rate



Interpretation of Power

- As with P values and confidence intervals, this probability is defined over repetitions of the same study design and so is a frequency probability
- Reporting the power of a study as part of the results "post-hoc power calculation" makes no sense.. When you have the study results there is no need to hypothesize about the magnitude of the association. You can estimate it. The confidence interval conveys all the information on precision.

Misunderstanding #25

- **If you accept the null hypothesis because the null P value exceeds 0.05 and the power of your test is 90 %, the chance you are in error (the chance that your finding is a false negative) is 10 %.**
- If the null hypothesis is false and you accept it, the chance you are in error is 100 %, not 10 %. Conversely, if the null hypothesis is true and you accept it, the chance you are in error is 0 %. The 10 % refers only to how often you would be in error over very many uses of the test across different studies when the particular alternative used to compute power is correct and all other assumptions used for the test are correct in all the studies. It does not refer to your single use of the test or your error rate under any alternative effect size other than the one used to compute power.

Dichotomania

- The compulsion to replace quantities with dichotomies
  - Degrading P values to "significant" or "non significant based on an arbitrary cutoff
  - Degrading confidence intervals to including the null value "yes or no".
- Example: "Our findings are conflicting with earlier results, our estimated risk ratio is 1.20 (95% CI 0.97-1.48) as opposed to a previously reported risk ratio of 1.20 (1.09-1.33)."
- In most scientific settings, the arbitrary classification of results into ''significant'' and ''non-significant'' is unnecessary; estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference.

Misunderstanding #3

- **A significant test result (P < 0.05) means that the test hypothesis is false or should be rejected.**
- No. P < 0.05 only means that a discrepancy from the hypothesis prediction would be as large or larger than that observed no more than 5 % of the time if only chance were creating the discrepancy.
- A small P value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; it may be small because there was a large random error or because some assumption other than the test hypothesis was violated (for example, the assumption that this P value was not selected for presentation because it was below 0.05).

Misunderstanding #4

- **A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted.**
- No. A large P value only suggests that the data are not unusual if all the assumptions used to compute the P value (including the test hypothesis) were correct.
- The same data would also not be unusual under many other hypotheses.
- Even if the test hypothesis is wrong, the P value may be large because it was inflated by a large random error or because of some other erroneous assumption (for example, the assumption that this P value was not selected for presentation because it was above 0.05).
- P>0.05 only means that a discrepancy from the hypothesis prediction would be as large or larger than that observed more than 5 % of the time if only chance were creating the discrepancy.

Misunderstanding #15

- **When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis.**
- No. This belief is often used to claim that a literature supports no effect when the opposite is case. In reality, every study could fail to reach statistical significance and yet when combined show a statistically significant association and persuasive evidence of an effect.
- For example, if there were five studies each with P = 0.10, none would be significant at 0.05 level; but when these P values are combined using the Fisher formula, the overall P value would be 0.01.
- Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

Misunderstanding #16

- **When the same hypothesis is tested in two different populations and the resulting P values are on opposite sides of 0.05, the results are conflicting.**
- No. Statistical tests are sensitive to many differences between study populations that are irrelevant to whether their results are in agreement, such as the sizes of compared groups in each population.
- As a consequence, two studies may provide very different P values for the same test hypothesis and yet be in perfect agreement (e.g., may show identical observed associations).
- Differences between results must be evaluated by directly, for example by estimating and testing those differences to produce a confidence interval and a P value comparing the results (analysis of heterogeneity).

Misunderstanding #17

- **When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement.**
- No. Two different studies may exhibit identical P values for testing the same hypothesis yet also exhibit clearly different observed associations.
- For example, suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce P = 0.003 in both; yet the test of the hypothesis of no difference in effect across studies gives P = 0.03, reflecting the large difference (12.00 - 3.00 = 9.00) between the mean differences.

Nullisim (Pseudo-skepticism)

- Nullisim: Assuming that the null is true in most settings; the null is treated as true until it is proven false.
  - To avoid false leads, statistical tests are designed to counter the natural tendency to see patterns.
  - Tests criteria minimize false-positives. At the cost of missing true effects. For example, requiring a 5% false positive rate (type I error) and a 20% false negative rate (type II error for a 80% power).
- However, some times the prior probability supports an effect (e.g., that anti-inflammatory drugs reduce inflammation) and the cost of false negatives may be larger than the false positives (e.g., that coxibs increase cardiovascular events).

Nullisim

- Explained by human aversion to admitting ignorance or uncertainty
  - Instead of believing one hypothesis until falsified, refutationism involves never asserting a hypothesis is true; recognize that available evidence is inconclusive.
- Null bias increases with multiple comparison adjustment. (Note that some procedures such as shrinking methods are justified for prediction model selection and exploration, e.g., genomics.)

Misunderstanding #5

- **A large P value (e.g., 0.70) is evidence in favor of the test hypothesis.**
- No. Any P value less than 1 implies that the test hypothesis is not the hypothesis most compatible with the data, because any other hypothesis with a larger P value would be even more compatible with the data.
- A P value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller P values.
- Furthermore, a large P value often indicates only that the data are incapable of discriminating among many competing hypotheses.
- The hypothesis most compatible with the data would be that with P = 1. But even if P = 1, there will be many other hypotheses that are highly consistent with the data, so that a definitive conclusion of ''no association'' cannot be deduced from a P value, no matter how large.

Misunderstanding #6

- **A null-hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.**
- No. Observing P>0.05 for the null hypothesis only means that the null is one among the many hypotheses that have P>0.05. Thus, unless the point estimate (observed association) equals the null value exactly, it is a mistake to conclude from P>0.05 that a study found ''no association'' or ''no evidence'' of an effect.
- If the null P value is less than 1 some association must be present in the data, and one must look at the point estimate to determine the effect size most compatible with the data under the assumed model.

Statistical Reification & Overconfidence in CI

- Treating statistical models as if they reflected physical laws.
- Significance tests and confidence intervals do not by themselves provide a basis for concluding an effect is present or absent with a given probability.
- Lack of statistical significance must not be interpreted as lack of association. The same results may be even more compatible with alternative hypotheses.
- A statistically significant association may be due to chance. (Regardless of whether based on P value cut-off or the confidence interval excluding the no-effect value.)

Statistical Reification & Overconfidence in CI

- Statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations.
  - Lack of statistical significance does not imply lack of effect. (small sample)
  - Statistically significant effects may be weak unimportant effects (large sample)

Misunderstanding #7

- **Statistical significance indicates a scientifically or substantively important relation has been detected**.
- No. Especially when a study is large, very minor effects or small assumption violations can lead to statistically significant tests of the null hypothesis.
- A small null P value simply flags the data as being unusual if all the assumptions used to compute it (including the null hypothesis) were correct; but the way the data are unusual might be of no clinical interest. One must look at the confidence interval to determine which effect sizes of scientific or other substantive (e.g., clinical) importance are relatively compatible with the data, given the model.

Misunderstanding #8

- **Lack of statistical significance indicates that the effect size is small**.
- No. Especially when a study is small, even large effects may be ''drowned in noise'' and thus fail to be detected as statistically significant by a statistical test.
- A large null P value simply flags the data as not being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; but the same data will also not be unusual under many other models and hypotheses besides the null.
- Again, one must look at the confidence interval to determine whether it includes effect sizes of importance.

Misunderstanding #13

- **Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.**
- This misinterpretation is promoted when researchers state that they have or have not found ''evidence of'' a statistically significant effect. The effect being tested either exists or does not exist. ''Statistical significance'' is a dichotomous description of a P value (that it is below the chosen cut-off) and thus is a property of a result of a statistical test; it is not a property of the effect or population being studied.

Misunderstanding #20

- **An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data.**
- As with the P value, the confidence interval is computed from many assumptions, the violation of which may have led to the results. Thus it is the combination of the data with the assumptions, along with the arbitrary 95 % criterion, that are needed to declare an effect size outside the interval is in some way incompatible with the observations. Even then, judgements as extreme as saying the effect size has been refuted or excluded will require even stronger conditions.

A model, assumptions beyond statistics

- Crucial assumption: The analyses themselves were not guided toward finding non significance or significance (analysis bias), and that the analysis results were not reported based on their non significance or significance (reporting bias and publication bias).
- Selective reporting renders false the meanings of statistical significance, P values, and confidence intervals.
- Because author decisions to report and editorial decisions to publish results often depend on whether the P value is above or below 0.05, selective reporting has been identified as a major problem.

Discussion

- Statistical tests are usually misinterpreted; what are their benefits?
- A mechanism by which chance could be put out of the equation and the researcher freed to focus on systematic errors and biologic plausibility for assessment of causality would have been a benefit.
- They were originally intended to account for random variability as a source of error, as a note of caution against overinterpretation of observed associations as true effects.
- Then use was turned on its head to provide fallacious support for null hypotheses in the form of ''failure to achieve'' or ''failure to attain'' statistical significance.

Discussion

- Neyman and Pearson wrote that ''it is doubtful whether the knowledge that [a P value] was really 0.03 (or 0.06), rather than 0.05…would in fact ever modify our judgment'' and that ''The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision.'' Decision not based on a fixed level of significance but in the light of evidence and ideas.
- Hill lamented that too often we deduce 'no difference' from 'no significant difference.'''

Discussion

- Some misinterpretations are harmless in tightly controlled experiments.
- Harms of statistical testing in more uncontrollable research settings (such as health) have far outweighed its benefits, leading to calls for banning such tests in research reports

Recommendation

- Given the absence of generally accepted alternative methods, there have been many attempts to salvage P values by detaching them from their use in significance tests.
- One approach is to focus on P values as continuous measures of compatibility and avoid comparison of P values with arbitrary cutoffs such as 0.05.
- Provide P values for relevant alternative hypotheses; for example, one could provide P values for those effect sizes that are recognized as scientifically reasonable alternatives to the null.
- Interpret them with reference to all the assumptions, including biases

Recommendation

- Shift emphasis from hypothesis testing to estimation
- Examine the effect size and precision of the estimate (confidence limits)
- But, it is crucial to consider the full range of the interval to discuss what uncertainty would remain even if there were no biases. Do NOT focus on whether it contained the null.
- As with P values, cautions are needed to avoid misinterpreting confidence intervals as providing sharp answers when none are warranted.

Recommendation

- Critically examine the assumptions used for the statistical analysis—not just the statistical assumptions, but also the hidden assumptions about how results were generated and chosen for presentation.
- When many possible associations are examined using a criterion of $p \leq 0.05$, the probability of finding at least one that meets the critical point increases in proportion to the number of associations that are tested.
- State analysis protocols a priori, enforce registration of trials (and observational studies with arguable limitations), along with open data and analysis code from all completed studies.

Recommendation

- Statistical tests should never constitute the sole input to inferences or decisions about associations or effects.
- No inference should be based on a single study.
- Examine and synthesize all results relating to a scientific question, rather than focus on individual findings.
- Epidemiologists do not need to decide whether the effect is present or absence, important or unimportant, based on their single estimate. Decisions in science, or policy, are more complex.

References

- Eur J Epidemiol. 2016;31:337-50. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. **Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG.**
- Am J Epidemiol. 2017 186(6):639-645 The Need for Cognitive Science in Methodology. **Greenland S**.
- The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 2016; 70:2, 129-133, **Ronald L. Wasserstein & Nicole A. Lazar**
- J Gen Intern Med. 2014 Jul;29:1060-4. Six persistent research misconceptions. **Rothman K**J.
- Am J Public Health. 2005;95 Suppl 1:S144-50. Causation and causal inference in epidemiology. **Rothman KJ, Greenland S**.
- Am J Epidemiol. 2017 186(6):627-635. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. **Lash TL.**
- Eur J Epidemiol. 2010;25(4):223-4. Curbing type I and type II errors. **Rothman KJ.**
- Biom J. 2017 ;59(5):854-872. A critical evaluation of the current "p-value controversy". **Wellek S**
- Contribution to the discussion of ""A critical evaluation of the current p-value controversy"". Biom J. 2017 ;59(5):892-894. **Senn S**