P-values:
What They Are
and What They Are Not

P-values:
What They Are
and What They Are Not

Fridtjof Thomas, PhD

Associate Professor, Division of Biostatistics

TN-CTSI seminar on statistical reasoning

in biomedical research

https://tnctsi.uthsc.edu/

Additional Seminars in the Series:

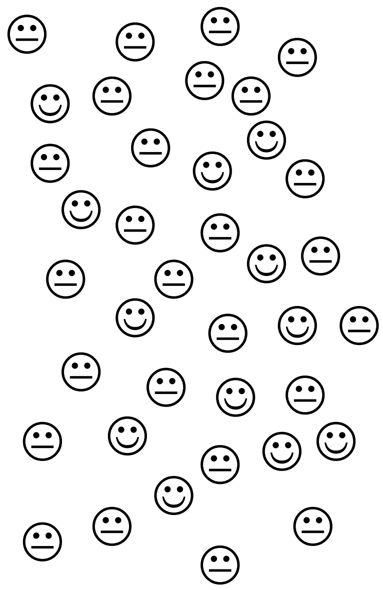
- May 7th Should We eliminate P-Values or Use More of Them: A Discussion on the P-Value Controversy (Saunak Sen, PhD)
- May 14th The Bayesian Approach to Data Analysis (Fridtjof Thomas, PhD)
- May 21st Multiple Testing and the False Discovery Rate (Saunak Sen, PhD)
- May 28th The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- June 4th Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

Outline

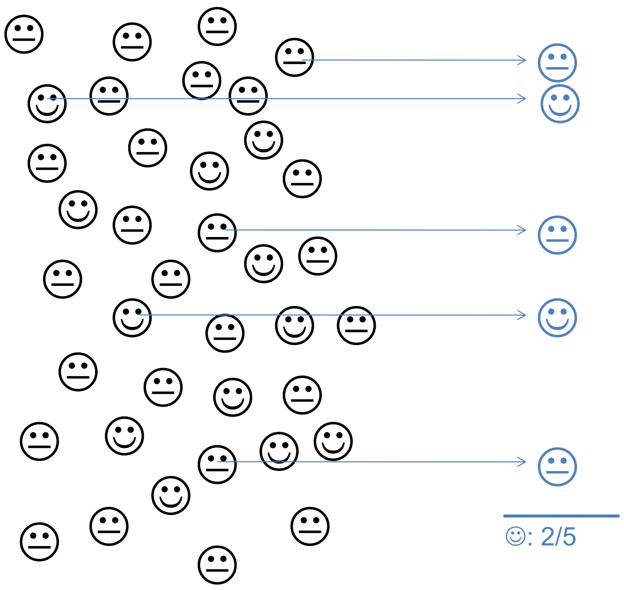
P-values are the bread and butter of applied statistics

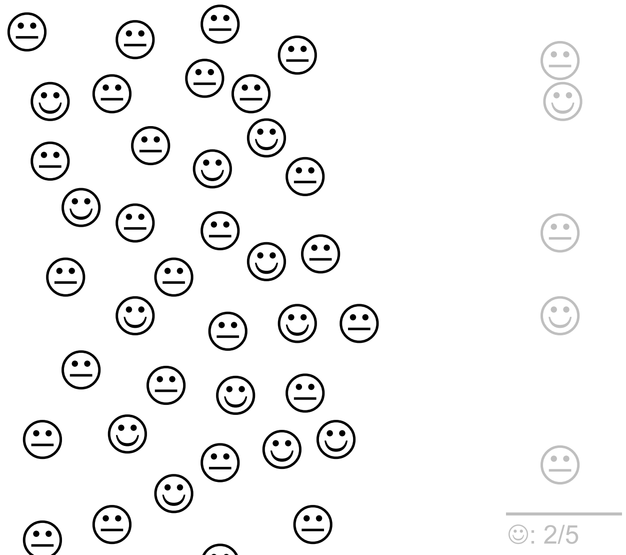
- ➢ Good examples of using p-values and how to interpret them
- ➢ Widely understood problems with p-values:
  - Statistical significance and sample size
  - Multiplicity in testing
- ➢ Mudding the waters I: P-values in tables comparing demographics in two randomized groups
- ➢ Mudding the waters II: P-values in observational studies
- ➢ What we like in p-values
- ➢ What we would like to have but p-values don't give us

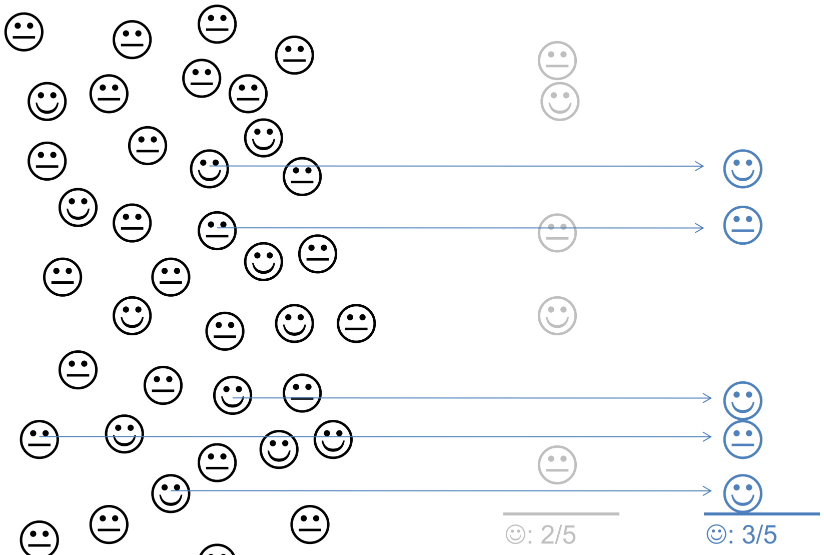
Title of talk taken from Schervish MJ. The American Statistician. 1996;50(3):203-6.



Where does our observed variability in data come from?







We have to keep in mind that variability enters our research through *sampling variation*.

If we want to test a null hypothesis such as that there are exactly 50% ☺-faces around, we have to realize that it is unlikely that we are going to observe exactly 50% ☺-faces in our sample (*VERY "unlikely" indeed if we collect 5 individuals…*).

? What %-observations are compatible with our "exactly 50% STATEMENT?"

In typical statistical tests one computes a *p-value*.

How likely would it be to observe what I in fact did observe under the assumption that my (null)hypothesis is correct?

Example:
I recruit N = 400 individuals and randomly assign them to a treatment and a control arm in my study. I observe a well specified dichotomous outcome and want to test whether the treatment had any effect at all on that outcome.

- Null hypothesis: Frequency of the wanted (alternatively unwanted) trait is the same in both groups (meaning "no effect of treatment").
- Alternative hypothesis: Frequencies are not the same in both groups.

Example 1: Computed p-values based on observed data: 0.075

Conclusion: Likely enough to continue the work under the assumption of "treatment has no effect."

Example 2: Computed p-values based on observed data: 0.029 (< 0.05)

Conclusion: Unlikely enough – I assume I "have found something."

Example 2: Computed p-values based on observed data: 0.029 (< 0.05)

Conclusion: Unlikely enough – I assume I "have found something."

If we use that thinking and judge as "unlikely enough" everything that does not make the 5% cut (p-value < 0.05) we commit a "type I error" in 5% of the cases/trials if there is, in fact, no effect at all.

If we do 100 (independent) tests, we will erroneously declare "significance" in 5 cases. Consequence: Will be revealed in the following subsequent studies.

Remark: Since most ideas "don't work out" we expect that most null hypotheses (no effect) are actually not far from the truth and we are more concerned about the type I error than about the type II error (alternative is true but goes unnoticed).

Good examples of using p-values and how to interpret them

Like in the example above:

- N is predetermined.
- Controlled random assignment to study arms.
- Null-hypothesis follows logically from research question.

“Acceptance” or “rejection” of null-hypothesis following a pre-specified cut-off value (typically 0.05) leads to known properties in “all trials” in the scientific discoveries.

P-values are valuable to control the “performance” of my “research machinery.” (Incl. power/sample size determination)

Widely understood problems with p-values

Statistical significance and sample size

Rejection of point-hypotheses is guaranteed for large enough sample sizes.

Multiplicity in testing ☞ May 21: More on Multiple Testing with Dr. Sen

P-values and type I error rates do not correspond to each other when more than one test is conducted.

Very many tests mean BIG trouble (example follows).

Multiplicity in testing comes in many facets, among them:

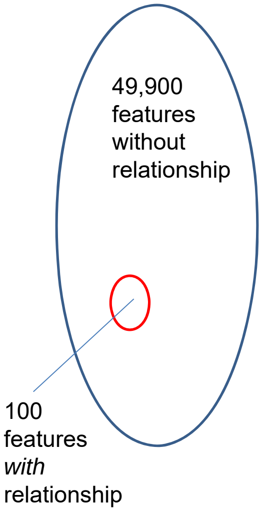
- Model building
  - Which covariates should be included?
  - Which variable-transformations should be applied?
- Determine “best thresholds” to create ordinal variable from interval variable when thresholds for grouping are not pre-specified (“grouping” should generally be avoided for other reasons but can sometimes be motivated)

Multiplicity problem: It’s a number game

If we do 50,000 (independent) tests, we will erroneously declare “significance” in about 2,500 cases (5% of 50,000).

BUT WAIT! Some real significance is hidden as well – no?

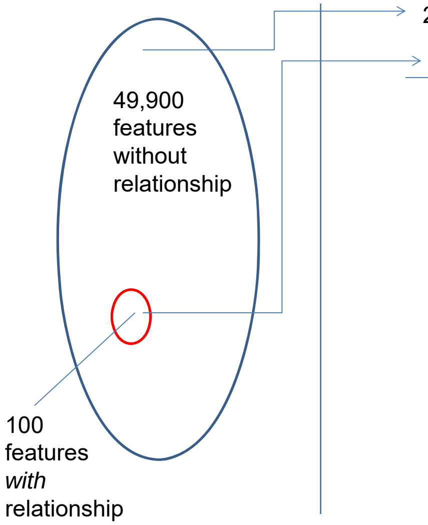
It’s a number game:



49,900 features without relationship

100 features *with* relationship

It’s a number game:



49,900 features without relationship

100 features *with* relationship

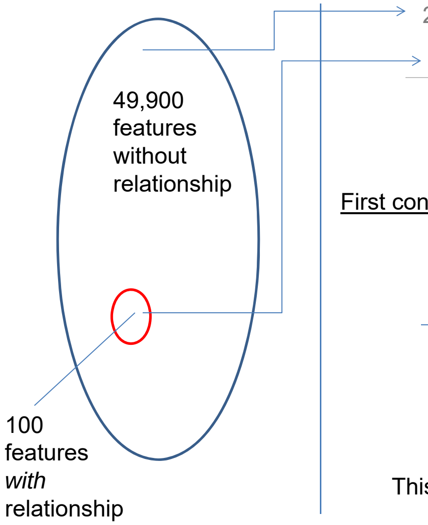
First exploratory study

2,495 false positives

100 true positives

3.8% true findings hidden within all significant results

It’s a number game:



49,900 features without relationship

100 features *with* relationship

Exploratory study

2,495 false positives

100 true positives

3.8% true findings hidden within all significant results

First confirmatory study using 2,595 potential findings

125 false positives (5% of 2,495)

100 true positives

44% true findings hidden within all significant results

This is called the *multiplicity problem*.

The number game for more detailed arrays:

2,000,000 features, 30 “true” relationships

1st	2nd	3rd	4th	
99,998	4,999	250	12	
30	30	30	30	
0.03% true significant	0.59% true significant	10.7% true significant	71% true significant	
3 years	6 years	9 years	12 years	time line

Mudding the waters I: Tables comparing demographics

Showing demographics and baseline values in two groups, such as age, gender, race, weight/BMI, household income, diabetes status etc.

Properly randomized trial with treatment and control group:
We know that all differences must be coincidental regardless what a p-value suggests. We know more than the p-value will ever tell us!

Why would we use it?

- Quality control that the randomization was conducted correctly?
- Checking whether correct randomization “worked”?

Different story:

- Demographics/anthropometrics of completers vs. dropouts.
- Demographics/anthropometrics of participants with respect to information not part of the randomization, e.g. study site.
- Demographics/anthropometrics of randomized trial participants vs. “general population” (external validity of trial).

Mudding the waters II: Observational studies

Where do the hypotheses come from?

Do I really believe that any health related outcome is COMPLETELY INDEPENDENT of age, body weight, etc.?

Would I ever believe that the population in City A has EXACTLY identical mean weight than population in City B?

Tentative answer: Statistically significant covariates are often considered “important” without further motivation (statistical significance vs. practical significance)

But:

- ➢ Do p-values measure “importance”? (No, at least not practical importance)
- ➢ Do p-values measure the support for a hypothesis? (No)

To the rescue: At least a non-significant covariate will not seriously distort a relationship if left out (maybe…)

Mudding the waters II: Observational studies (cont.)

Similar arguments apply to randomized clinical trials when more complex modeling is involved, e.g. inclusion of covariates was not predetermined.

P-values to check on achieved balance between groups? (E.g., propensity score matching for observational studies)

- Other measures should be used (e.g., standardized differences).

☞ Tip: BIOE 864 Statistical Methods for Observational Studies
(1 credit hour Fall term/ UTHSC CGHS - Master of Science in Epidemiology)

What we like about p-values

Existence:
P-value exists for all uniformly most powerful unbiased (UMPU) tests and all hypotheses *H* and observed data *X*.

Uniformly most powerful: Test with greatest power $1 - \beta$ among all tests of given size α.

Very practical and relevant when planning a randomized study. (Type 1 error)

One single value for all sorts of tests (t-test, chi-square, Wilcoxon rank sum test, Fisher’s exact test)

Interpretation:

A. As probability: obtaining data “as extreme as the observed one” in an independent replication of the experiment if null hypothesis is indeed true.
B. Connection to hypothesis testing: greatest lower bound on the set of all significance levels alpha such that we would reject *H* at level α.

Do p-values measure the support for a hypothesis?

General notation for hypotheses and p-values:

P-value: $p_{a,b}(x)$ where $a \in [-\infty, \infty)$, $b \in [a, \infty]$, and *x* is the data.

Point null:	$a = b = \mu_0$	$H_0: \mu = \mu_0$	$H_a: \mu \neq \mu_0$
One-sided:	$a = -\infty, b = \mu_0$	$H_0: \mu \leq \mu_0$	$H_a: \mu > \mu_0$
	$a = \mu_0, b = \infty$	$H_0: \mu \geq \mu_0$	$H_a: \mu < \mu_0$
Interval:	$a = \mu_1, b = \mu_2 > \mu_1$	$H_0: \mu \in [\mu_1, \mu_2]$	$H_a: \mu \notin [\mu_1, \mu_2]$

We only consider the simplest example where the test statistic’s sample distribution is a standard normal distribution.

Test statistic: function of the sample alone.

Based on Schervish MJ. P Values: What They Are and What They Are Not. The American Statistician. 1996;50(3):203-6.

Do p-values measure the support for a hypothesis? (cont.)

Example of a p-value:

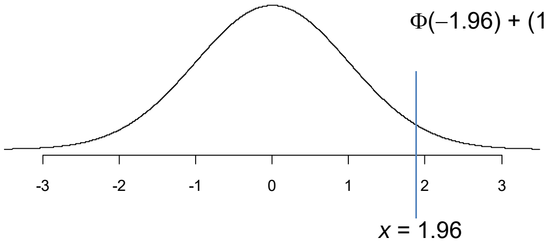
Point null: a = b = 0 $\quad H_0$: $\mu = 0 \quad H_a$: $\mu \neq 0$

$x = 1.96$

P-value

"Probability of obtaining something at least as extreme", i.e. either $x > 1.96$ or $x < -1.96$:

$\Phi(-1.96) + (1 - \Phi(1.96)) = 2 \times \Phi(-1.96) = 0.05$



Do p-values measure the support for a hypothesis? (cont.)

P-value for interval UMPU test is given by (Lehmann 1986, Schervish 1996):

$$p_{\mu1,\mu2}(x) = \Phi(x - \mu_1) + \Phi(x - \mu_2) \text{ if } x < 0.5\,[\mu_1 + \mu_2]$$
$$\text{or } = \Phi(\mu_1 - x) + \Phi(\mu_2 - x) \text{ if } x \geq 0.5\,[\mu_1 + \mu_2]$$

Check for our earlier example:
$\mu_1 = \mu_2 = 0$ and $x = 1.96 \geq 0.5\,[0 + 0] = 0$

$\Phi(\mu_1 - x) + \Phi(\mu_2 - x) = \Phi(0 - 1.96) + \Phi(0 - 1.96) = 2 \times \Phi(-1.96) = 0.05$

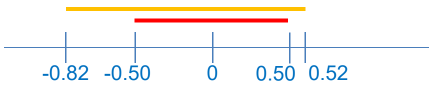
Do p-values measure the support for a hypothesis? (cont.)

What do we require from a "measure of support"?

Suppose *H* implies *H'* . Tests for *H* and *H'* are said to be *coherent* if rejection of *H'* always implies rejection of *H*. Think: *H* - The picture shows a cat; *H'* – The picture shows a mammal. *H* cannot be true if *H'* is not true.

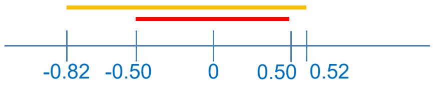
A measure of support for hypotheses *H* and *H'* should be such that whenever *H* implies *H'* then the support for *H'* is at least as large as the measure of support for *H*.

Example: For any given data and a numerical quantity to test for there must be at least as much support for the "true value" being within the wider interval (***H'***) as there is for the shorter interval (***H***) if that interval is entirely embedded in the wider one.



Do p-values measure the support for a hypothesis? (cont.)

Example 1:



Let $x = 2.18$, *H*: $\mu \in [-0.50, 0.50]$, and *H'*: $\mu \in [-0.82, 0.52]$

Since *H* implies *H'* any coherent measure of support must give at least as much support to *H'* than to *H*. A P-value would be required to be at least as large for *H'* than for *H.*

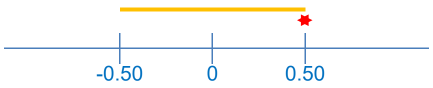
Do the math:
For *H:* $\Phi(-0.50 - 2.18) + \Phi(0.50 - 2.18) = 0.0502$
For *H':* $\Phi(-0.82 - 2.18) + \Phi(0.52 - 2.18) = 0.0498$

Consequence: If we test on 5%-level, we would reject H but accept H' and act as if our value of interest is outside the interval [–0.82, 0.52] but inside the interval [–0.50, 0.50]. This is absurd.

Conclusion: P-values are not coherent measures of support for hypotheses.

Do p-values measure the support for a hypothesis? (cont.)

Example 2:



Let $x = 2.18$, H: $\mu \in [-0.50, 0.50]$, and H': $\mu = 0.50$

Do the math:
For H: $\Phi(-0.50 - 2.18) + \Phi(0.50 - 2.18) = 0.0502$
For H': $\Phi(0.50 - 2.18) + \Phi(0.50 - 2.18) = 0.0930$

We find that the P-value for *H'* gives more support to a single point than the P-value for *H* for an interval including that point!

Again, the P-values are not coherent.

These are not rare examples! Schervish (1996) shows that whenever *x* is outside the interval of interest proper subsets can be constructed that receive higher support by the P-values than the superset. This can never be the case for coherent measures of support. P-values are not coherent.

Do p-values measure the support for a hypothesis? (cont.)

The following can be shown (Schervish 1996):

Interval P-values are incoherent.

One-sided P-values are coherent (the larger the hypothesis is the more support there is).

Also true is this:
P-values decrease as the data goes further away from the hypothesis.

But measures of support should tell us something meaningful about different hypotheses given the data (not something about the same hypothesis when different data are encountered).

P-values are not meaningful measures of support.

What we would like to have but p-values don't give us

We would like to interpret P-values as measures of support for various hypotheses. Instead, P-values tell us something about the same hypothesis when different data are encountered.

But P-values are not coherent measures of support and should not be used for that purpose.

☞ May 14: The Bayesian Approach to Data Analysis

The recent P-value controversy

We would like to interpret P-values as measures of support for various hypotheses. Instead, P-values tell us something about the same hypothesis when different data are encountered.

But P-values are not coherent measures of support and should not be used for that purpose.

☞ May 7: Should We eliminate P-Values or Use More of Them:
A Discussion on the P-Value Controversy with Dr. Sen

ASA statement on P–values (June 2016) and recent discussions

Nature (Feb 26, 2015): **Psychology journal bans *P* values -** Test for reliability of results 'too easy to pass', say editors.
http://www.nature.com/news/psychology-journal-bans-p-values-1.17001

Is statistical significance enough?

Example homeopathy:

- Old approach dating back to 1796 (Hahnemann)
- Based on the idea that "like cures like" (still popular with some?)
- Homeopathic dilution in alcohol or distilled water ("potentization")
- Interestingly, high dilutions are referred to as "higher potency"
- Regularly, the dilutions for therapeutic use are such that most doses do not contain a single molecule of the "active" ingredient. E.g., Oscillococcinum is used against influenza-like symptoms and derived from duck liver and heart, diluted to 1 part "duck" to 10^{400} parts water. (According to Wikipedia used in >50 countries and in production for over 65 years.)
- The molecules are claimed to leave an "imprint" in the dilution

Problem 1: There is no mechanism know to science that could be used by the "imprint"

Problem 2: 5 out of 100 studies will show a statistically significant difference between intervention and control groups in an absolutely correctly run randomized controlled and double blinded trial (**type 1 error**).

➢ Some argue that such "significant findings" should not be published.

➢ But: Shall we only publish what is in agreement with current knowledge?

Levels of evidence

1a	Systematic review of high quality RCTs with similar results and effect sizes for many different RCTs.
1b	Individual high quality RCT with high precision (narrow confidence interval)
1c	All or none
2a	Systematic review of <u>cohort studies</u> with similar results and effect sizes.
2b	Individual cohort study or low quality RCT (e.g., <80% follow-up)
2c	"Outcomes Research" and ecological studies (based on average exposures etc. of populations of geographical or temporal units)
3a	Systematic review of case-control studies
3b	Individual case-control study
4	Case-series and poor quality cohort and case-control studies
5	Expert opinion (unless critically appraised or based on "first principles")

Source: Oxford Centre for Evidence-based Medicine https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

All or none: Example "Bubble Boy" disease

- Babies born without functional immune system.
- SCID-X1: 1 in 50,000-100,000 affected; caused by a mutation in a gene (IL2RG)
- Most die within first year of life. (Only about 20% have access to a suitable sibling for a bone-marrow transplant as the existing cure.)

St. Jude announced April 18, 2019: Gene therapy cure for babies with X-linked severe combined immunodeficiency

"The gene therapy, produced in the Children's GMP, LLC, manufacturing facility on the St. Jude campus, involved use of a virus to transport and insert a correct copy of a gene into the genome of patients' blood stem cells. Following the treatment, the children began producing functioning immune cells for the first time, according to St. Jude, and most patients were discharged from the hospital within one month."

https://www.stjude.org/inspire/news/bubble-boy-scid-x1-cure.html

Summary

➢ There are situations where p-values are entirely adequate and should be used.

➢ Widely understood problems with p-values:
- Statistical significance and sample size
- Multiplicity in testing

➢ P-values become more difficult to interpret correctly in observational studies

➢ Gold Standard experiments will declare hogwash to be "statistically significant" in 5% of conducted experiments (assuming most optimistically that hogwash at least does not do harm…)

Thank you!

Additional Seminars in the Series:

- May 7th Should We eliminate P-Values or Use More of Them: A Discussion on the P-Value Controversy (Saunak Sen, PhD)
- May 14th The Bayesian Approach to Data Analysis (Fridtjof Thomas, PhD)
- May 21st Multiple Testing and the False Discovery Rate (Saunak Sen, PhD)
- May 28th The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- June 4th Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

BERD Clinics: https://tnctsi.uthsc.edu/