The Significance of Significance

And Other Statistical Tales

September 25, 2019

Before we start the lecture, I have some questions for you to consider…

How do we prove that something is different from another or one thing causes another?

Examples:
How do we prove that a protein is overexpressed in high grade cancer compared to low grade cancer?

How do we prove that a specific gene mutation increases risk for cystic fibrosis?

Let’s ask Sir Austin Bradford Hill (1897-1991) for some advice.

Applying Hill’s Criteria for Causation to the case of smoking and lung cancer.

1. Strength of Association. The lung cancer rate for smokers was quite a bit higher than for nonsmokers.
2. Temporality. Smoking in the vast majority of cases preceded the onset of lung cancer.
3. Consistency. Different methods (e.g., prospective and retrospective studies) produced the same result. The relationship also appeared for different kinds of people (e.g., males and females)
4. Theoretical Plausibility. Biological theory of smoking causing tissue damage which over time results in cancer in the cells was a highly plausible explanation.
5. Coherence. The conclusion (that smoking causes lung cancer) “made sense” given the current knowledge about the biology and history of the disease.
6. Specificity. Lung cancer is best predicted from the incidence of smoking.
7. Dose Response Relationship. Data showed a positive, linear relationship between the amount smoked and the incidence of lung cancer.
8. Experimental Evidence. Tar painted on laboratory rabbits’ ears was shown to produce cancer in the ear tissue over time. Hence, it was clear that carcinogens were present in tobacco tar.
9. Analogy. Induced smoking with laboratory rats showed a causal relationship. It, therefore, was not a great jump for scientists to apply this to humans.

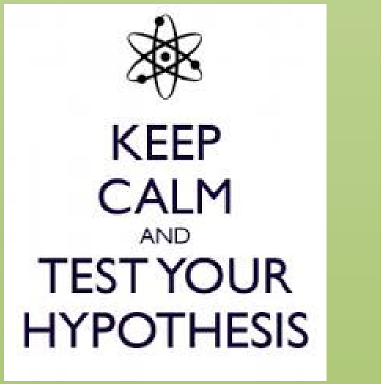
Too many researchers rely on the p-value to tell them what is important (significant) or not.

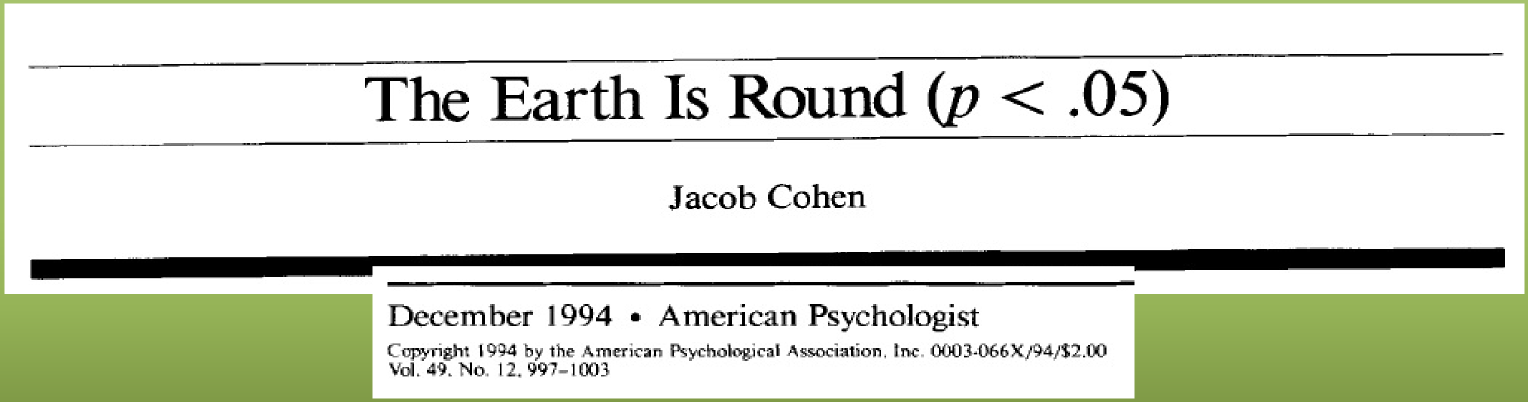
Objectives

Learn about null hypothesis significance testing (NHST)

Learn what a p-value really means

Understand there is a difference between statistical and clinical/biological significance




The Earth Is Round ($p < .05$)

Jacob Cohen

December 1994 • American Psychologist

Copyright 1994 by the American Psychological Association, Inc. 0003-066X/94/$2.00
Vol. 49, No. 12, 997–1003

What is a p-value?

In November 2015, a reporter for the website www.fivethirtyeight.com attended the inaugural METRICS conference at Stanford, which brought together some of the world's leading statistical experts on the study of studies.

She asked attendees to explain p-values in plain English in a single sentence.

http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/

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

I think this is clear as mud.

What is a p-value?

The official definition:

In statistical significance testing, the p-value is the probability of obtaining a test result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

So, what is a null hypothesis?

What is Null Hypothesis Significance Testing (NHST)?

Goal: Make statement(s) regarding unknown population parameter values based on sample data

Elements of a hypothesis test:

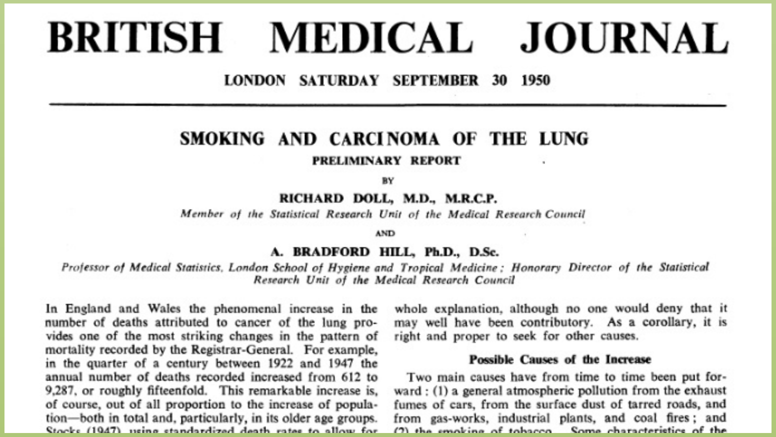
Null hypothesis - Statement regarding the value(s) of unknown parameter(s). Typically will imply no association between variables

Alternative hypothesis - Statement contradictory to the null hypothesis

Test statistic - Quantity based on sample data and null hypothesis used to test between null and alternative hypotheses

Rejection region - Values of the test statistic for which we reject the null in favor of the alternative hypothesis

Example of Null (H_0) and Alternative (H_A) Hypotheses



BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded increased from 612 to 9,287, or roughly fifteenfold. This remarkable increase is, of course, out of all proportion to the increase of population—both in total and, particularly, in its older age groups. Stocks (1947) using standardized death rates to allow for

whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

Two main causes have from time to time been put forward: (1) a general atmospheric pollution from the exhaust fumes of cars, from the surface dust of tarred roads, and from gas-works, industrial plants, and coal fires; and (2) the smoking of tobacco. Some characteristics of the

Chi-square test of proportions

H_0 : proportion of smokers in people who have lung cancer = proportion of smokers in people who do not have lung cancer

H_A : the proportions are different

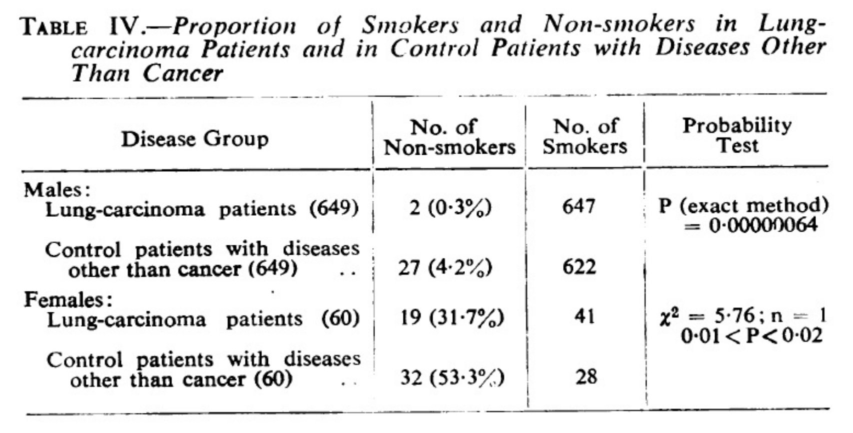


TABLE IV.—*Proportion of Smokers and Non-smokers in Lung-carcinoma Patients and in Control Patients with Diseases Other Than Cancer*

Disease Group	No. of Non-smokers	No. of Smokers	Probability Test
Males:			
Lung-carcinoma patients (649)	2 (0·3%)	647	P (exact method) = 0·00000064
Control patients with diseases other than cancer (649) ..	27 (4·2%)	622	
Females:			
Lung-carcinoma patients (60)	19 (31·7%)	41	$\chi^2 = 5{\cdot}76$; n = 1 $0{\cdot}01 < P < 0{\cdot}02$
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	

Notice I did not say the proportion of smokers in people will be greater in people with lung cancer, even though that is what we believe

What are one-sided and two-sided tests?
AKA one-tailed and two-tailed tests

One-sided H_A: smoking proportion in people without cancer is hypothesized to be lower than in people with lung cancer

H_0 *:* proportion of smokers in people who have lung cancer = proportion of smokers in people who do not have lung cancer

H_A *:* the proportion in people without cancer is lower

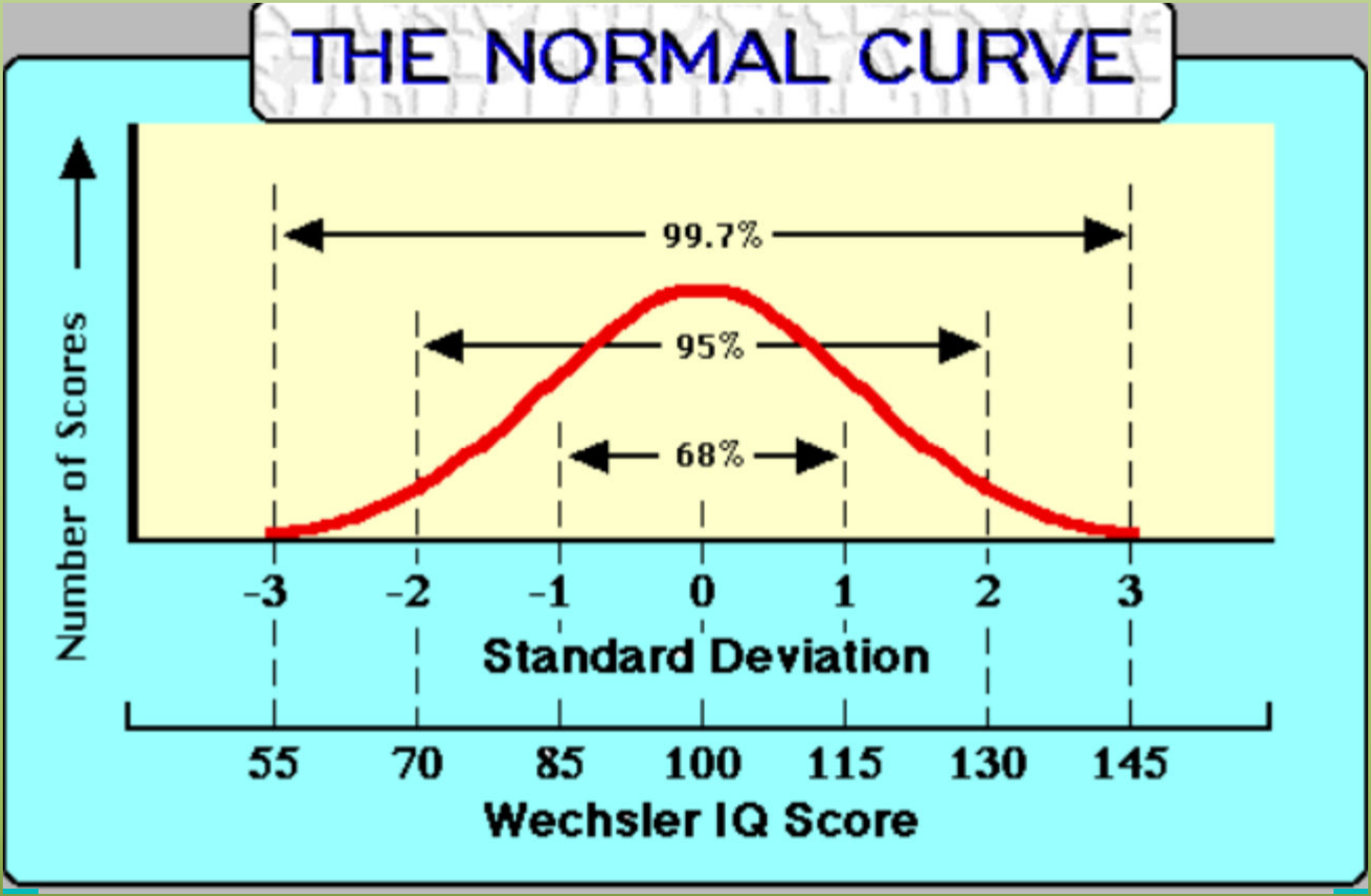
Two-sided H_A: smoking proportion in people with lung cancer can either be higher or lower than the proportion in people without lung cancer

H_0 : proportion of smokers in people who have lung cancer = proportion of smokers in people who do not have lung cancer

H_A : the proportions are different

Statisticians almost always prefer a two-sided test

Remember the normal distribution?

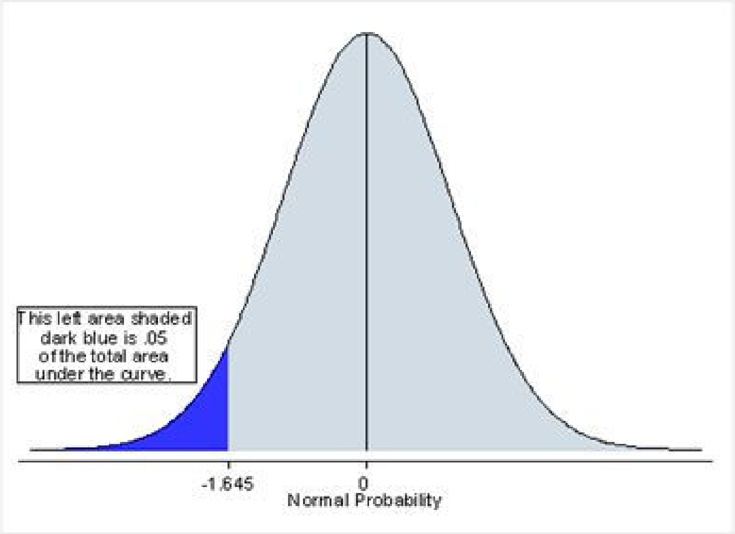


One-tailed vs two-tailed t-test

Null hypothesis: $\mu_1 = \mu_2$

One-tailed t-test

AKA One-Sided

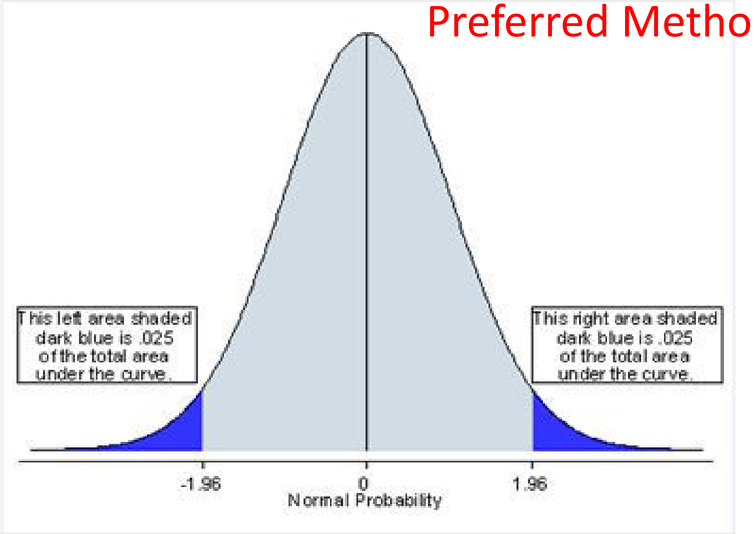


A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

Two-tailed t-test

AKA Two-Sided

Preferred Method



A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

slideplayer.com/slide/5380003

Another Example

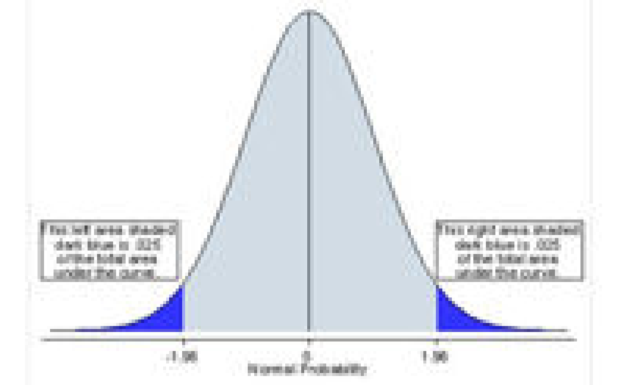
Research Hypothesis: Men who take statistics classes are taller than men who do not study statistics

H_0 : the mean height of men who take statistics classes = the mean height of men who do not study statistics

H_A : the means are different

$\mu_1 = \mu_2$

$\mu_1 - \mu_2 = 0$



$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

Two-sided

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

One-sided

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 < \mu_2$

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 > \mu_2$

Statistical Hypotheses

The two hypotheses cannot overlap

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

OK

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \geq \mu_2$

NOT OK

Example: "Body Weight"

In the 1970s, 20–29 year old men in the U.S. had a mean body weight of 170 pounds. We want to test whether mean body weight in the population now differs from 170. We believe men may have become heavier.

Null hypothesis $H_{0:}$ mean weight of men today = 170 ("no difference from what was found before")

The **alternative hypothesis** can be one of these statements:

$H_{A:}$ mean weight now > 170 (**one-sided test**) or
$H_{A:}$ mean weight now < 170 (**one-sided test**) or
$H_{A:}$ mean weight now ≠ 170 (**two-sided test**)

What is NHST? – cont.

Goal: Make statement(s) regarding unknown population parameter values based on sample data

Elements of a hypothesis test:

Null hypothesis - Statement regarding the value(s) of unknown parameter(s). Typically will imply no association between variables

Alternative hypothesis - Statement contradictory to the null hypothesis

Test statistic – A quantity calculated from a statistical test based on sample data

Rejection region - Values of the test statistic for which we reject the null in favor of the alternative hypothesis

What is a Test Statistic?

Each inferential statistical test has an underlying equation.

Example: “t” in the equation below is the test statistic for the independent t-test (testing the difference between two independent means)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = mean of sample 1
s_1 = the standard deviation of sample 1
n_1 = the sample size of sample 1
$\bar{x}_2$ = mean of sample 2
Etc.

What makes a t value larger?

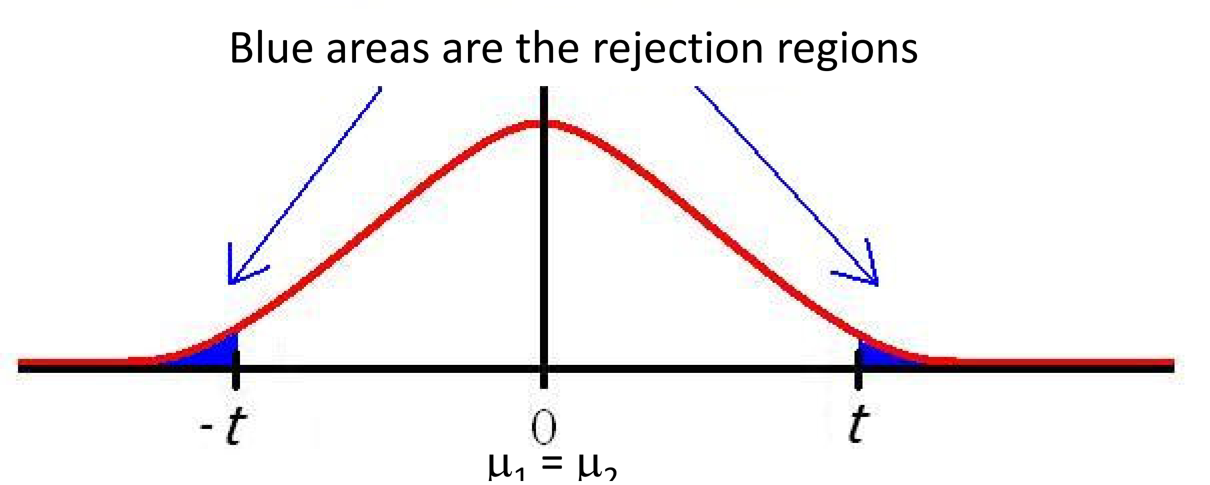
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

A bigger difference between $\bar{x}_1$ and $\bar{x}_2$ (aka Effect size)
Less variability and larger sample sizes

What is a Rejection Region?
Using the t-test t statistic as an example

“t” statistic is compared to the t distribution (a probability distribution, example below). If the “t” value you calculated is large (or small) enough that it falls in one of the blue areas (AKA rejection regions for a two sided test; reject the null hypothesis), then p≤0.05

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Blue areas are the rejection regions

-2.042

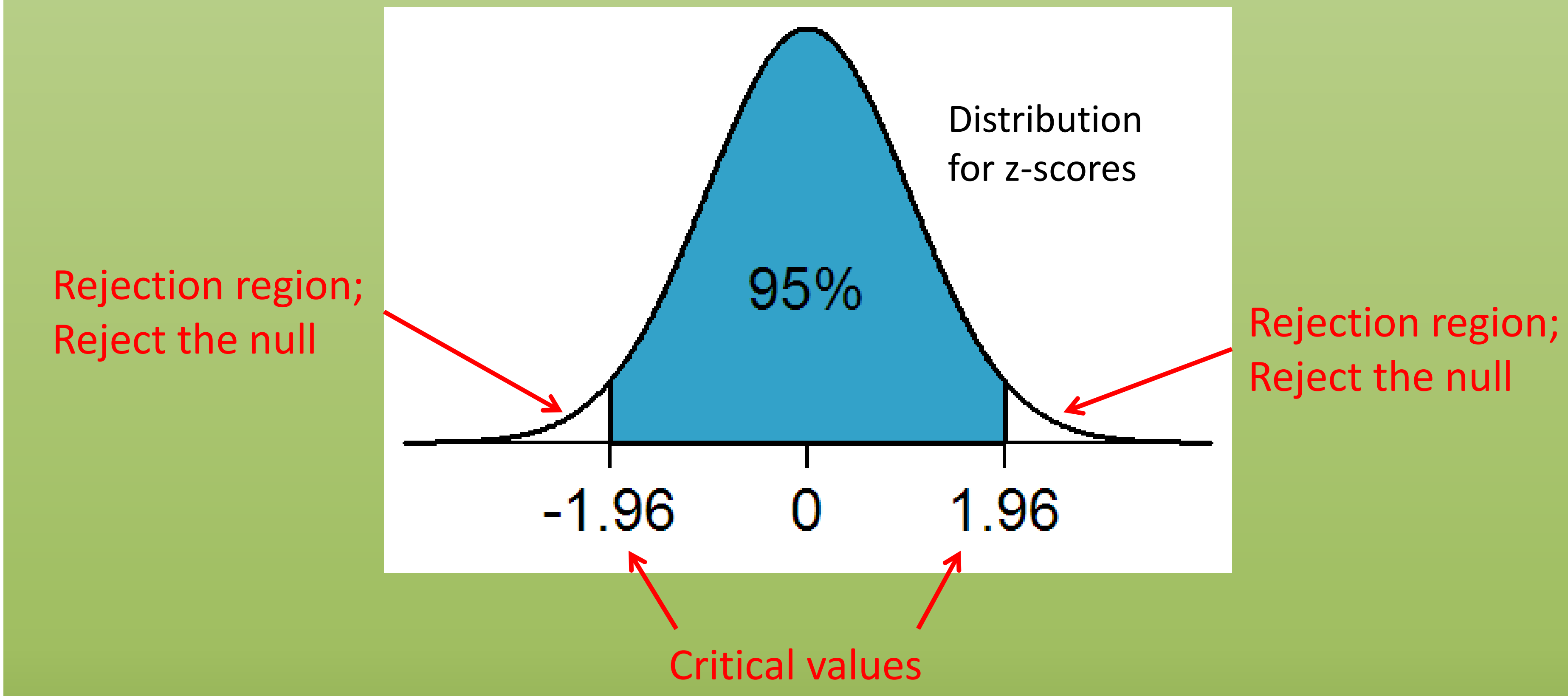
$\mu_1 - \mu_2 = 0$

2.042

These are called “critical values”
They represent a sample size of 31

Critical values for z scores (two-sided test, α=0.05) and rejection region for the null hypothesis

Fail to reject the null for values in the blue area

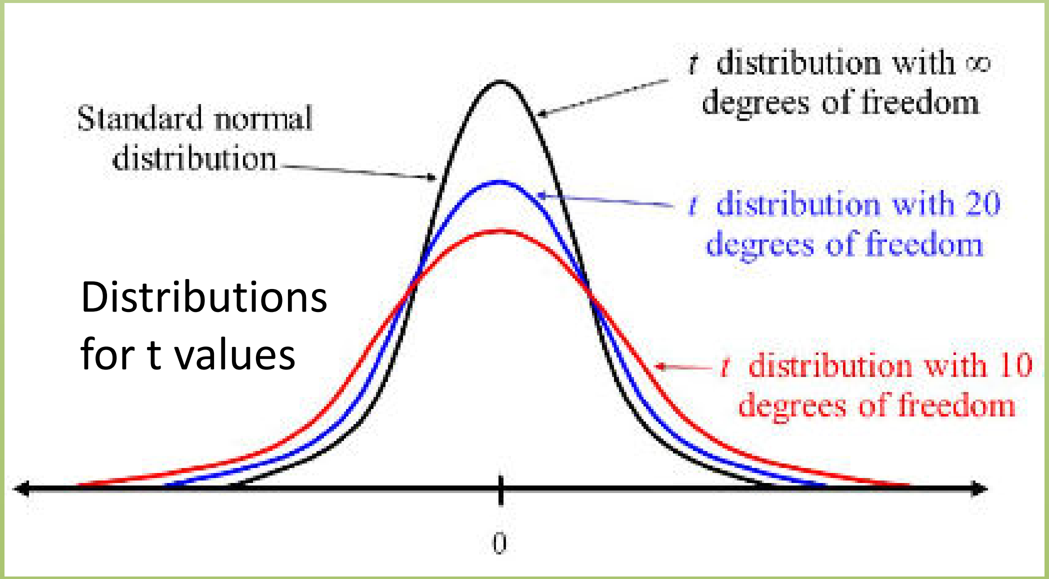


Rejection region;
Reject the null

Rejection region;
Reject the null

Critical values

The critical values change with the sample size
(Degrees of freedom = n-1)



Distributions
for t values

Two-sided t-test critical values for different sample sizes

N	DF	Critical values
11	10	-2.228, 2.228
21	20	-2.086, 2.086
∞	∞ -1	-1.960, 1.960

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

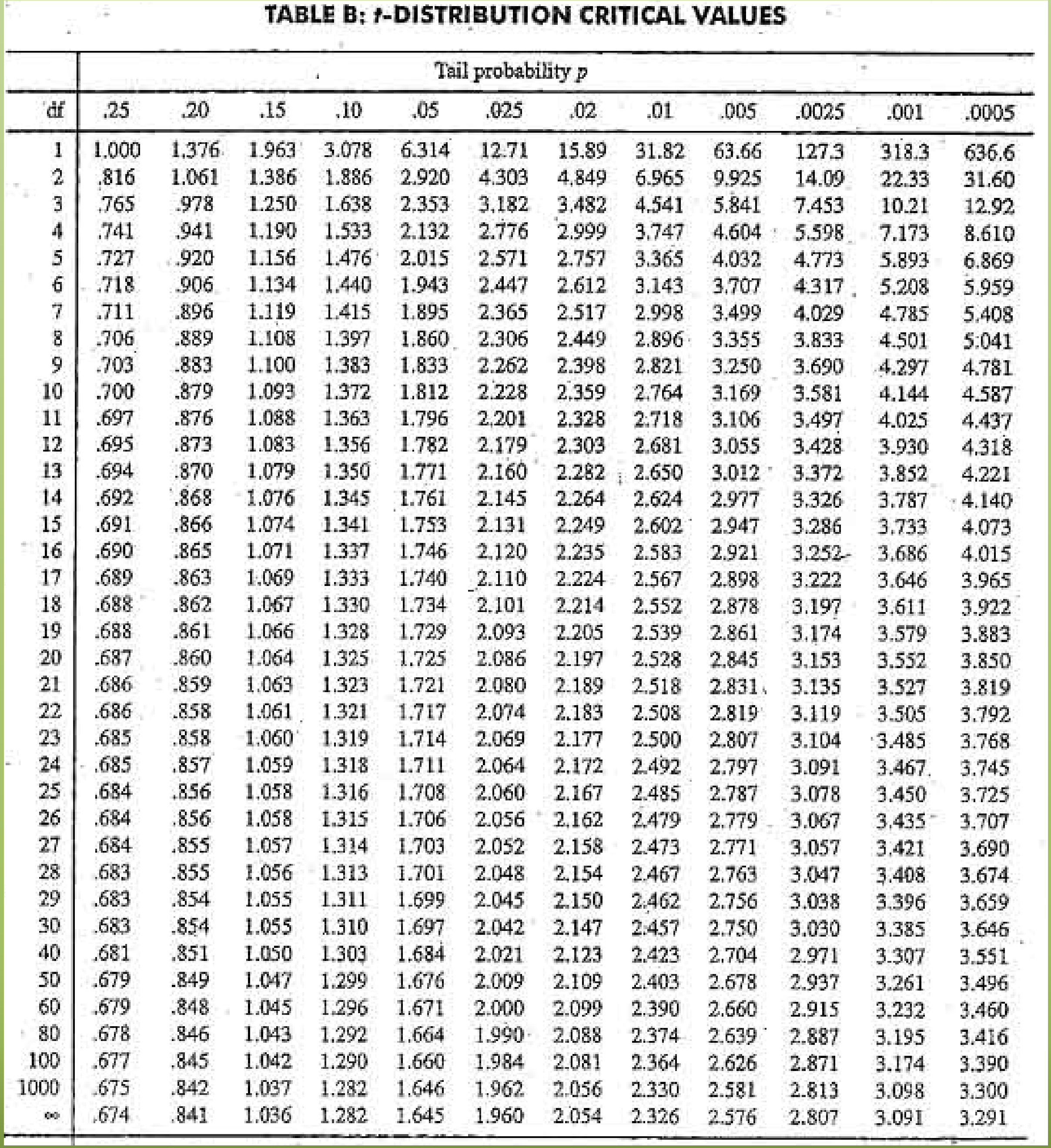


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

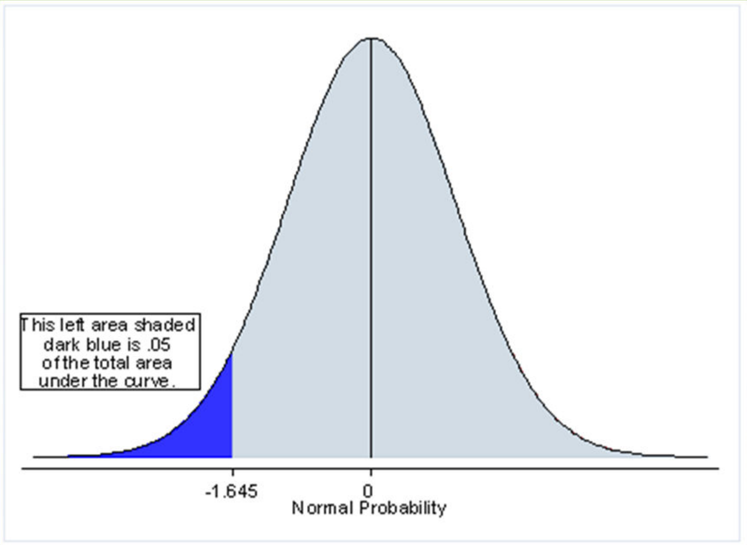
Two-sided t-test critical values for different sample sizes

N	DF	Critical values $X_1 \neq X_2$
11	10	-2.228, 2.228
21	20	-2.086, 2.086
∞	∞ -1	-1.960, 1.960

One-sided t-test critical values for different sample sizes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

		Critical values	
N	DF	$X_1 > X_2$	$X_1 < X_2$
11	10	1.812	-1.812
21	20	1.725	-1.725
∞	∞ -1	1.645	-1.645



This left area shaded dark blue is .05 of the total area under the curve.

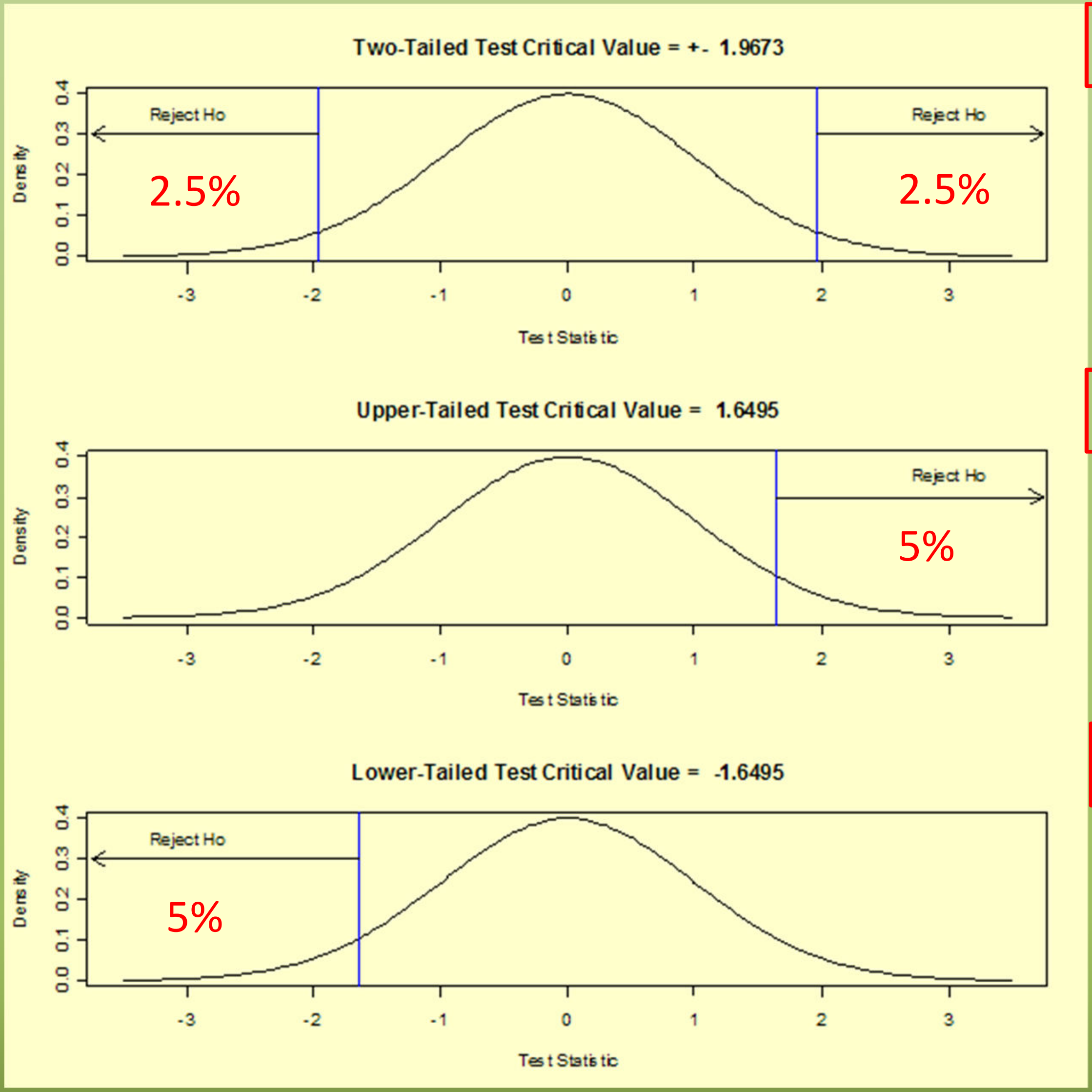
What are degrees of freedom?
…Freedom to vary, of course

Degrees of freedom are generally not something you *need* to understand to perform a statistical analysis—unless you are studying statistical theory

Degrees of freedom are often broadly defined as the number of "observations" (pieces of information) in the data that are free to vary when estimating parameters

Related to n (usually n-1)

Your t statistic is 1.8222



Two-Tailed Test Critical Value = +- 1.9673

Upper-Tailed Test Critical Value = 1.6495

Lower-Tailed Test Critical Value = -1.6495

≥1.9673 or ≤ - 1.9673

$p>0.05$

≥ 1.6495

$p\leq0.05$

≤ - 1.6495

$p>0.05$

The p-value: Reject or Fail to Reject

We set 0.05 as the threshold for significance, then

If from the analysis of our sample we find that $p \leq 0.05$

We *reject* the null hypothesis and accept the alternative

If we find $p > 0.05$

We *fail to reject* the null (the null might be true after all)

The p-value is about the null hypothesis

The conclusion about the alternative hypothesis is based on what you conclude about the null hypothesis based on the p-value

The p-value is <u>only</u> about the strength of evidence for or against the null hypothesis AND NOT ABOUT THE SIGNIFICANCE OR IMPORTANCE OF YOUR RESULTS

To reiterate:

If the statistical analysis shows that the significance level is below the cut-off value (e.g., 0.05 or 0.01), **we reject the null hypothesis** and **accept the alternative hypothesis**.

NOTE: you cannot **accept the null hypothesis**, but only find evidence against it.

> The implication with “fail” is that there was not enough evidence or large enough sample size to allow rejection of H_0

A small p-value is only evidence against the null.

Example -

You’ve done an experiment and found that your marker of interest is expressed 10% higher in people with high grade cancer compared to those with low grade.

You did a statistical test and got a p-value = 0.01.

You celebrate, thinking about the paper you will publish… the grants you will get … a promotion …

But what does that $p = 0.01$ really mean?

What does $p=0.01$ mean?

Assuming that you set statistical significance at ≤ 0.05, your $p=0.01$ means you are able to reject the null hypothesis and accept the alternative hypothesis.

A warning to my students…

I am going to insist that when you report the results from a p-value (in class, assignments, exams), you must state that you either reject (and accept the alternative) or fail to reject the null hypothesis (and conclude there are no differences – or whatever your null hypothesis is).

$P \leq 0.05$

Reject the null hypothesis and accept the alternative

$P > 0.05$

Fail to reject the null; unable to prove it is wrong

You probably will not report your results this way anywhere else outside of this class.

I am insisting on this to remind you what the p-value really means.

Hypothesis Testing and Error

Ultimately, you don't know for certain if H_0 is true or not (or the H_A)

No matter which decision you make (reject or fail to reject H_0), you could be making an error

Type I error: reject H_0 when it is true (no difference in population)

False positive

Type II error: failure to reject H_0 when it is false (difference exists)

False negative

Type I and Type II error:
Closely tied to hypothesis testing and p-values

Type I (false positive)

The probability of finding a difference and there really isn’t one (the null hypothesis is true)

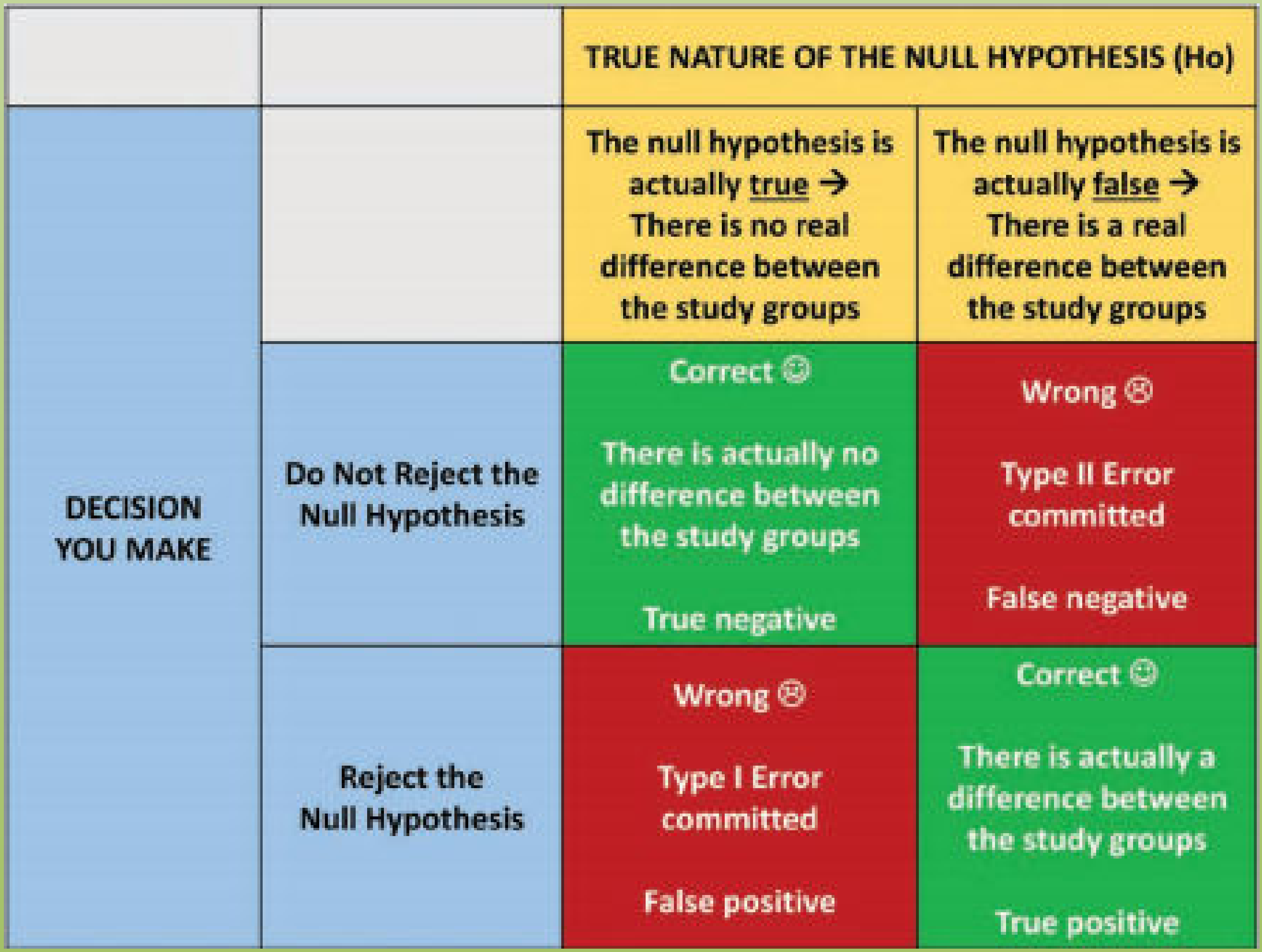
Known as the α (or probability of “type I error”)
Usually set at 5% (or 0.05)
α is set when you plan your analyses

Type II (false negative)

The probability of not finding a difference that actually exists (the null hypothesis is false)

Known as the β (or probability of a “type II error”)
Power, the probability of NOT making a Type II error, is 1- β

If we collect enough data (*i.e.*, large enough samples),
we can reduce the probabilities of both types of errors



		TRUE NATURE OF THE NULL HYPOTHESIS (Ho)	
		The null hypothesis is actually true → There is no real difference between the study groups	The null hypothesis is actually false → There is a real difference between the study groups
DECISION YOU MAKE	Do Not Reject the Null Hypothesis	Correct ☺ There is actually no difference between the study groups True negative	Wrong ☹ Type II Error committed False negative
	Reject the Null Hypothesis	Wrong ☹ Type I Error committed False positive	Correct ☺ There is actually a difference between the study groups True positive

Usually unknowable truth

$$\alpha = P(Type\ I\ \ Error) \qquad \beta = P(Type\ II\ \ Error)$$

Figure 1. Decision in a study versus truth in the population of interest. This figure shows the difference between a type I error and a type II error when making a decision to either reject or not reject the null hypothesis.

Significance, Errors, Power, and Sample Size: The Blocking and Tackling of Statistics

Edward J. Mascha, PhD,* and Thomas R. Vetter, MD, MPH†

February 2018 • Volume 126 • Number 2 www.anesthesia-analgesia.org

Avoid Confusing α and *p*-values

You set α in advance, it does not rely on your sample data.

The *p*-value is computed from the sample data.

α = prob (p-value < alpha | H_0 is true)

“The probability that the p-value from the sample data is less than alpha, given that the null hypothesis is true”

p-value = prob (getting a result equal to or more extreme than the one you observed | H_0 is true)

Avoid Confusing α and *p*-values

Alpha is the *acceptable* probability of a type I error: declaring significance when the null hypothesis is true (false positive).

If you conduct a null hypothesis test with alpha=0.05 and obtain *p*=0.003, the probability that you have made a type I error is not 0.003. It is 0.05. This holds true for all statistical tests you do on the data set.

How to Avoid Type I and Type II errors

Type I:
Set a lower alpha level (*i.e.*, 0.01 and not 0.05)

Type II:
Increase the sample size

> To have a good chance of rejecting the null when it is false

Of course reducing bias in study design, sample selection, and data collection and analysis will also reduce the probability of making errors.

Once Again, the Steps of NHST

1. Set up a null hypothesis and an alternative hypothesis appropriate to your test, i.e.,

   H_0 = no difference in expression levels between 2 groups

   H_A = there is a difference

2. Calculate a *p*-value (e.g., using a t-test)

3. Reject the null if a threshold is passed ($p \leq 0.05$) and accept the alternative hypothesis. If the threshold is not crossed, you fail to reject the null hypothesis and the hypothesis of no difference holds.

The question you may ask, why is the threshold 0.05?

Evolution of the p-value

Ronald Fisher
(1890-1962)

UK statistician Ronald Fisher introduced the p-value in the 1920s

- A way to judge whether evidence was significant
- Set up a “null hypothesis” to disprove
  (null hypothesis is usually there is no difference)
- Calculate the probability (p) of getting results at least as extreme as what was actually observed (assuming null is true)
- The smaller the p-value, the greater the likelihood that the straw-man null hypothesis is false.

Fisher intended the p-value to be just one part of a process that blended data, observation, and background knowledge to lead to scientific conclusions.

Evolved into null hypothesis significance testing (NHST)

Why $p<0.05$?

In *Statistical Methods for Research Workers* (1925), Ronald Fisher suggests the value for p should be 0.05, or 1 in 20.

In 1926, he wrote, “if one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.”

And there you have it. With no theoretical justification, these few sentences drove the standard significance level of 0.05 that we use to this day.

The Meaning of “Significant” According to Fisher

Fisher proposed the use of the term “significant” to be attached to small p-values. He meant it was *something worthy of notice*. He wrote in his book, *Statistical Methods for Research Workers* about the 5% level:

“A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.”

He thought that the meaning of a $p \leq 0.05$ was that the *experiment should be repeated* to determine if similar low p-values would be found with repeated experiments.

If subsequent experiments also yielded the same results and significant p-values, it could be concluded there was strong evidence that the observed differences were real.

So “significance” meant only “worthy of attention” in the form of doing more experimentation to add to the evidence of a potentially significant difference, but a single $p \leq 0.05$ was not proof in itself.

Evolution of the 0.05 Threshold

Why $p \leq 0.05$?

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an American Statistical Association (ASA) discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Ronald L. Wasserstein & Nicole A. Lazar (2016): The ASA's
statement on p-values: context, process, and purpose, The American Statistician, DOI:
10.1080/00031305.2016.1154108

Misstatements about the p-value

“Our findings were even more important because the p-value was <0.0001.”

“Our findings showed that drug X was more effective that drug Z because the p-value for drug X was smaller.”

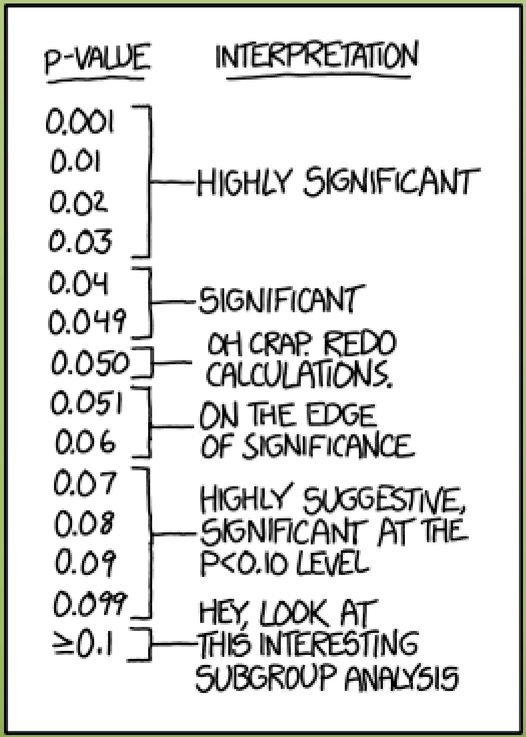
Wrong in both cases.

A p<0.0001 only says that you have stronger evidence to reject the null hypothesis. This level of significance may simply be due to the fact you have a large sample size. BUT, the difference in measures of effect you found may not be clinically or biologically relevant.

A smaller p-value only means you have more confidence you can reject the null hypothesis

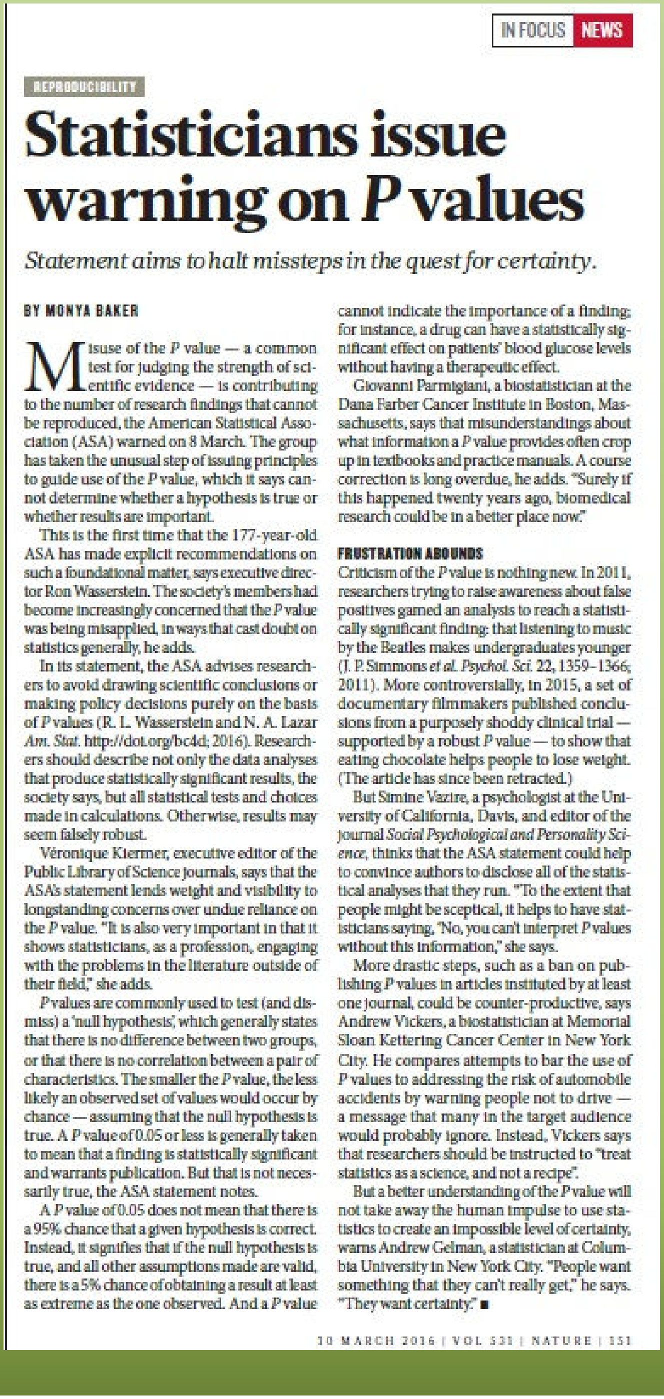
A list of descriptions to use when your p-value is >0.05 culled from peer-reviewed journal articles in which (a) the authors set themselves the threshold of 0.05 for significance, (b) failed to achieve that threshold value for p and (c) described it in such a way as to make it seem more interesting

all but significant (p=0.055)
almost but not quite significant (p=0.06)
approaches but fails to achieve a customary level of statistical significance (p=0.154)
at the brink of significance (p=0.06)
weakly statistically significant (p=0.0557)
well-nigh significant (p=0.11)
uncertain significance (p>0.07)
vaguely significant (p>0.2)
suggestive of statistical significance (p=0.06)
roughly significant (p>0.1)
probably significant (p=0.06)
on the cusp of significance (p=0.058)
not remarkably significant (p=0.236)
not insignificant (p=0.056)
not currently significant (p=0.06)
mildly significant (p=0.07)
leaning towards significance (p=0.15)



P-VALUE	INTERPRETATION
0.001 0.01 0.02 0.03	HIGHLY SIGNIFICANT
0.04 0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051 0.06	ON THE EDGE OF SIGNIFICANCE
0.07 0.08 0.09 0.099	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

mchankins.wordpress.com/2013/04/21
/still-not-significant-2/



IN FOCUS NEWS

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions purely on the basis of *P* values (R. L. Wasserstein and N. A. Lazar *Am. Stat.* http://doi.org/bc4d; 2016). Researchers should describe not only the data analyses that produce statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

Véronique Kiermer, executive editor of the Public Library of Science journals, says that the ASA's statement lends weight and visibility to longstanding concerns over undue reliance on the *P* value. "It is also very important in that it shows statisticians, as a profession, engaging with the problems in the literature outside of their field," she adds.

P values are commonly used to test (and dismiss) a 'null hypothesis', which generally states that there is no difference between two groups, or that there is no correlation between a pair of characteristics. The smaller the *P* value, the less likely an observed set of values would occur by chance — assuming that the null hypothesis is true. A *P* value of 0.05 or less is generally taken to mean that a finding is statistically significant and warrants publication. But that is not necessarily true, the ASA statement notes.

A *P* value of 0.05 does not mean that there is a 95% chance that a given hypothesis is correct. Instead, it signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed. And a *P* value cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

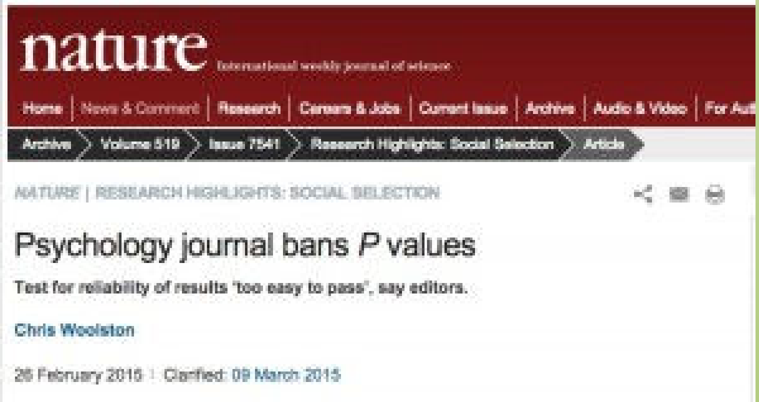
FRUSTRATION ABOUNDS

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gamed an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* 22, 1359–1366; 2011). More controversially, in 2015, a set of documentary filmmakers published conclusions from a purposely shoddy clinical trial — supported by a robust *P* value — to show that eating chocolate helps people to lose weight. (The article has since been retracted.)

But Simine Vazire, a psychologist at the University of California, Davis, and editor of the journal *Social Psychological and Personality Science*, thinks that the ASA statement could help to convince authors to disclose all of the statistical analyses that they run. "To the extent that people might be sceptical, it helps to have statisticians saying, 'No, you can't interpret *P* values without this information,'" she says.

More drastic steps, such as a ban on publishing *P* values in articles instituted by at least one journal, could be counter-productive, says Andrew Vickers, a biostatistician at Memorial Sloan Kettering Cancer Center in New York City. He compares attempts to bar the use of *P* values to addressing the risk of automobile accidents by warning people not to drive — a message that many in the target audience would probably ignore. Instead, Vickers says that researchers should be instructed to "treat statistics as a science, and not a recipe".

But a better understanding of the *P* value will not take away the human impulse to use statistics to create an impossible level of certainty, warns Andrew Gelman, a statistician at Columbia University in New York City. "People want something that they can't really get," he says. "They want certainty." ■


nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au
Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection > Article

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

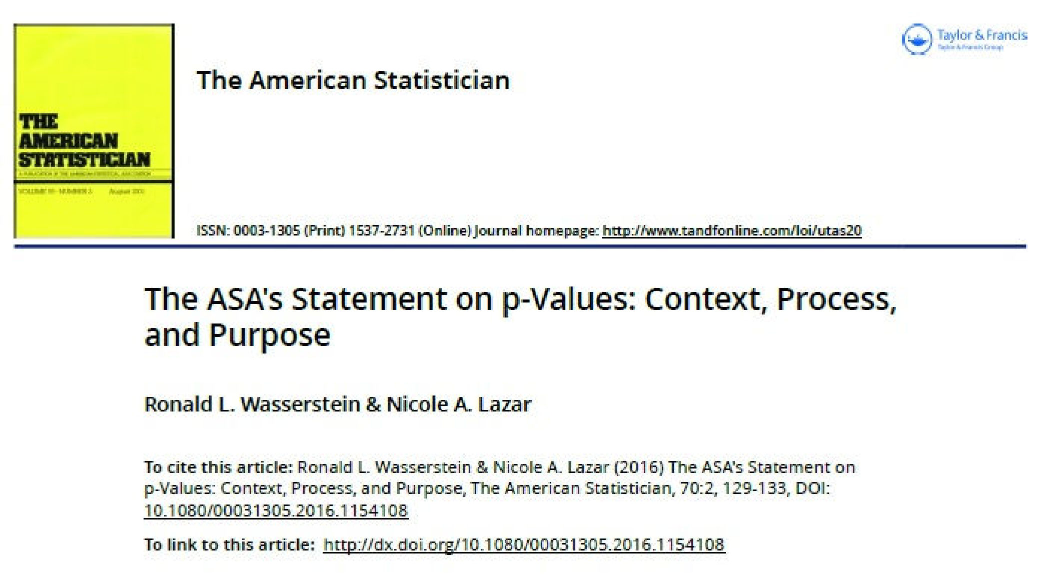
26 February 2015 | Clarified: 09 March 2015

Some are saying we should abandon the p-value and null-hypothesis significance testing (NHST) all together.

Problems with the p-value

In October 2015, the American Statistical Association (ASA) gathered experts to develop a consensus statement on statistical significance and p-values.

The consensus project was spurred by a growing worry that in some scientific fields, p-values have become a litmus test for deciding which studies are worthy of publication. As a result, research that produces p-values that surpass an arbitrary threshold are more likely to be published, while studies with greater or equal scientific importance may remain in the file drawer, unseen by the scientific community.



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1154108

This is the first time that the ASA has made explicit recommendations on such a foundational matter in statistics.

There was increasing concern that the *p* value was being misapplied in ways that cast doubt on statistics generally.

“Practices that reduce data analysis or scientific inference to mechanical ‘bright-line’ rules (such as ‘$p \leq 0.05$’) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become ‘true’ on one side of the divide and ‘false’ on the other.”

ASA statement of use of p-values, 2016

It would not make sense to reach different conclusions about two studies on the same topic if one had p=0.49 and one had p=0.51.

The takeaway from the ASA statement:

p-values are not badges of truth and $p \leq 0.05$ is not a line that separates real results from false ones.

P-values are imply one piece of a puzzle that should be considered in the context of other evidence.

That is what Fisher said almost 100 years ago…

Making a Stat Less Significant

Wall Street Journal 4/2/11

SUPREME COURT OF THE UNITED STATES
MATRIXX INITIATIVES, INC., et al. *v* . SIRACUSANO et al.
No. 09–1156. Argued January 10, 2011—Decided March 22, 2011

A group of mathematicians has been trying for years to have a core statistical concept debunked. Now, the Supreme Court might have done it for them.

Last month, the court considered a case brought by investors in Matrixx Initiatives Inc. They alleged the company failed to disclose material information by neglecting to reveal it had received reports an over-the-counter medicine, Zicam Cold Remedy, caused a loss of sense of smell.

When those reports came to light, the company's stock fell. Eventually, the Food and Drug Administration warned consumers not to use certain Zicam products. The company argued in court that the initial reports of the possible side effect didn't rise to the level of "statistical significance," and therefore didn't need to be disclosed.

The Court rejected that argument, ruling unanimously that the case against Matrixx pending in a lower court could proceed. In their opinion, the justices said companies *can't only rely on statistical significance when deciding what they need to disclose to investors*.

If not p-values, then what?

The 20 commentaries published with the ASA statement present a range of ideas about where to go from here. Some committee members argued that there should be a move to rely more on other measures, such as confidence intervals or Bayesian analyses. Others felt that switching to something else would only shift the problem around.

“The solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation.”

In the meantime, we will continue to talk about (and use) p-values and confidence intervals help (more on this later)

One Pitfall of Statistical Analyses in Published Papers

Reporting only *p* values.

The mean, median, SD, confidence interval, relative risk, or odds ratio, etc. should be reported, to allow the reader to critique for him/herself the validity of the results.

Look at magnitudes of the measures of effect rather than just *p* values.

The measures of effect are what will help determine clinical/biological significance

Int J Emerg Med. Sep 2008; 1(3): 197–199.

Measures of Effect/Effect Size

Effect size is a quantitative measure of the strength of a phenomenon
Effect size shows the size of the difference

Examples

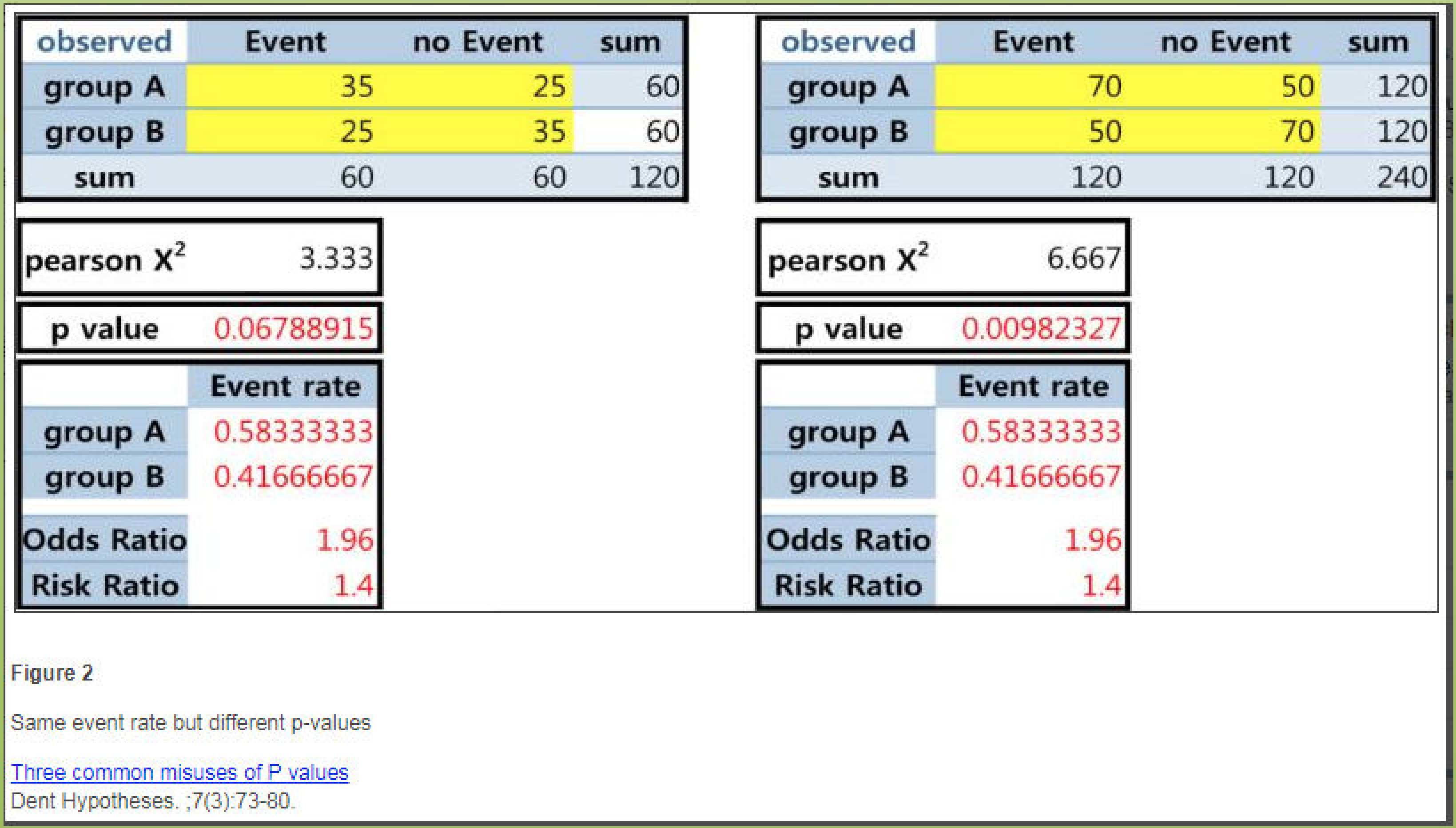
correlation coefficient (correlation between variables)
means or proportions and how they differ between groups
regression coefficient in a regression model
survival proportions
odds ratios, hazard ratios, relative risks

Statistical Significance is not Necessarily Clinical/Biological Significance

The effect of increasing sample size on detectable differences with equal SD

Sample Size	Sample 1 Mean	Sample 2 Mean	p
4	100.0	110.0	0.05
25	100.0	104.0	0.05
64	100.0	102.5	0.05
400	100.0	101.0	0.05
2,500	100.0	100.4	0.05
10,000	100.0	100.2	0.05

Increasing sample size increases importance?



observed	Event	no Event	sum
group A	35	25	60
group B	25	35	60
sum	60	60	120

pearson X^2	3.333
p value	0.06788915

	Event rate
group A	0.58333333
group B	0.41666667
Odds Ratio	1.96
Risk Ratio	1.4

observed	Event	no Event	sum
group A	70	50	120
group B	50	70	120
sum	120	120	240

pearson X^2	6.667
p value	0.00982327

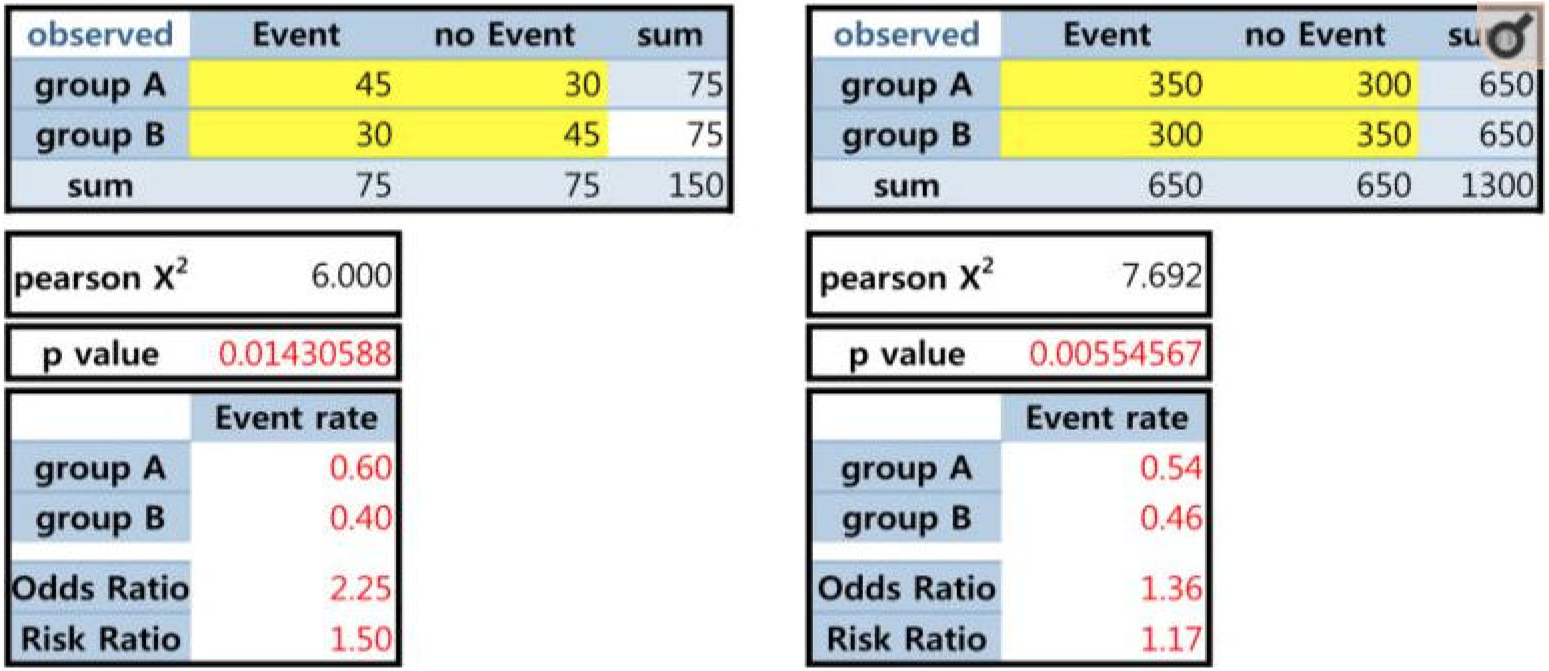
	Event rate
group A	0.58333333
group B	0.41666667
Odds Ratio	1.96
Risk Ratio	1.4

Figure 2

Same event rate but different p-values

Three common misuses of P values
Dent Hypotheses. ;7(3):73-80.

Increasing sample size increases importance?



observed	Event	no Event	sum
group A	45	30	75
group B	30	45	75
sum	75	75	150

pearson X^2	6.000
p value	0.01430588

	Event rate
group A	0.60
group B	0.40
Odds Ratio	2.25
Risk Ratio	1.50

observed	Event	no Event	su
group A	350	300	650
group B	300	350	650
sum	650	650	1300

pearson X^2	7.692
p value	0.00554567

	Event rate
group A	0.54
group B	0.46
Odds Ratio	1.36
Risk Ratio	1.17

Smaller effect size with smaller p-value

I can “prove” that the difference between mean height of 70.0 and 69.9 inches are significantly different (p=0.04)

	Change	Group 1			Group 2			t-test
Total n	in number	n	mean	SD	n	mean	SD	p
1017		903	70.0	2.57	114	69.9	2.66	0.53
2034	2X	1806	70.0	2.57	228	69.9	2.66	0.38
4068	4X	3612	70.0	2.57	456	69.9	2.66	0.21
8135	8X	7223	70.0	2.57	912	69.9	2.66	0.08
11183	11X	9929	70.0	2.57	1254	69.9	2.66	0.04

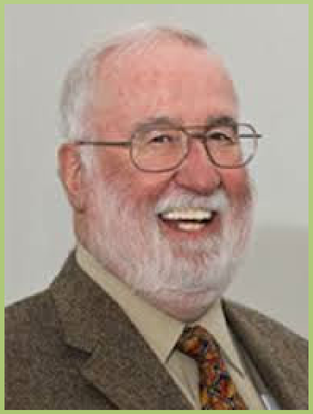
But is the result biologically significant?

If I just saw p=0.04, I would say the difference in mean height is significant. But once I see the difference is only 0.1 inch, I don’t think this difference is meaningful. I wonder why you wasted so much time and resources measuring over 11,000 men.

The p-value

"*Too small a sample and you can prove nothing.*
Too large and you can prove anything."

David Sackett



P-value Conclusions

P-values are one of many tools we have as data analysts and scientists to conduct our research

As the editors of Epidemiology pointed out in 2001: “The question is not whether the p value is intrinsically bad, but whether it is too easily substituted for the thoughtful integration of evidence and reasoning."

You need to utilize p-values in conjunction with other evidence

Remember what a p-value really means!