

TESTING

A clearly is the better team – or?



HYPOTHESIS TESTING

Formulate your theory “in a testable way”

- Null Hypothesis
- Alternative Hypothesis

Identify your test

- Test statistics
- One vs two-sided

Identify how certain you want to be

- Level of Significance

Do the math

WHAT IS A (TESTABLE) HYPOTHESIS?

A hypothesis is a claim (assumption) about a population parameter:

- population mean

Example: The mean monthly cell phone bill of this city is $\mu = \$42$

- population proportion

Example: The proportion of adults in this city with cell phones is $p = .68$

NULL VS. ALTERNATIVE HYPOTHESIS

The FDA or “science” needs to decide on a new theory, drug, treatment...

H_0 : The null hypothesis - the current theory, drug, treatment, is as good or better

H_1 : The alternative hypothesis - the new theory, drug, treatment, should replace the old one

Researchers do not know which hypothesis is true. They must make a decision on the basis of evidence presented.

THE NULL HYPOTHESIS, H_0

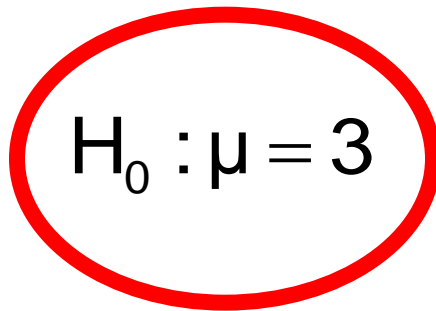
Usually refers to a general statement or default position that there is no relationship or no difference

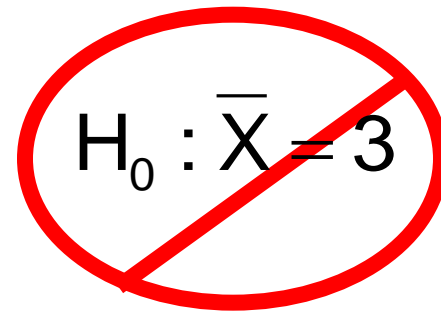
THE NULL HYPOTHESIS, H_0

States the assumption (numerical) to be tested

Example: The average number of TV sets in U.S. Homes is equal to three ($H_0 : \mu = 3$)

Is always about a population parameter, not about a sample statistic


$$H_0 : \mu = 3$$


$$H_0 : \bar{X} = 3$$

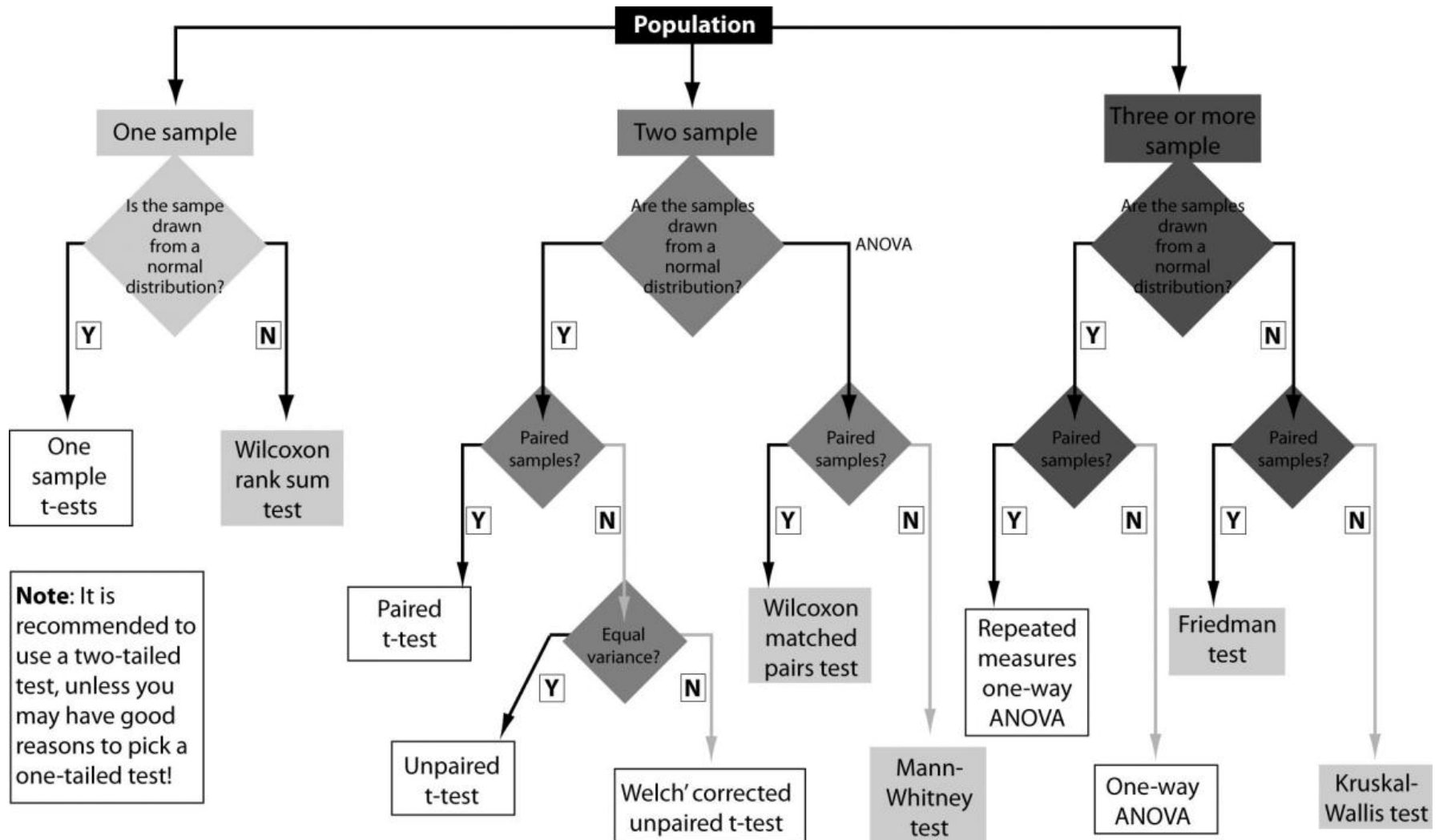
SELECT YOUR TEST

Testing is a bit like finding the right recipe based on these ingredients:

- Question
- Data type
- Sample size
- Variance known? Variance of several groups equal?

Good news: Plenty of tables available, e.g.,

- http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm (with examples in R, SAS, Stata, SPSS)
- http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf



Example of a table of tests

Summary Table for Statistical Techniques

	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
1	Estimating a Mean	One Population Mean μ	Sample mean \bar{y}	Numerical	<ul style="list-style-type: none"> What is the average weight of adults? What is the average cholesterol level of adult females? 	1-sample t-interval $\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	Stat >Basic statistics >1-sample t	<ul style="list-style-type: none"> data approximately normal or have a large sample size ($n \geq 30$)
2	Test about a Mean	One Population Mean μ	Sample mean \bar{y}	Numerical	<ul style="list-style-type: none"> Is the average GPA of juniors at Penn State higher than 3.0? Is the average Winter temperature in State College less than 42° F? 	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$ or $H_a: \mu > \mu_0$ or $H_a: \mu < \mu_0$ The one sample t test: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$	Stat >Basic statistics >1-sample t	<ul style="list-style-type: none"> data approximately normal or have a large sample size ($n \geq 30$)
3	Estimating a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	<ul style="list-style-type: none"> What is the proportion of males in the world? What is the proportion of students that smoke? 	1-proportion Z-interval $\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	Stat >Basic statistics >1-sample proportion	<ul style="list-style-type: none"> have at least 5 in each category
4	Test about a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	<ul style="list-style-type: none"> Is the proportion of females different from 0.5? Is the proportion of students who fail Stat 500 less than 0.1? 	$H_0: \pi = \pi_0$ $H_a: \pi \neq \pi_0$ or $H_a: \pi > \pi_0$ or $H_a: \pi < \pi_0$ The one proportion Z-test: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Stat >Basic statistics >1-sample proportion	<ul style="list-style-type: none"> $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$

A common test: One-sample t-test

- **When:** Estimating a mean, comparing mean to a hypothetical value
- **Requirements:** Data approx. normal or sample size > 30

- **Setup:**

$\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}$	$\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}$	$\begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}$
$T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \sim t_{n-1}$		
$t = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$		

- **Evaluation:** Compare t-statistic (table, excel) to your value and accept / reject null hypothesis

example

- Manufacturer claims battery lasts for 3.5 hrs
 $\bar{x} = 3.25, s = 0.31, n = 10$

- Our sample data

$$H_0 : m \geq 3.5, H_1 : m < 3.5$$

- Hypothesis: Life time is less than 3.5 hrs

$$t = \sqrt{n} \frac{\bar{x} - m}{s} \approx -2.55$$

cum. prob one-tail two-tails	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681</									

cum. prob one-tail two-tails	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

example

- Manufacturer claims battery lasts for 3.5 hrs
 $\bar{x} = 3.25, s = 0.31, n = 10$

- Our sample data

$$H_0 : m \geq 3.5, H_1 : m < 3.5$$

- Hypothesis: Life time is less than 3.5 hrs

$$t = \sqrt{n} \frac{\bar{x} - m}{s} \approx -2.55$$

- t-statistic: $t_{crit}(0.05, 9) = -1.833$

- Critical t-value at 0.05 level of significance:

$t < t_{crit}$: Null hypothesis rejected! Battery drains faster!

An Alternative Way

Beer Consumption XXXXXXXXXX Human Attractiveness to Malaria Mosquitoes

Beer (25):

27 20 21 26 27 31 24 21 20 19
23 24 28 19 24 29 18 20 17 31
20 25 28 21 27

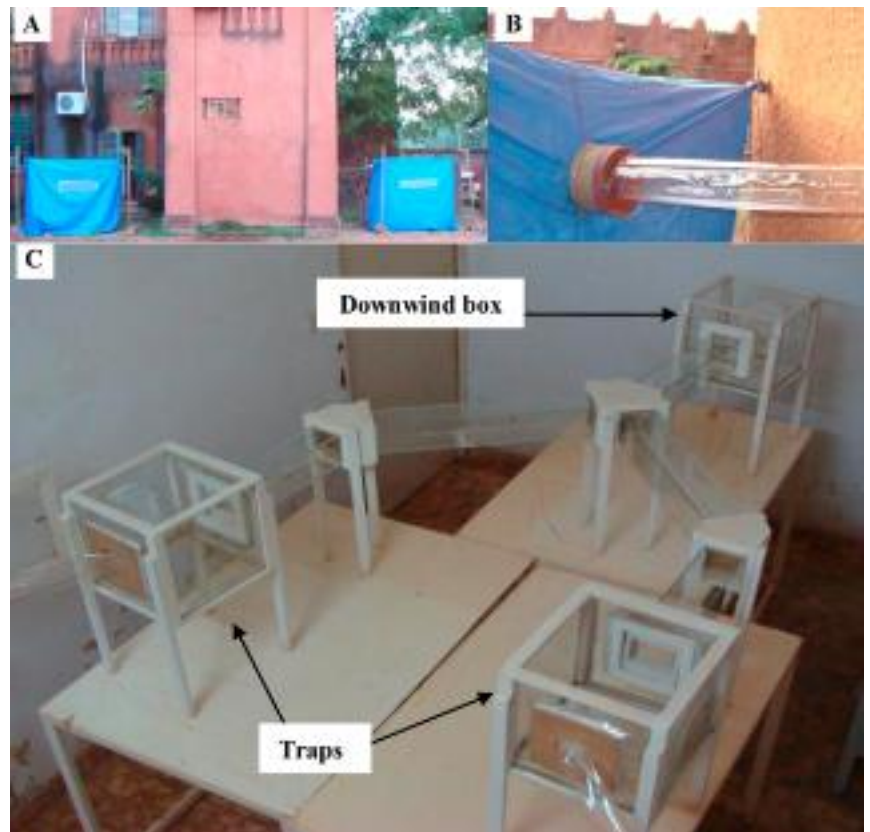
Mean: 23.6

Water (18):

21 22 15 12 21 16 19 15 22 24
19 23 13 22 20 24 18 20

Mean: 19.2

Is a difference of 4.4 significant?



Permutation Test

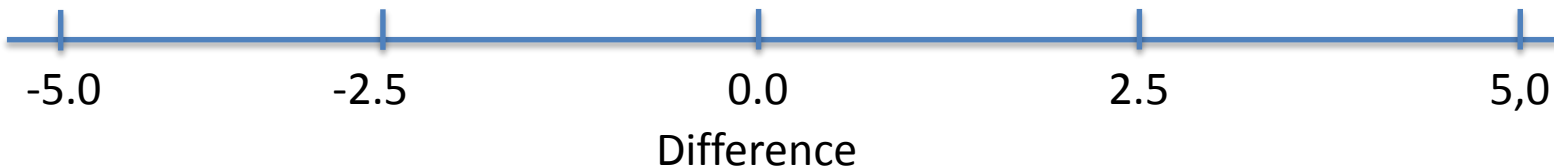
Beer (25)

27	23	20	31	29
20	24	25	24	18
21	28	28	21	20
26	19	21	20	17
27	24	27	19	31

Water (18)

21	19	16	24
22	23	19	18
15	13	15	20
12	22	22	
21	20	24	

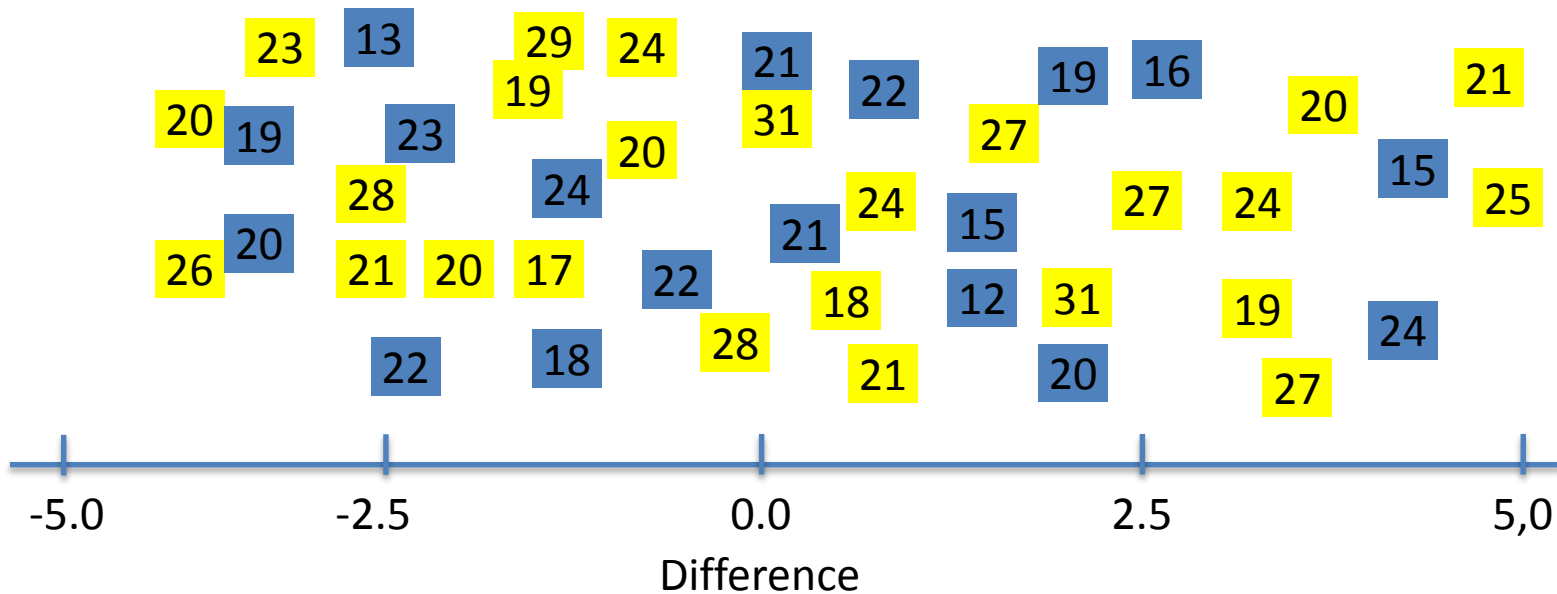
Difference: 4.4



Permutation Test

Beer (25)

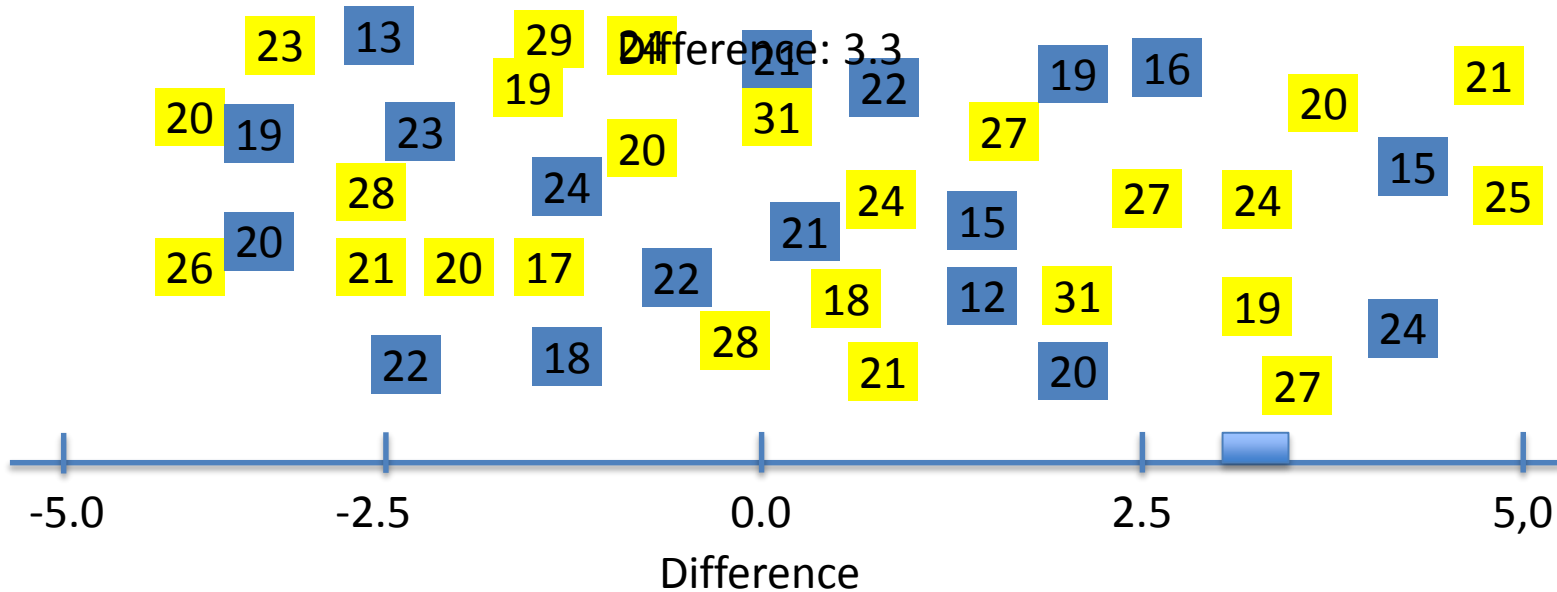
Water (18)



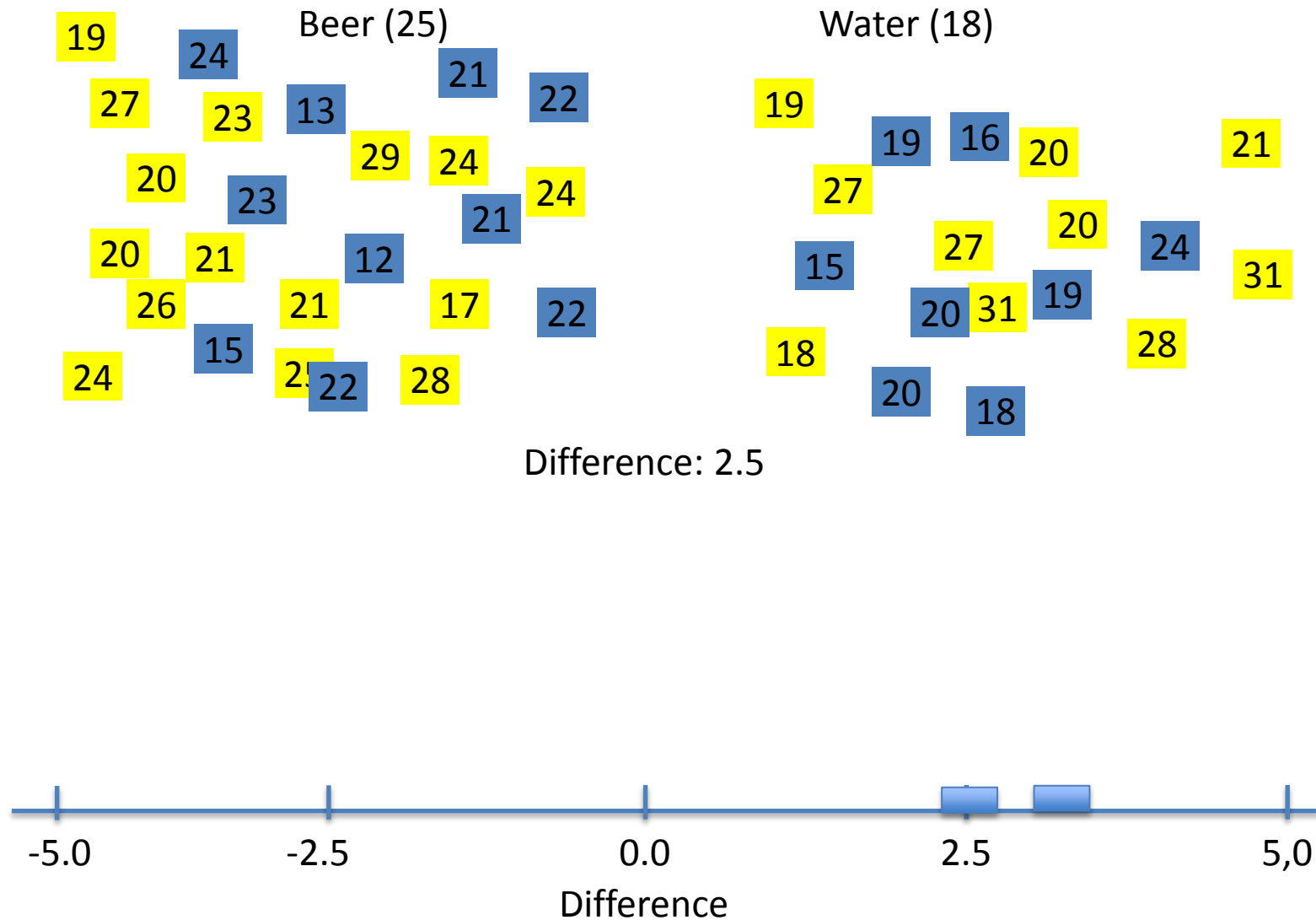
Permutation Test

Beer (25)

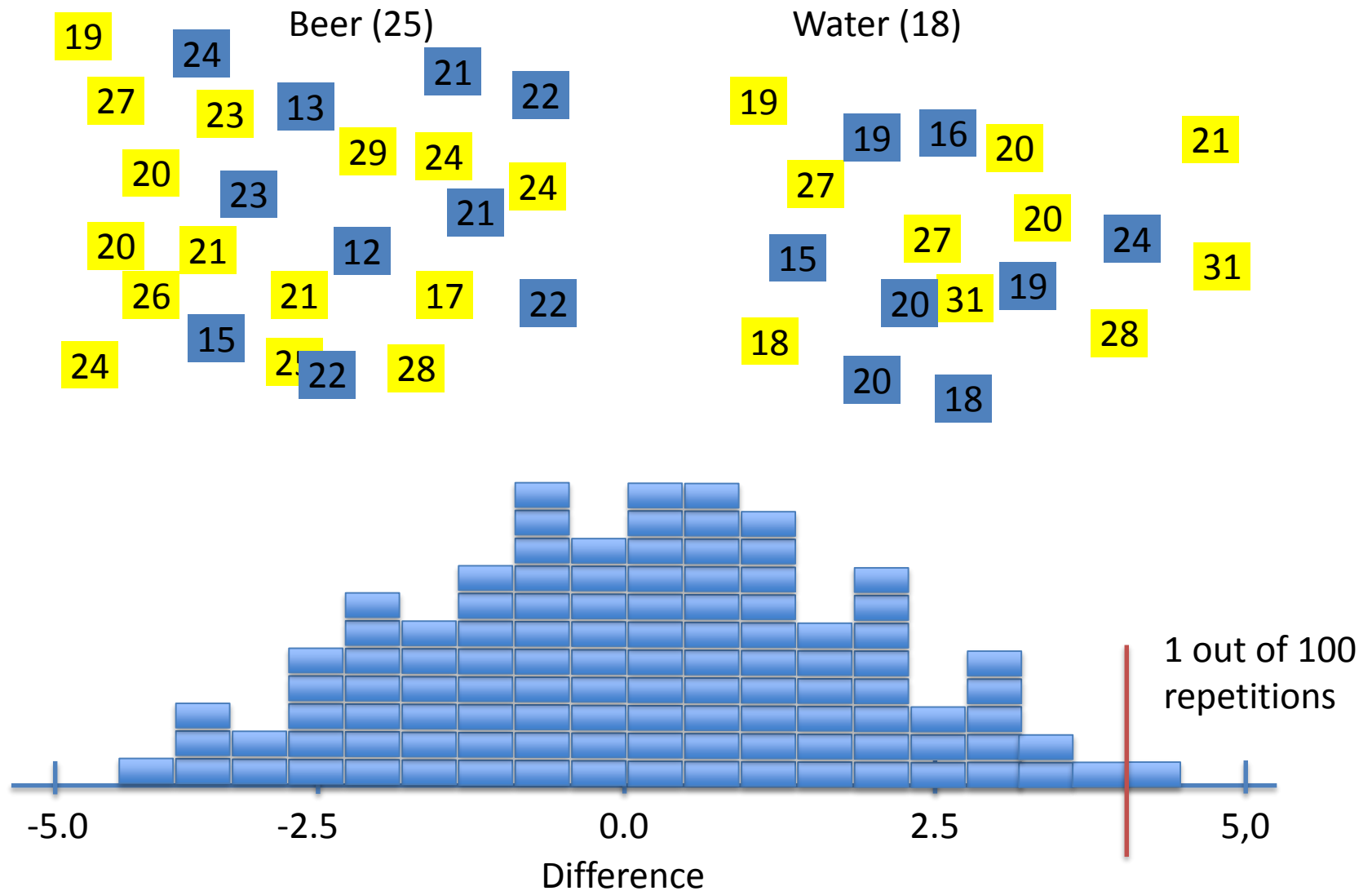
Water (18)



Permutation Test



Permutation Test



ONE TAILED AND TWO TAILED TESTS

One-tailed tests

- Based on a uni-directional hypothesis

Example: The average height of an adult in 2010 is higher than 6 feet

Two-tailed tests

- Based on a bidirectional hypothesis

Example: The average height of an adult in 2010 different from 6 feet

LEVEL OF SIGNIFICANCE

How certain do you want to be?

Example: Significance level of 0.05

- Unidirectional
 - 5% of the time we will higher mean **by chance**
 - 95% of the time the higher mean will be real
- Bidirectional
 - 2.5 % of the time we will find higher mean by chance
 - 2.5% of the time we will find lower mean by chance
 - 95% of time difference will be real

LEVEL OF SIGNIFICANCE AND THE REJECTION REGION

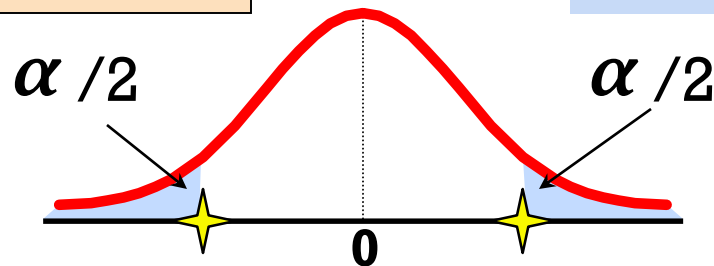
Level of significance = α

Represents critical value

$$H_0: \mu = 3$$

$$H_1: \mu \neq 3$$

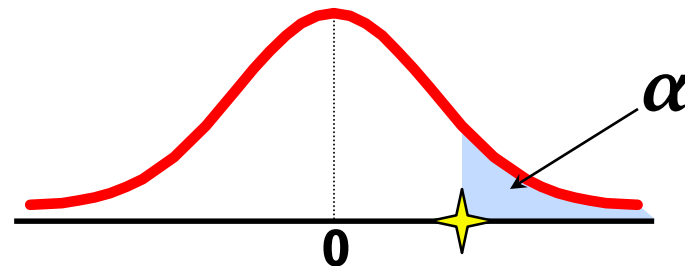
Two-tail test



$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$

Upper-tail test

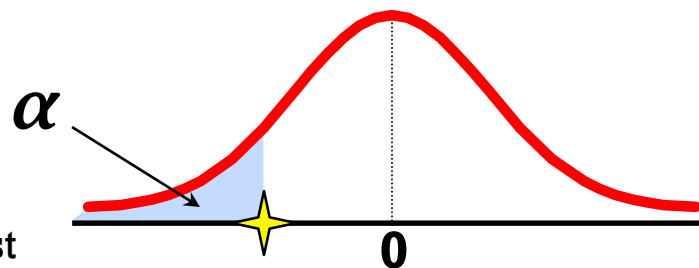


Rejection region is shaded

$$H_0: \mu \geq 3$$

$$H_1: \mu < 3$$

Lower-tail test



*The **p value** is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true*

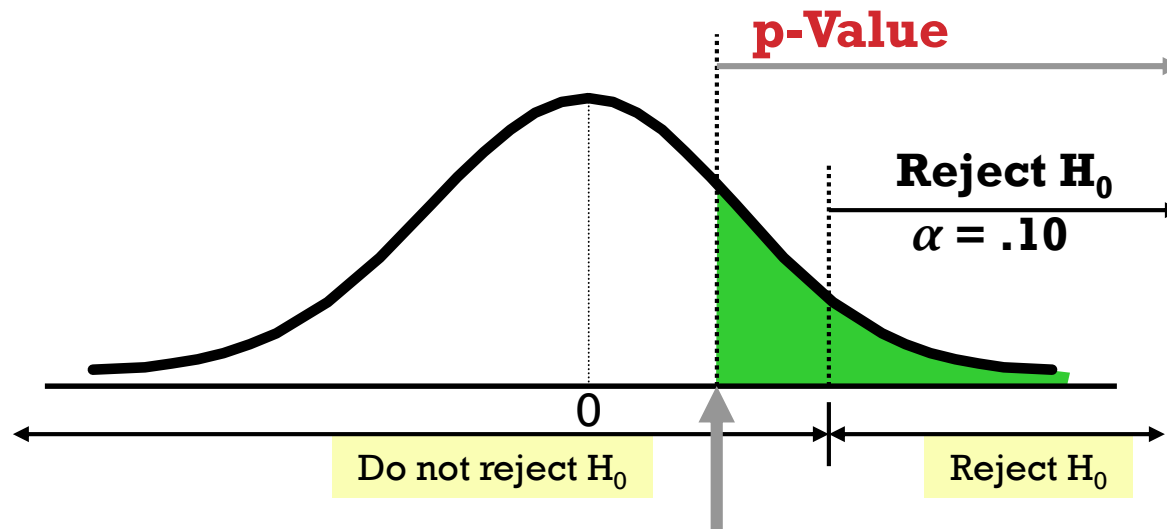
P-VALUE APPROACH TO TESTING

p-value: Probability of obtaining a test statistic more extreme (\leq or \geq) than the observed sample value **given H_0 is true**

- Also called **observed level of significance**
- Smallest value of α for which H_0 can be rejected

P-VALUE

Calculate the p-value and compare to α



OUTCOMES AND PROBABILITIES

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error ($1 - \alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	No Error ($1 - \beta$)

Key:
Outcome
(Probability)

OR IN MORE COMMON WORDS

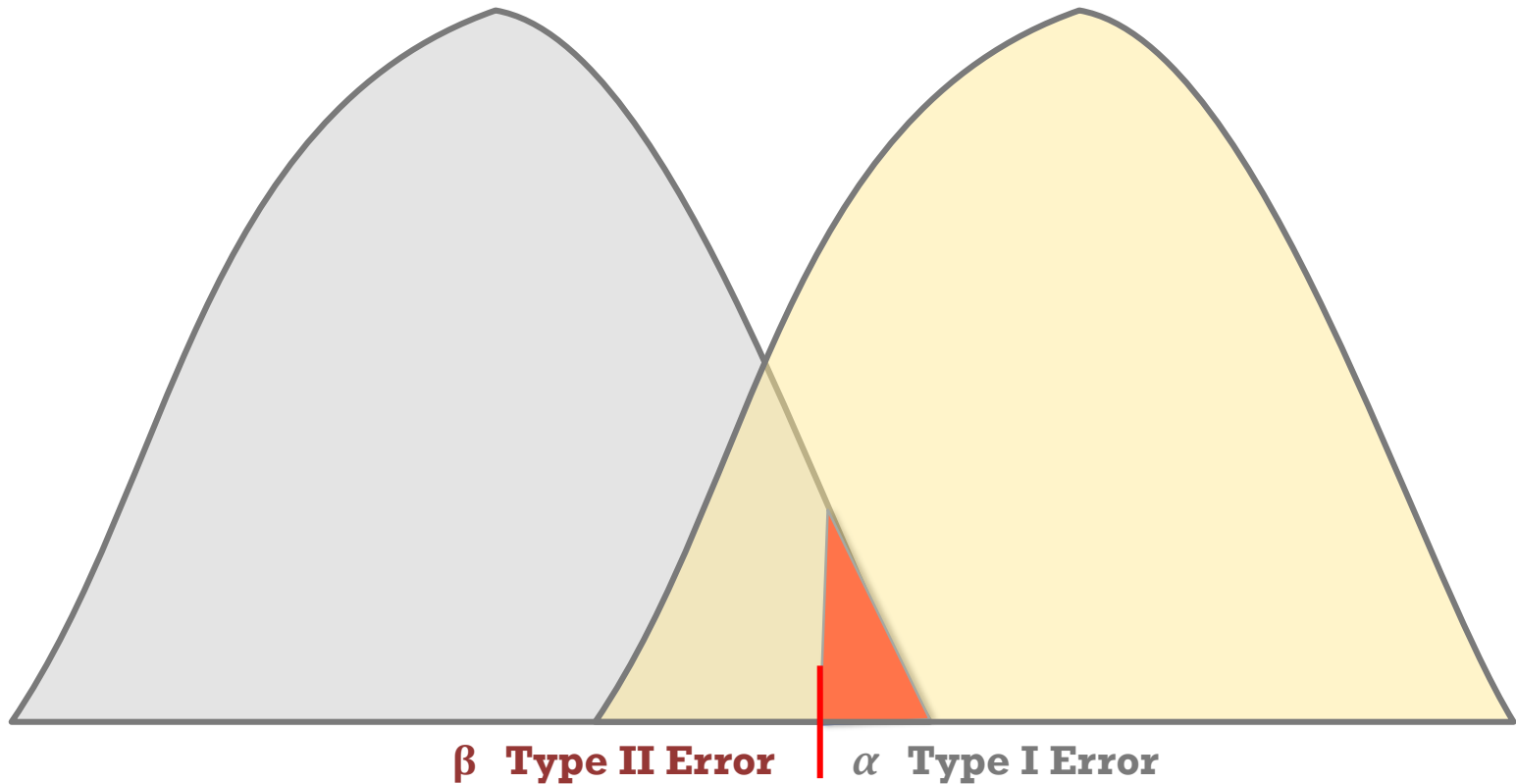
Not seeing something
that's there

	Actual Situation	
Decision	No Fire	Fire
No Alarm	No error (1 - α)	Defective fire alarm (β)
Alarm	Wrong alarm (α)	No error (1 - β)

Seeing something that's
not there

Key:
Outcome
(Probability)

TYPE I VS TYPE II ERROR



CLICKER: CHANGING SIGNIFICANCE LEVELS

Decrease level of significance from 0.01 to 0.05

- a) Probability of finding a difference that does NOT exist goes up
- b) Probability of finding a difference that does exist goes up
- c) Probability of NOT finding a difference that does exist goes up
- d) Probability of NOT finding a difference that does NOT exist goes up

If you believe several answers are correct, select the one which is referred to as either the type I or II error

CLICKER: CHANGING SIGNIFICANCE LEVELS

What happens if we increase our significance from 0.01 to 0.001

- a) Probability of finding a difference that does NOT exist goes up
- b) Probability of finding a difference that does exist goes up
- c) Probability of NOT finding a difference that does exist goes up
- d) Probability of NOT finding a difference that does NOT exist goes up

If you believe several answers are correct, select the one which is referred to as either the type I or II error

P-VALUE HAS PROBLEMS!

Never ever use it as the only measure

footnote: many disciplines used to or still do this

CONSEQUENCES: PUBLICATION BIAS

“In the last few years, several meta-analyses have reappraised the efficacy and safety of antidepressants and concluded that the therapeutic value of these drugs may have been significantly overestimated.”

“Although publication bias has been documented in the literature for decades and its origins and consequences debated extensively, there is evidence suggesting that this bias is increasing.”

“A case in point is the field of biomedical research in autism spectrum disorder (ASD), which suggests that **in some areas negative results are completely absent**”

(emphasis mine)

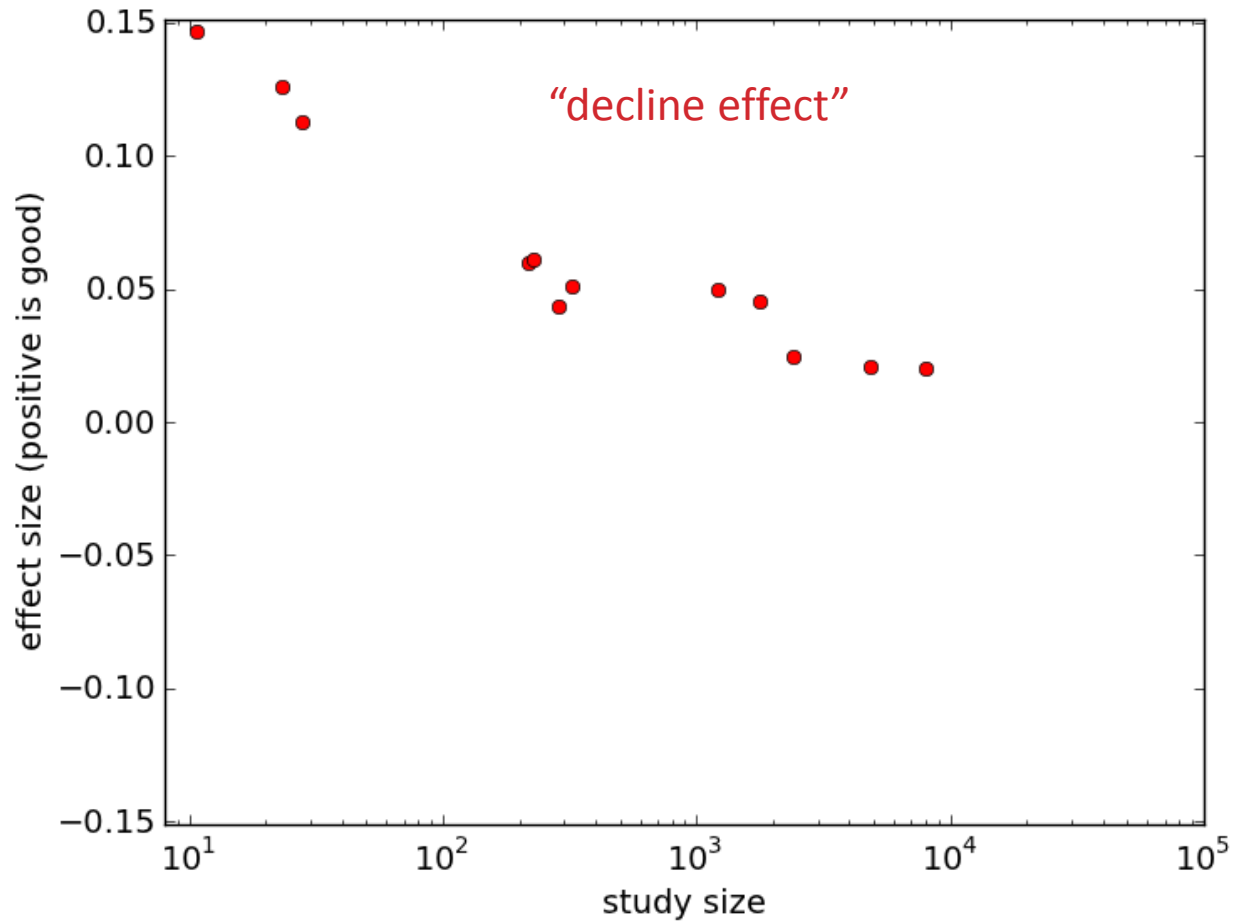
“... a highly significant correlation ($R^2 = 0.13$, $p < 0.001$) between impact factor and overestimation of effect sizes has been reported.”

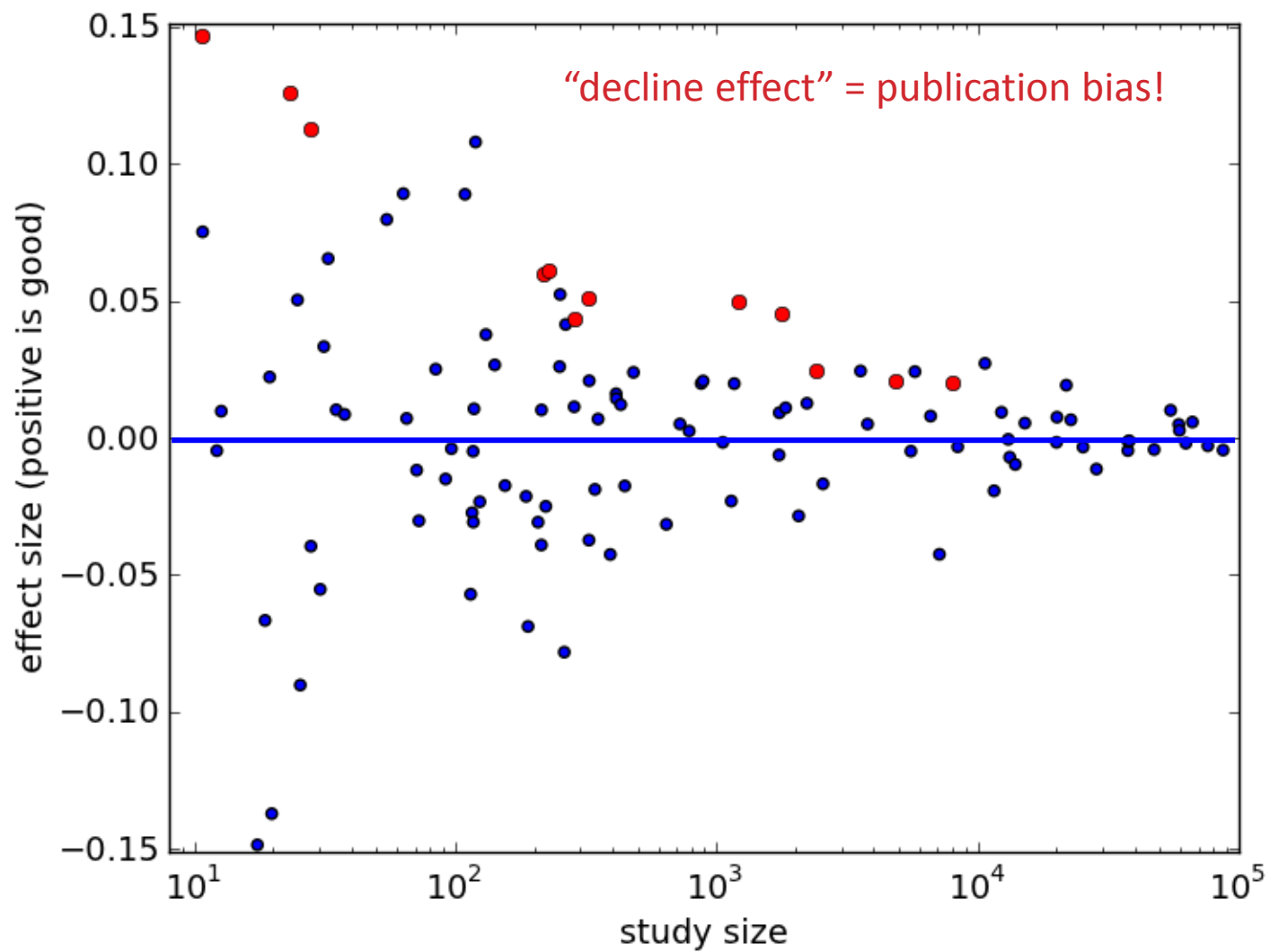
Publication bias: What are the challenges and can they be overcome?

Ridha Joobar, Norbert Schmitz, Lawrence Annable, and Patricia Boksa

J Psychiatry Neurosci. 2012 May; 37(3): 149–152. doi: 10.1503/jpn.120065

PUBLICATION BIAS





BACKGROUND: EFFECT SIZE

Expressed in relevant units

Not just “significant” – how significant?

Used prolifically in meta-analysis to combine results from multiple studies

- But be careful – averaging results from different experiments can produce nonsense

$$\text{Effect size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{standard deviation}}$$

Caveat: Other definitions of effect size exist: odds-ratio, correlation coefficient

EFFECT SIZE

Standardized Mean Difference

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{pooled}}$$

Lots of ways to estimate the pooled standard deviation

$$\sigma_{pooled} = \hat{\sigma}_2$$

Glass, 1976

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}}$$

e.g., Hartung et al., 2008

Choen's Heuristic:

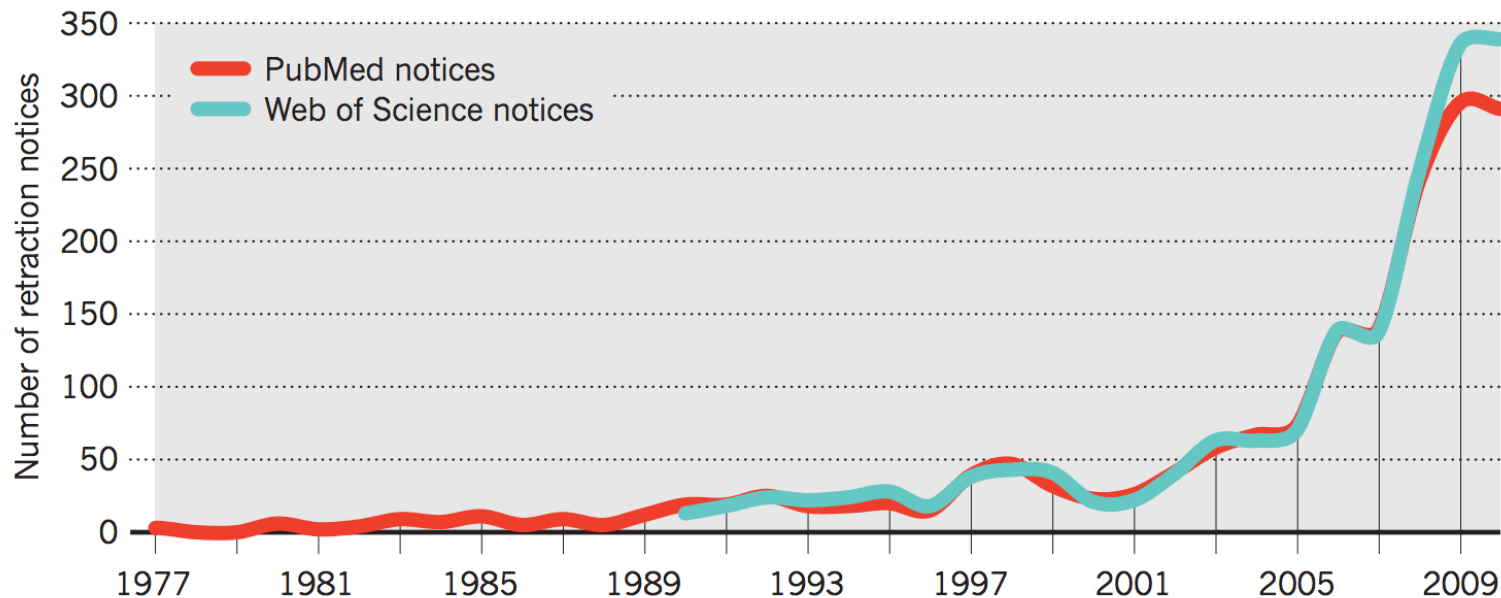
Small=0.20

Medium= 0.50

Large = 0.8

CONSEQUENCES: MISTAKES AND FRAUD

- 2001 – 2011: • 10X increase in retractions
• only 1.44X increase in papers



Richard Van Noorden, 2011, Nature 478

The Rise of the Retractions

<http://www.nature.com/news/2011/111005/pdf/478026a.pdf>

CONSEQUENCES: MULTIPLE HYPOTHESIS TESTING

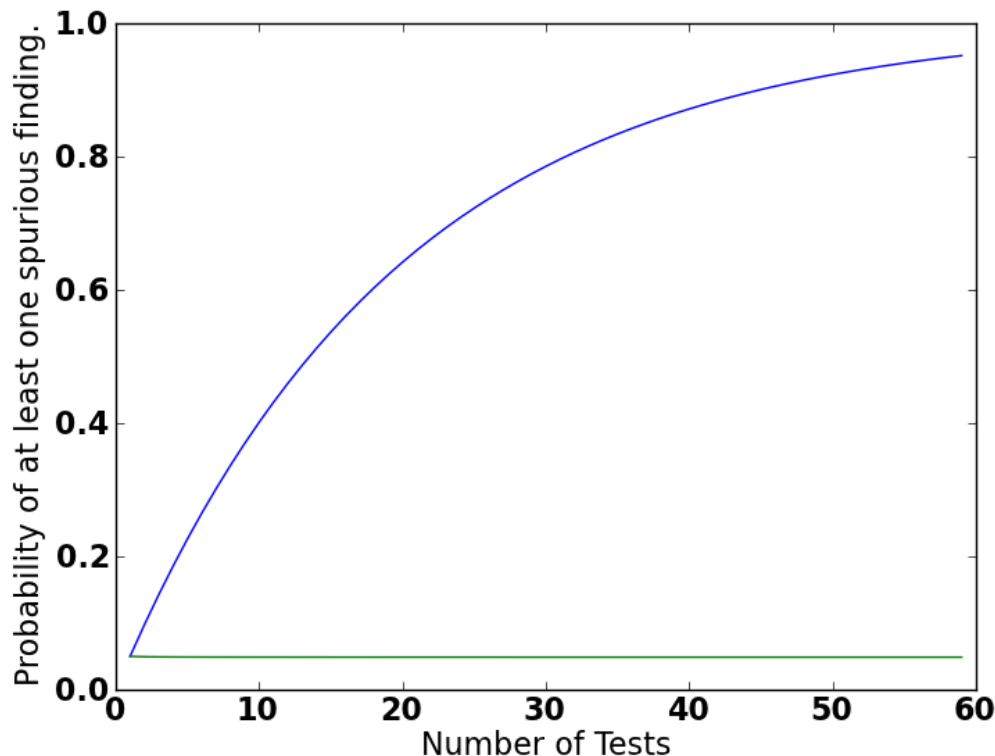
- If you perform experiments over and over, you're bound to find something
- This is a bit different than the publication bias problem: Same sample, different hypotheses
- Significance level must be adjusted down when performing multiple hypothesis tests

$P(\text{detecting an effect when there is none}) = \alpha = 0.05$

$P(\text{detecting an effect when it exists}) = 1 - \alpha$

$P(\text{detecting an effect when it exists on every experiment}) = (1 - \alpha)^k$

$P(\text{detecting an effect when there is none on at least one experiment}) = 1 - (1 - \alpha)^k$



$\alpha = 0.05$

“Familywise Error Rate”

FAMILY-WISE ERROR RATE CORRECTIONS

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

P-Value Has Problems!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1-2, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991



Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

*The **p value** is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true*

Misconception 1

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true. “

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

Misconception #1

“If $P=.05$, the null hypothesis has only a 5% chance of being true”

Let us suppose we flip a penny four times and observe four heads, two-sided $P = .125$. This does not mean that the probability of the coin being fair is only 12.5%.

Misconception #2

A non significant difference (eg, $P .05$) means there is no difference between groups.

- A non significant difference only means the null effect is statistically consistent with the observation
- It does not make the null effect most likely
- In fact, the observed effect best explains the effect regardless the significance.

Misconception #3

A statistically significant finding is (clinical) important

The P value carries no information about the magnitude of an effect, which is captured by the effect estimate and confidence interval.

Misconception #4

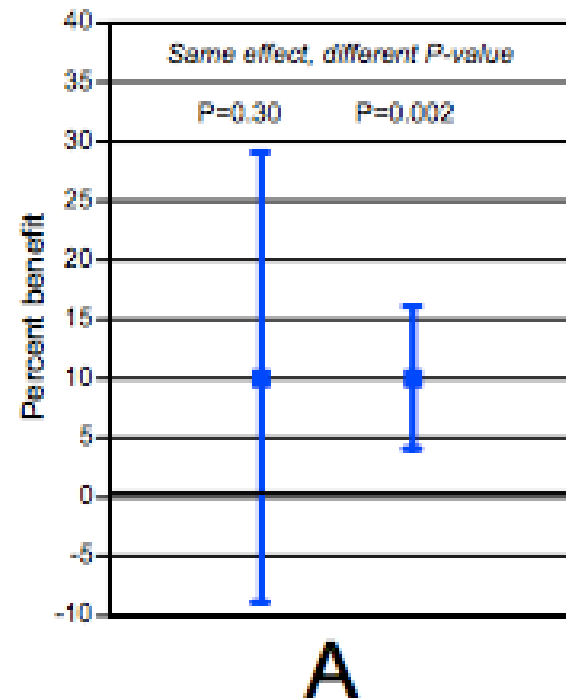
“Studies with P values on opposite sides of .05 are conflicting”

H_0 : Drug T has no effect

H_1 : Drug T has a positive effect

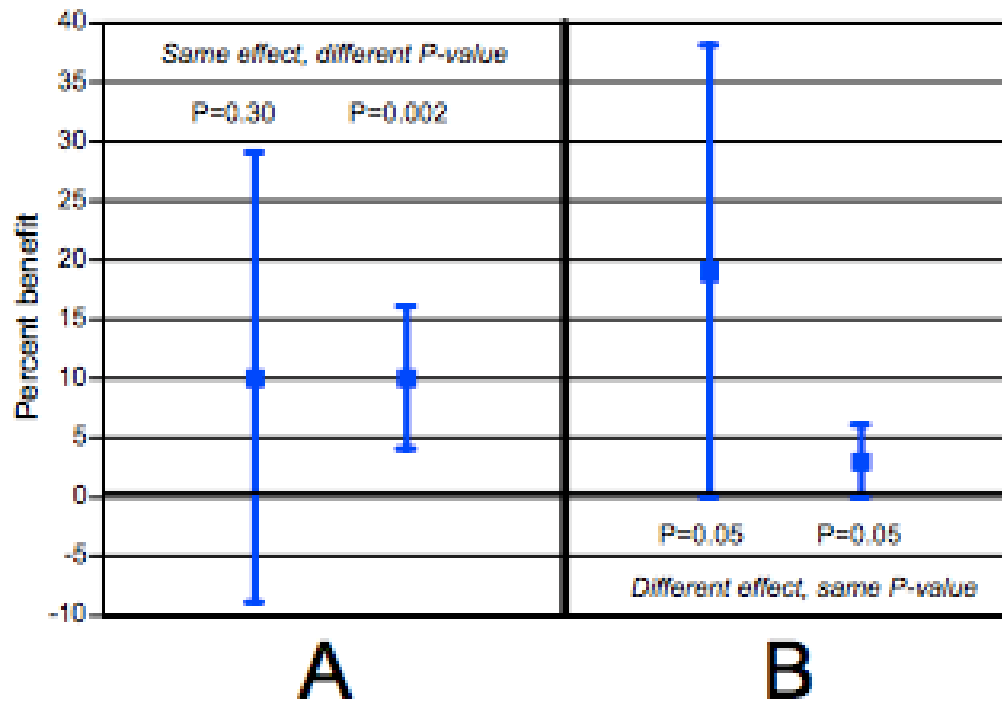
Study I: $P=0.3$

Study II: $P=0.002$



Misconception #5

Studies with the same P value provide the same evidence against the null hypothesis



Clicker:

What is the interpretation of $p = 0.05$

- A) The chances are greater than 1 in 20 that a difference would be found again if the study were repeated.
- B) The probability is less than 1 in 20 that a difference this large could occur by chance alone.
- C) The probability is greater than 1 in 20 that a difference this large could occur by chance alone.
- D) The chance is 95% that the study is correct
- E) None of the above

Misconception #6

P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis

The probability of the observed data, plus more extreme data, under the null hypothesis.