19

STATISTICAL SIGNIFICANCE

"P-Value"

KSU COLLEGE OF MEDICINE
2019 - 2020

ACKNOWLEDGMENTS

TEAM MEMBERS

RAHAF ALTHNYAN

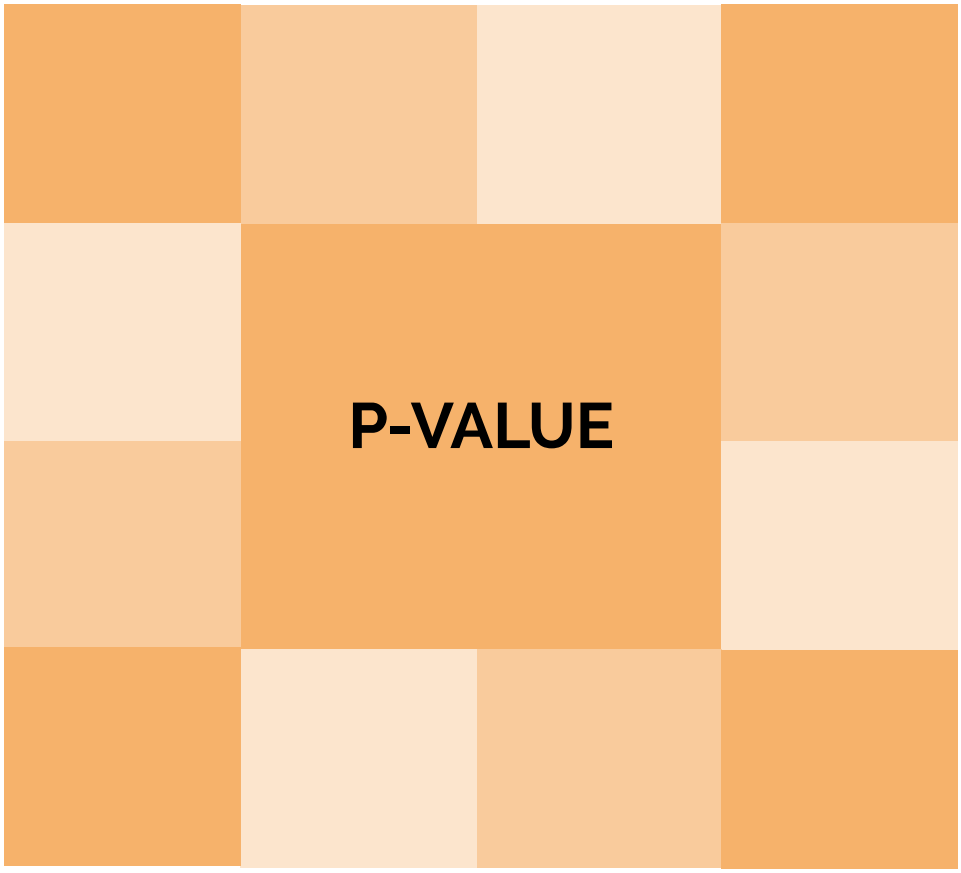
KHALED ALSMARI

REVIEWER

ASEEL BADUKHON

Special thanks to SARAH ALENEZY & 436 TEAM

TABLE OF **CONTENTS**



LECTURE **OBJECTIVES**

By the end of this lecture, I am able to:

- Able to understand the concepts of statistical inference and statistical significance.
- Able to apply the concept of statistical significance(p-value) in analyzing the data.
- Able to interpret the concept of statistical significance(p-value) in making valid conclusions

OVERVIEW

Why use inferential statistics at all?

Average height of all 25-years-old men (population) in KSA is a PARAMETER. Whatever you calculate is called parameter

The height of the members of a sample of 100 such men are measured; the average of those 100 numbers is STATISTIC.

Using inferential statistics, we make inferences about population (taken to be unobservable) based on a random sample taken from the population of interest. (we can generate the parameter from the statistic)

Is risk factor X associated with disease Y?

From the sample, we compute an estimate of the effect of X (risk factor) on Y (outcome) (e.g. risk ratio if cohort study):

- Is the effect real? Did chance play a role

Why worry about chance?

because of Sampling variability...

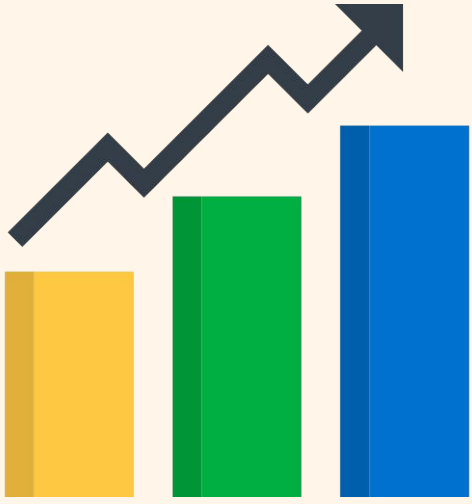
- You only get to pick one sample!

Interpreting the results

Make inferences from data collected using laws of probability and statistics. You have to use these two concepts

- tests of significance (p-value)
- confidence intervals

Parameter	Statistics
-Descriptive measure of a population -Not always possible to measure because it needs the actual value in the population	-Descriptive measure of a sample -Always possible to measure because it doesn't need the actual value in the population



Significance testing

- The interest is generally in comparing two groups (e.g., risk of outcome in the treatment and placebo group) Significance testing can only be done if we have 2 comparison groups (it can't be applied to purely descriptive research)
- The statistical test depends on the type of data and the study design (eg. odds ratio in case-control or cross-sectional studies, and relative risk in RCTs and cohort studies)

Hypothesis Testing

The Null Hypothesis Ho	The Alternative Hypothesis HA
• There is no association between the predictors (associated factors) and outcome variable in the population • Assuming there is no association, statistical tests estimate the probability that the association is due to chance. (this is what we will estimate) • States the assumption (numerical) to be tested • Begin with the assumption that the null hypothesis is **TRUE** • Always contains the '=' sign	• The proposition that there is an association between the predictors and outcome variable • We do not test this directly but accept it by default if the statistical test rejects the null hypothesis • Is the opposite of the null hypothesis • Challenges the status quo • Never contains just the '=' sign • Is generally the hypothesis that is believed to be true by the researcher

- We always **test the null hypothesis**. if it's rejected we automatically accept the alternative hypothesis.

Hypothesis Testing

One and Two sided tests

Hypothesis tests can be one or two sided (tailed):

One tailed tests are directional:
H0: μ1- μ2= 0
HA: μ1- μ2 > 0 or HA: μ1- μ2< 0

Two tailed tests are not directional:

H0: μ1- μ2= 0

HA: μ1- μ2≠ 0

When To Reject H_0 ?

Rejection region: set of all test statistic values for which H_0 will be rejected

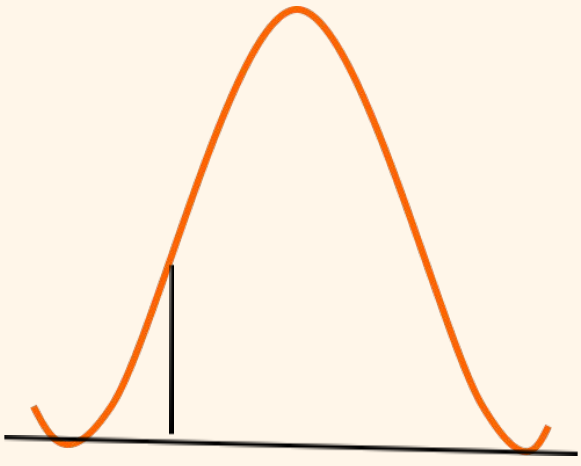
Example of one-sided: we are giving an intervention that will reduce body weight in obese subjects. We know the weight is going to reduce and therefore it is one sided.
If we don't know what is going to happen either (an increase or decrease)
In statistics we always use 2-sided tests to avoid bias

One sided test
A statistical hypothesis test in which alternative hypothesis has only one end. So, it will tell you if there is a relationship between variables in single direction.
The hypothesis directional
Region of rejection is either left or right

Two sided test
A significance test in which alternative hypothesis has two ends. So, if there is a relationship between variables in both direction.
The hypothesis non-directional
Region of rejection are both left & right

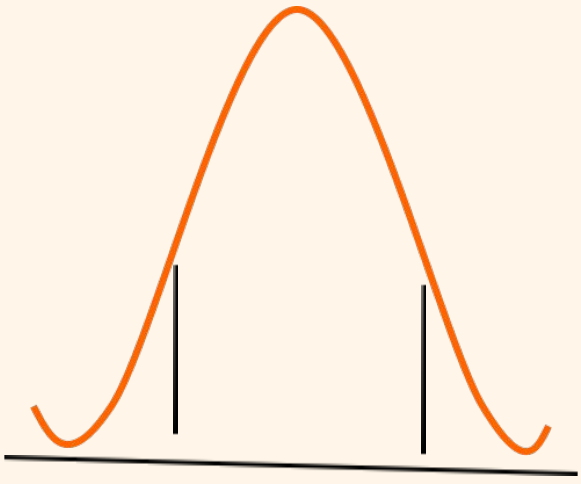
Level of significance, α: Specified before an experiment to define rejection region

One Sided: α = 0.05



Critical Value = -1.64

Two Sided: α/2 = 0.025



Critical Value = -1.96 and +1.96

Hypothesis Testing

Type-I & Type-II Errors

α

a = Probability of rejecting H_0 when H_0 is true

a is called significance level of the test

β

b = Probability of not rejecting H_0 when H_0 is false

1-b is called statistical power of the test

Diagnosis and statistical reasoning:

Significance Difference is

Test result	Present (H0 not true)	Absent (H0 is true)
Reject H0	No error (1-B)	Type I error (a)
Accept H0	Type II error (B)	No error (1-a)

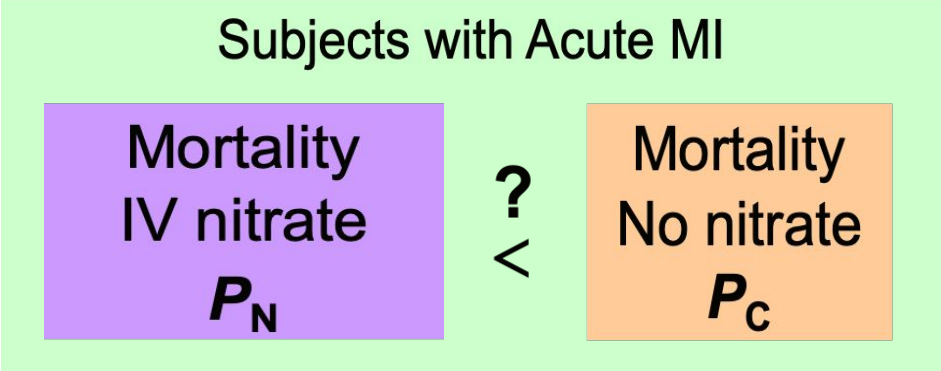
Disease Status

Test result	Present	Absent
+ve	True +ve (Sensitivity)	False +ve
-ve	False -ve	True -ve (Specificity)

- a : significance level
- 1-b : power

• No test will give 100% sensitivity and 100% specificity. There will be always some number in false negative and false positive.
• You prefix the type one error (α) as 5% (0.05), why? Recall normal distribution concept, in normal distribution we considered normal as 95%, remaining 5% we considered it abnormal.
• In clinical medicine, we accept 95% as normal, and 5% as type one error.

SIGNIFICANCE TEST



- Suppose we do a clinical trial to answer the above question
- Even if IV nitrate has no effect on mortality, due to sampling variation, it is very unlikely that $P_N = P_C$
- Any observed difference b/w groups may be due to treatment or a coincidence (or chance)
- Your sample size is people who have acute MI. You subdivide your subjects who have acute MI into two groups. The first group you are giving them IV nitrate while the second group you are not giving them IV nitrate. Why you do that? Because you want to see the effect of IV nitrate on mortality and answer your question which is IV nitrate decrease the mortality rates in people who have MI.

Null Hypothesis (H_o)

- There is no association between the independent and dependent/outcome variables

  - Formal basis for hypothesis testing

- In the example, H_o :"The administration of IV nitrate has no effect on mortality in MI patients" <u>or</u> $P_N - P_C = 0$

Obtaining *P* values:

Trial	Number dead / randomized: Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
			How do we get this *p*-value? ↓		
Chiche	3/50	8/45	0.33	(0.09,1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08,0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33,2.12)	0.70
Jaffe	4/57	2/57	2.04	(0.39,10.71)	0.40
Lis	5/64	10/76	0.56	(0.19,1.65)	0.29
Jugdutt	24/154	44/156	0.48	(0.28, 0.82)	0.007

In the table, there are the 6 studies in the first column, sample size of iv nitrate patients and control in the second and third column. So in IV nitrate (in chiche study) 50 patients were randomized, yet 3 have died (people who died\ total) and we are interested to know how we got the p value and its interpretation?

SIGNIFICANCE TEST

Example of significance testing

- In the Chiche trial:
  - p_N = 3/50 = 0.06; p_C = 8/45 = 0.178

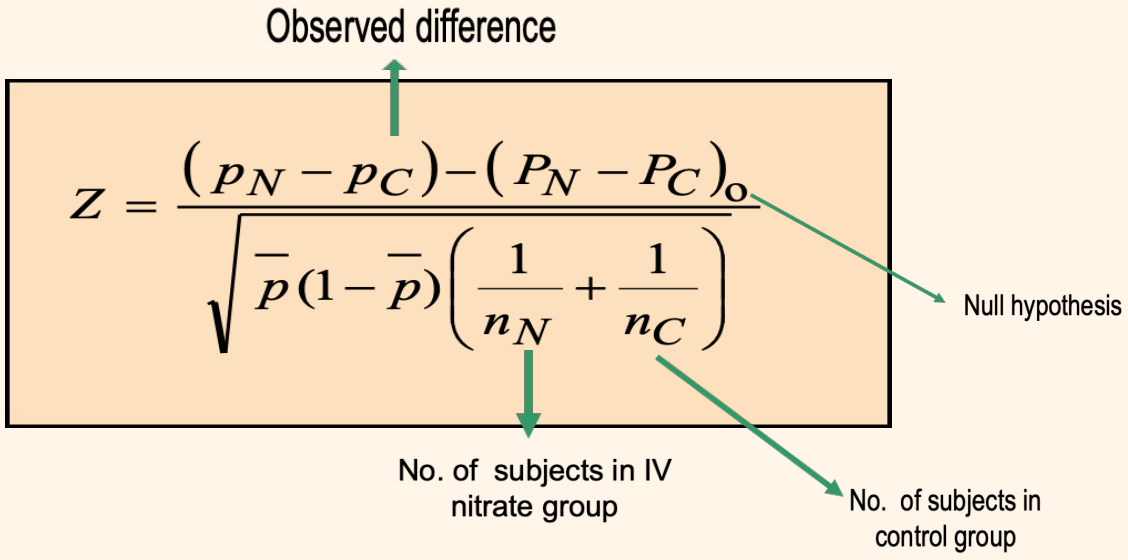
-P: proportion of N: nitrate IV patients & C: control. There is a difference between the two proportions, but is it real? Or by chance? We need a statistical evidence.

- Null hypothesis: No difference.
  - H_0: $p_N - p_C = 0$ or $p_N = p_C$

- Statistical test:
  - Two-sample proportion

Test statistic for Two Population Proportion

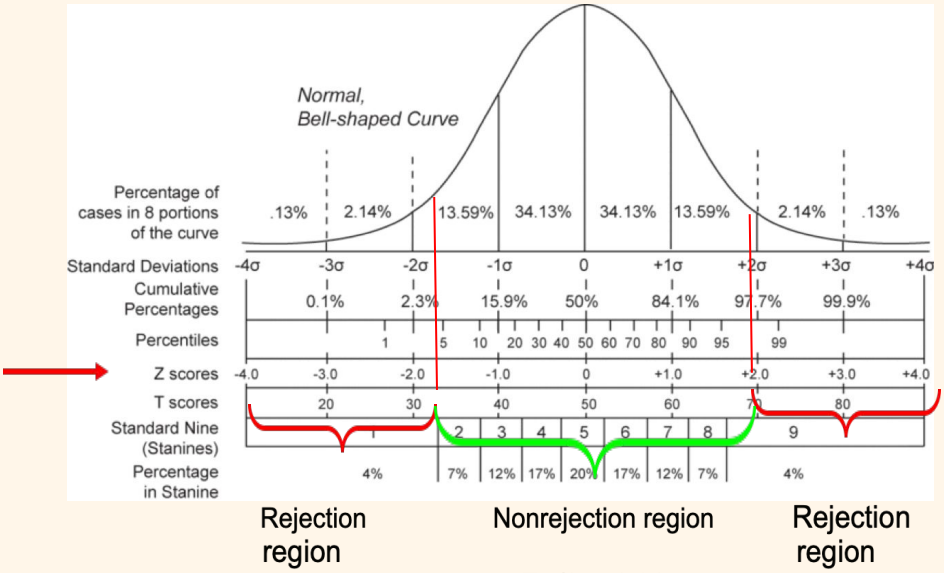
The test statistic for $p_1 - p_2$ is a Z statistic:



Where $\bar{p} = \frac{X_N + X_C}{n_N + n_C}$, $p_N = \frac{X_N}{n_N}$, $p_C = \frac{X_C}{n_C}$

SIGNIFICANCE TEST

Testing significance at 0.05 level



$Z\alpha/2 = 1.96$

Reject H_0 if $Z < -Z\ \alpha/2$ or $Z > Z\ \alpha/2$

Two Population Proportions

$$Z = \frac{(0.06 - 0.178)}{\sqrt{0.116(1-.116)\left(\frac{1}{50} + \frac{1}{45}\right)}} = -1.79$$

This is z score. Where it falls in the normal distribution? According to the normal distribution, So we fail to reject the null hypothesis. No statistical significance in the proportion of mortality between IV nitrate patients and control

Where $\bar{p} = \frac{3+8}{45+50} = 0.116$, $p_N = \frac{3}{45} = 0.06$, $p_C = \frac{8}{50} = 0.178$

SIGNIFICANCE TEST

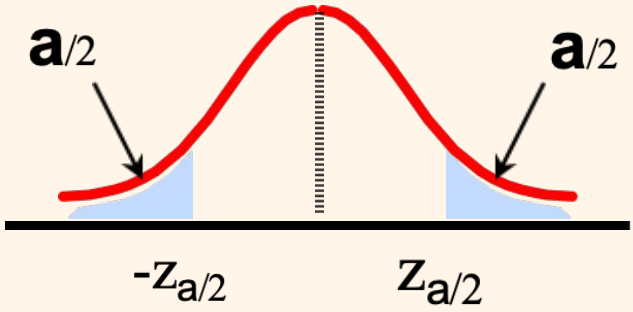
Statistical test for $p_1 - p_2$

What is the corresponding probability value?

Two Population Proportions, Independent Samples:

Two-tail test:
H_0: $p_N - p_C = 0$
H_1: $p_N - p_C \neq 0$

$$Z = \frac{(0.06 - 0.178)}{\sqrt{0.116(1-.116)\left(\frac{1}{50} + \frac{1}{45}\right)}} = -1.79$$



$Z_{a/2} = 1.96$

Reject H_0 if $Z < -Z_{a/2}$
or $Z > Z_{a/2}$

Since -1.79 is > than -1.96, we fail to reject the null hypothesis.

But what is the actual *p*-value?

P (Z<-1.79) + P (Z>1.79)= 0.08

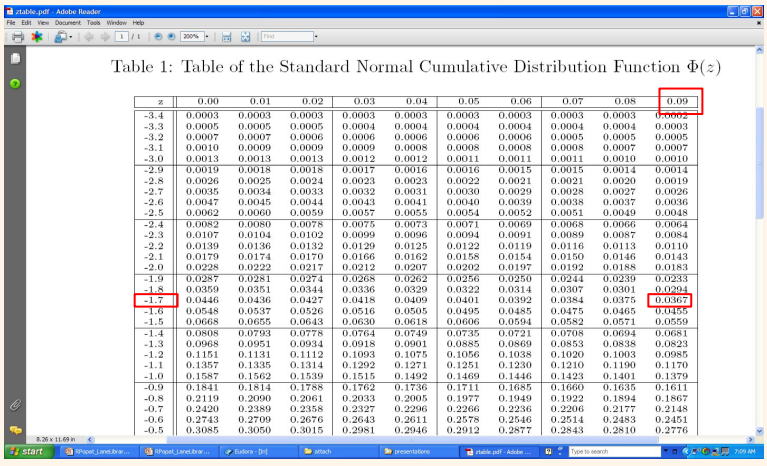


Table 1: Table of the Standard Normal Cumulative Distribution Function $\Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

A p-value of 0.08 mean that if the study is repeated 100 time. you will get the same results 8 time

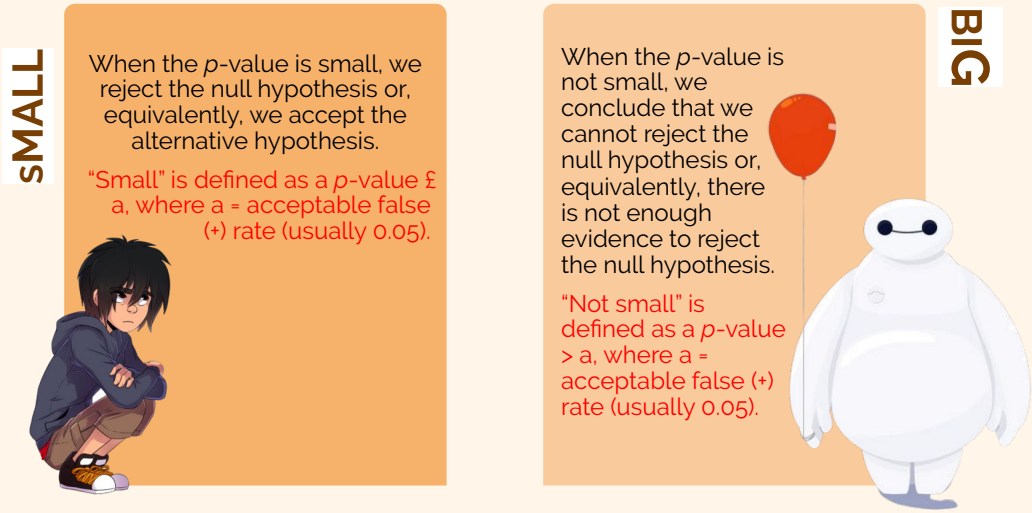
If the calculated p-value is greater than 0.05 then it is NOT statistically significant (at level α)
If p-value is less than 0.05 then it is significant level of α is set before the study

- After calculating a test statistic we convert this to a p-value by comparing its value to distribution of test statistics under the null hypothesis
- Measure of how likely the test statistic value is under the null hypothesis
- p-value $\leq \alpha \Rightarrow$ Reject H_0 at level α
- p-value $> \alpha \Rightarrow$ Do not reject H_0 at level α

P-VALUE

What is a *p*- value?

- 'p' stands for probability
- Tail area probability based on the observed effect
- Calculated as the probability of an effect as large as or larger than the observed effect (more extreme in the tails of the distribution), assuming null hypothesis is true
- Measures the strength of the evidence against the null hypothesis
- Smaller p- values indicate stronger evidence against the null hypothesis



- p ≤0.05 is an arbitrary cut-point
- Does it make sense to adopt a therapeutic agent because p-value obtained in a RCT was 0.049, and at the same time ignore results of another therapeutic agent because p-value was 0.051?
- Hence important to report the exact p-value and not ≤ 0.05 or >0.05
- Size of the p-value is related to the sample size
- Size of the p-value is related to the effect size or the observed association or difference
- P values give no indication about the clinical importance of the observed association
- A very large study may result in very small p-value based on a small difference of effect that may not be important when translated into clinical practice
- Therefore, important to look at the effect size and confidence intervals…

P-VALUE

Statistically Significant **V S** Not Statistically Significant

Statistically Significant	Not Statistically Significant
Reject Ho	Do not reject Ho
Sample value not compatible with Ho	Sample value compatible with Ho
Sampling variation is an unlikely explanation of discrepancy between Ho and sample value	Sampling variation is a likely explanation of discrepancy between Ho and sample value

Clinical Importance Vs. Statistical Significance

so, you should know both clinical and statistical significance.

Clinical importance	Statistical Significance
The *practical importance* of the treatment effect, whether it has a real, palpable, noticeable effect on daily life.	Ruled by the p-value (and confidence intervals). When we find a difference where p <0.05, we call this 'statistically significant'. If a difference is statistically significant, it simply means it was unlikely to have occurred by chance. It doesn't necessarily tell us about the *importance* of this difference or how meaningful it is for patients
Dependent on its implications on existing practice-treatment effect size being one of the most important factors that drives treatment decisions	Heavily dependent on the study's sample size; with large sample sizes, even small treatment effects (which are clinically inconsequential) can appear statistically significant; therefore, the reader has to interpret carefully whether this "significance" is clinically meaningful

*The source of the grey text is:
https://www.students4bestevidence.net/blog/2017/03/23/statistical-significance-vs-clinical-significance/
http://www.picronline.org/article.asp?issn=2229-3485;year=2015;volume=6;issue=3;spage=169;epage=170;aulast=Ranganathan

An example, you have 5000 sample size in both groups (standard and experimental treatment): there is only 2 unit difference in clinical statistics, however; p value is highly significant. so, clinically you will be not be interested . we got this because of the huge sample size.

P-VALUE

Clinical Importance Vs. Statistical Significance

- **Statistically significant AND clinically important.**
  This is where there is an important, meaningful difference between the groups and the statistics also support this. (The flip side of this is where a difference is neither clinically nor statistically significant).
- **Not statistically significant BUT clinically important.**
  This is most likely to occur if your study is underpowered and you do not have a large enough sample size to detect a difference between groups. In this case you might fail to detect an important difference between groups.
- **Statistically significant BUT NOT clinically important.**
  This is more likely to happen the larger sample size you have. If you have enough participants, even the smallest, trivial differences between groups can become statistically significant. It's important to remember that, just because a treatment is statistically significantly better than an alternative treatment, *does not necessarily meanÂ that these differences are clinically important or meaningful to patients*

EXAMPLE

	Yes	No
Standard	0	10
New	3	7

Absolute risk reduction = 30% ← Clinical

Fisher exact test: p = 0.211 ← Statistical

In this example, you have only 10 cases in each group: in the standard treatment there is no improvement. in the new treatment there are 3 cases that improved. clinically GOOD but statistically NOT GOOD. why? sample size is small

Statistical Significance

Clinical significance: Practical importance if observed effect	Statistical Significance: Not Significant	Statistical Significance: Significant
Not important		Annoyed
Important	Very sad	Elated

EXAMPLES

Trial	Number dead / randomized: Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
Chiche	3/50	8/45	0.33	(0.09,1.13)	0.08
Some evidence against the null hypothesis					
Flaherty	11/56	11/48	0.83	(0.33,2.12)	0.70
Very weak evidence against the null hypothesis…very likely a chance finding					
Lis	5/64	10/76	0.56	(0.19,1.65)	0.29
Jugdutt	24/154	44/156	0.48	(0.28, 0.82)	0.007
Very strong evidence against the null hypothesis…very unlikely to be a chance finding					

very common

very rare

Interpreting *P* values if the null hypothesis were true:

Trial	Number dead / randomized: Intravenous nitrate	Control	Risk Ratio	95% C.I.	P value
Chiche	3/50	8/45	0.33	(0.09,1.13)	0.08
…8 out of 100 such trials would show a risk reduction of 67% or more extreme just by chance					
Flaherty	11/56	11/48	0.83	(0.33,2.12)	0.70
…70 out of 100 such trials would show a risk reduction of 17% or more extreme just by chance…very likely a chance finding					
Lis	5/64	10/76	0.56	(0.19,1.65)	0.29
Jugdutt	24/154	44/156	0.48	(0.28, 0.82)	0.007
Very unlikely to be a chance finding					

1-0.33 = 0.67
Less than 1= protection
More than 1= risk

- In chiche and flaherty studies, the sample size is almost the same(45,48) but the p value is different why? because risk ratio (outcome of the study) is different.
- In lis and jugdutt studies,risk reduction is almost the same. However, the p value is different why? because sample size is different

EXAMPLES

Interpreting *P* values

Trial	Intravenous nitrate	Control	Risk ratio	95% confidence interval	*P* value
Chiche	3/50	8/45	0.33	(0.09, 1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08, 0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33, 2.12)	0.7
Jaffe	4/57	2/57	2.04	(0.39, 10.71)	0.4
Lis	5/64	10/77	0.56	(0.19, 1.65)	0.29
Jugdutt	12/77	44/157	0.48	(0.28, 0.82)	0.007

- Lis and Jugdutt trials are similar in effect (~ 50% reduction in risk)…but Jugdutt trial has a large sample size

Trial	Intravenous nitrate	Control	Risk ratio	95% confidence interval	*P* value
Chiche	3/50	8/45	0.33	(0.09, 1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08, 0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33, 2.12)	0.7
Jaffe	4/57	2/57	2.04	(0.39, 10.71)	0.4
Lis	5/64	10/77	0.56	(0.19, 1.65)	0.29
Jugdutt	12/77	44/157	0.48	(0.28, 0.82)	0.007

- Chiche and Flaherty trials approximately same size, but observed difference greater in the Chiche trial

- If a new antihypertensive therapy reduced the SBP by 1 mmHg as compared to standard therapy we are not interested in swapping to the new therapy.
- However, if the decrease was as large as 10 mmHg, then you would be interested in the new therapy.
- Thus, it is important to not only consider whether the difference is statistically significant by the possible magnitude of the difference should also be considered.