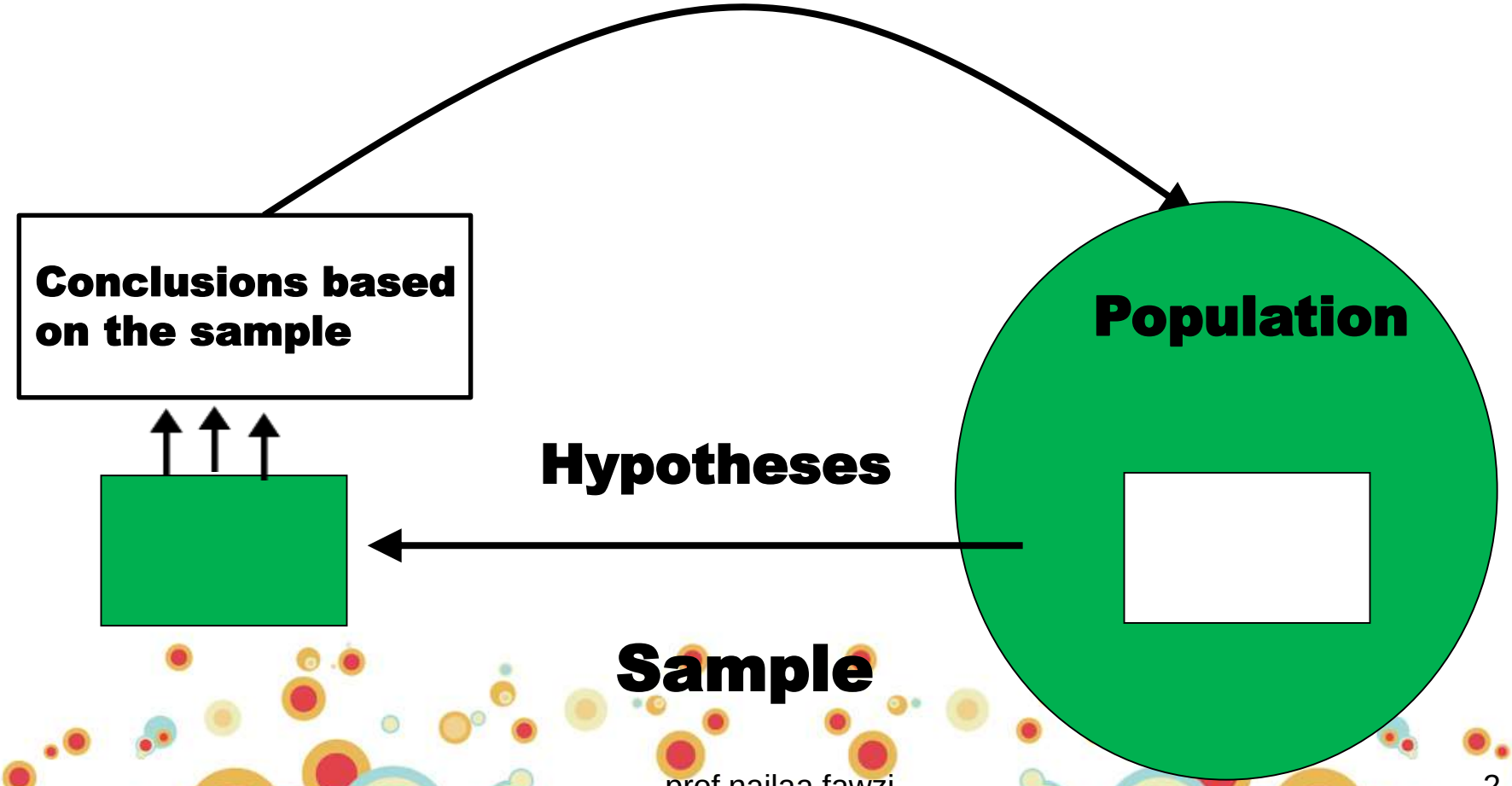
Inferential statistics

Prof Dr NAJLAA FAWZI

The idea of statistical inference

Generalisation to the population



Inferential statistics

- **Uses patterns in the sample data to draw inferences about the population represented, accounting for randomness.**
- **Two basic approaches:**
  - **Hypothesis testing**
  - **Estimation**
- **Common goal: conclude on the effect of an independent variable (exposure) on a dependent variable (outcome).**

- **The general goal of a hypothesis test is to rule out chance (sampling error) as a possible explanation for the results from a research study.**
- **Hypothesis testing is a technique to help determine whether a specific treatment has an effect on the individuals in a population.**

Hypothesis Testing

The hypothesis test is used to evaluate the results from a research study in which

1. **A sample is selected from the population.**
2. **The treatment is administered to the sample.**
3. **After treatment, the individuals in the sample are measured.**

- **If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect.**
- **However, it is also possible that the difference between the sample and the population is simply sampling error**

- **The purpose of the hypothesis test is to decide between two explanations:**

  1. The difference between the sample and the population can be explained by sampling error (there does not appear to be a treatment effect)

  2. The difference between the sample and the population is too large to be

  explained by sampling error (there does appear to be a treatment effect).

HYPOTHESIS TESTING

The purpose of hypothesis testing is to help the clinician , researcher or the administrator in reaching a decision concerning a population by examining a sample from that population.

It involves conducting a test of statistical significance and qualifying the degree to which sampling variability may account for the results observed in a particular study.

Hypotheses: may be defined simply as statement about one or more populations , it is usually concerned with the parameters of the population and by means of hypothesis , one can determine whether or not such statements are compatible with available data.

Researchers are concerned with two types of hypotheses :

1- Research Hypothesis: is the supposition that motivates the researcher , it may be the results of years of observations.

2- Statistical Hypothesis: is a hypothesis that is stated in such away that they may be evaluated by appropriate statistical techniques.

The aim of a statistical test

To reach a scientific decision (“yes” or “no”) on a difference (or effect), on a probabilistic basis, on observed data.

Procedure for Performing an Inferential Test

Here is a step-by-step procedure for performing inferential statistics.

1. Start with a theory

(TV violence reduces children's ability to detect violent behavior because it blurs the distinction between real and fantasy violence)

2. Make a research hypothesis

(Children who experience TV violence will fail to detect violent behavior in other children)

3. Operationalize the variables

4. Identify the population to which the study results should apply

5. Form a null hypothesis for this population

6. Collect a sample of children from the population and run the study

7. Perform statistical tests to see if the obtained sample characteristics are sufficiently different from what would be expected under the null hypothesis to be able to reject the null hypothesis.

Tests of Significance: **There are many tests of significance developed and utilized .**

Most common are " Z"(normal curve test), students " t" test, Chi-square test, F- test.

Procedure and steps

The steps involved in general in the utilization of any test of significance are :

1- Data

The nature of the data that form the basic for testing is determined and well understood.

The data must be determined whether quantitative or qualitative , and it is presented in mean and SD for quantitative data, and in frequency and proportion for qualitative data.

Parametric , and non parametric tests?

2- Assumptions

The assumptions that are of importance in hypothesis testing include:

a- randomly selection of sample

b- independence of samples

c- normality distribution of the population ****

d- equality of the variance of the population.****

***** Each test of significance has its own proper assumption.**

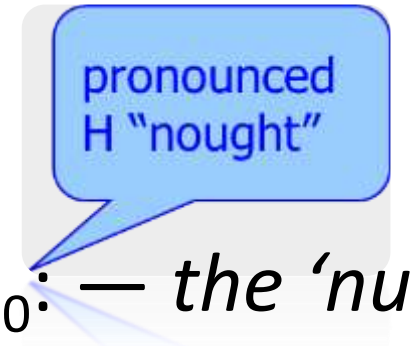
3- Hypotheses

We have ((2)) types of statistical hypotheses

a- Null hypothesis

b- Alternative hypothesis

null hypothesis



H_0: — *the 'null' hypothesis*

Which is defined as the hypothesis of no difference ,

it is a statement of agreement with (or no difference) .

The samples or populations being compared in an experiment ,

study or test are similar.

Any difference occurs is related to chance and not

to any other factor.

B- Alternative Hypothesis{ HA} : **or it is the hypothesis of difference**

(so it is the hypothesis adopted when we reject H0).

Gastroenteritis outbreak in city A:

“The risk of illness was higher among diners who ate home cooked meat(RR=3.6).”

Is the association due to chance?

The two hypothesis

There is NO difference between the two groups (=no effect)	Null Hypothesis (H_0) (e.g.: RR=1)

There is a difference between the two groups (=there is an effect)	Alternative Hypothesis (H_1) (e.g.: RR=3.6)

- **Null hypothesis (H0): “There is no association between consumption of cooked meat and gastroenteritis.”**

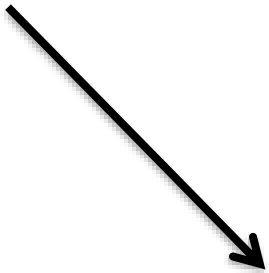
- **Alternative hypothesis (H1): “There is an association between consumption of cooked meat and gastroenteritis.”**

- **Tests of statistical significance**
- **Data not consistent with H_0 :**
  - **H_0 can be rejected in favour of some alternative hypothesis H_1 (the objective of our study).**
- **Data are consistent with the H_0 :**
  - **H_0 cannot be rejected**

You can not say that the H_0 is true.

You can only decide to reject it or not reject it.

H0 is either not rejected [mean that the data tested do not provide sufficient evidence to cause rejection] , or it is rejected [data are not compatible with , but are supportive to some other hypothesis which is the alternative hypothesis].



it is important to remember that when we fail to reject H0,

we do not say that it is true , but it may be true.

4- level of Significance

Choosing a level of significance prior to data collection, the investigator must decide on the level of significant that represents a tolerable risk of committing type I error ,

when comparing a new procedure or medication to an established treatment ,it is often desirable to select a very low value for (α typically = 0.01), especially if the new treatment is know to be associated with substantially greater risk of serious side effects.

Level of significance is the criteria used to reject null hypothesis , to find out significance difference between two variable.

It is defined as the probability of making a decision to reject the null hypothesis .

The decision is often made by using p value , p value denotes significance level.

The significance level is usually set at 0.05 (p= 0.05).

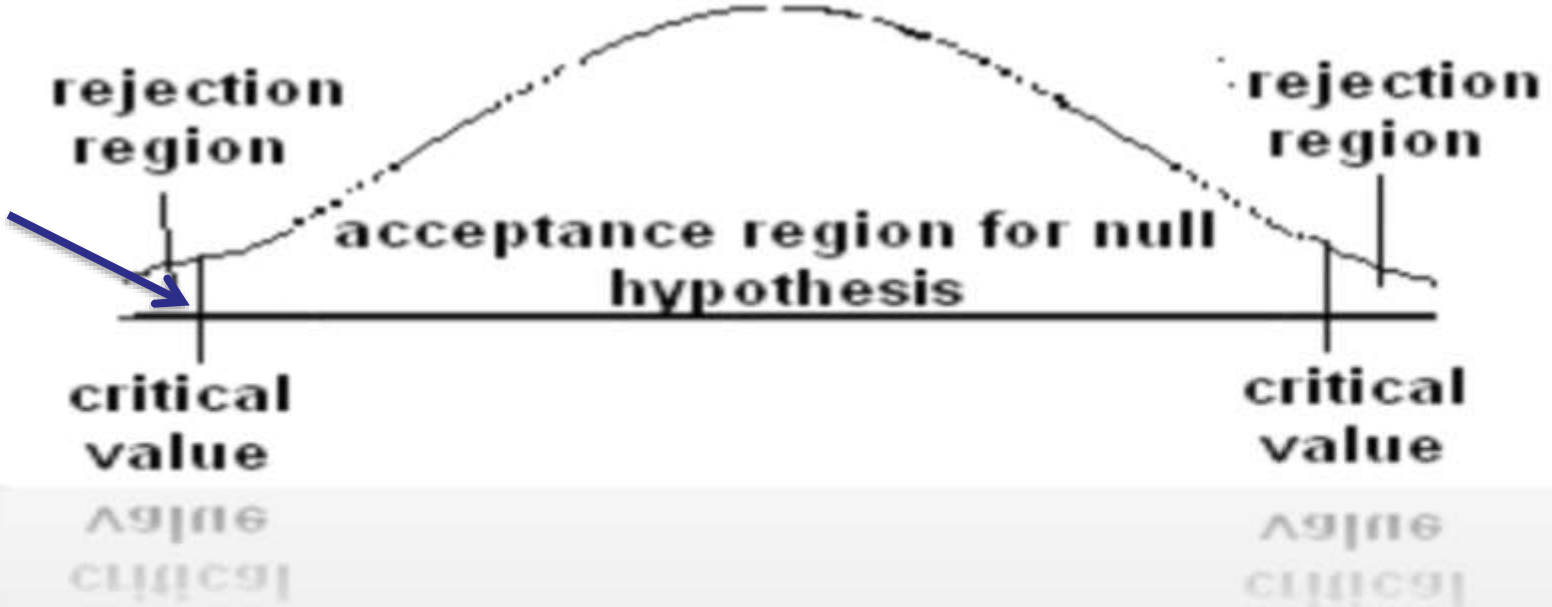
The smaller the p- value (significance level) , the more significant the result said to be.

Level of significance it specifies the area under the curve of the distribution of test statistics , that represent the basis on which determination of the rejection region and acceptance region for the tested data.

Acceptance region : a set of values of the statistical test lead to the acceptance of Ho.

Rejection region : a set of values of the statistical test lead to the rejection of Ho.

Critical value : value of test statistics that separate the rejected region from the acceptance region.



It is the tabulated value that lie on both extremes of the acceptance area

The null hypothesis to be accepted; calculated values must be less than the critical value (tabulated value) to fall in the acceptance area for null hypothesis & so to accept H0.

TABLE : Percentiles of the t Distribution

$p(t_{10} \leq 2.2281) = .975$

d.f.	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480
80	1.2922	1.6641	1.9901	2.374	2.6388
90	1.2910	1.6620	1.9867	2.368	2.6316
100	1.2901	1.6602	1.9840	2.364	2.6260
120	1.2887	1.6577	1.9799	2.358	2.6175
140	1.2876	1.6558	1.9771	2.353	2.6114
160	1.2869	1.6545	1.9749	2.350	2.6070
180	1.2863	1.6534	1.9733	2.347	2.6035
200	1.2858	1.6525	1.9719	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576

TABLE 2.1

Chi Square Values and Probability

Degrees of Freedom	$P = 0.99$	0.95	0.80	0.50	0.20	0.05	0.01
1	0.000157	0.00393	0.0642	0.455	1.642	**3.841**	**6.635**
2	0.020	0.103	0.446	1.386	3.219	**5.991**	**9.210**
3	0.115	0.352	1.005	2.366	4.642	**7.815**	**11.345**
4	0.297	0.711	1.649	3.357	5.989	**9.488**	**13.277**
5	0.554	1.145	2.343	4.351	7.289	**11.070**	**15.086**
6	0.872	1.635	3.070	5.348	8.558	**12.592**	**16.812**
7	1.239	2.167	3.822	6.346	9.803	**14.067**	**18.475**
8	1.646	2.733	4.594	7.344	11.030	**15.507**	**20.090**
9	2.088	3.325	5.380	8.343	12.242	**16.919**	**21.666**
10	2.558	3.940	6.179	9.342	13.442	**18.307**	**23.209**
15	5.229	7.261	10.307	14.339	19.311	**24.996**	**30.578**
20	8.260	10.851	14.578	19.337	25.038	**31.410**	**37.566**
25	11.524	14.611	18.940	24.337	30.675	**37.652**	**44.314**
30	14.953	18.493	23.364	29.336	36.250	**43.773**	**50.892**

From Fisher, R. A., and Yates, F. (1943) *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, London.

5- Selection of the appropriate test statistics

we have to apply test statistics which uses the sample data to reach a decision to reject or not reject null hypothesis.

6- Statistical decision

Comparison of the calculated test criterion value with that of the theoretical value(tab value) at 5% , 1%.

if the calculated value test criterion is **lower** than the theoretical value, **Ho is not rejected**

[avoid using the word accepted]. **WHY???.**

If the calculated test criterion value is higher than the theoretical value Ho is rejected, and the HA is accepted

7- Drawing of the conclusion [or inference]

On the basis of the level of significance is deciding whether the differenceobserved is due to chance or due to some other known factors.

Possible Out come of a Statistical Test

Statistical decision	Ho true	Ho false & HA true
Do not reject Ho	Correct decision	Incorrect decision ,or type II error [we don't know the concurrent risk of committing a type II error since B is usually unknown , but as we know that in most practical situation , it is larger than α]
Reject Ho	Incorrect decision or type I error (if we made α small , therefore the probability of a type I error is small)	Correct decision

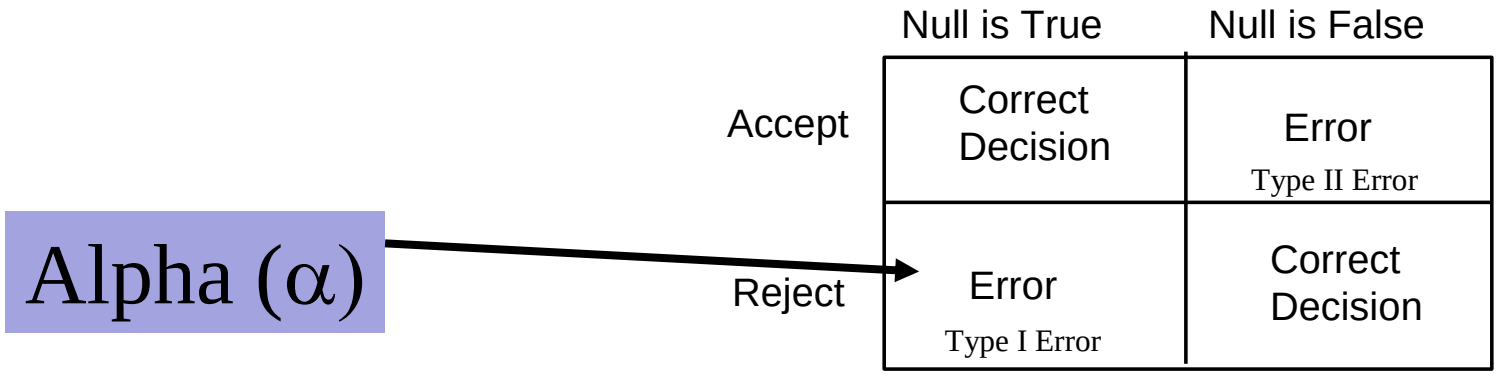
Statistical errors : Statistical errors are used to describe possible errors made in statistical decision .

There are two basic types of errors :

1- Type I error

2- Type II error

Possible Outcomes in Hypothesis Testing



Difference observed is really just sampling error
The prob. of type one error

Type I Errors

- **A Type I error occurs when the sample data appear to show a treatment effect when, in fact, there is none.**
- **In this case the researcher will reject the null hypothesis and falsely conclude that the treatment has an effect.**
- **Type I errors are caused by unusual, unrepresentative samples. Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though the treatment has no effect.**
- **The hypothesis test is structured so that Type I errors are very unlikely; specifically, the probability of a Type I error is equal to the alpha level.**

α error (Type I error) (rejecting true H0)

Incorrectly rejecting true H0.

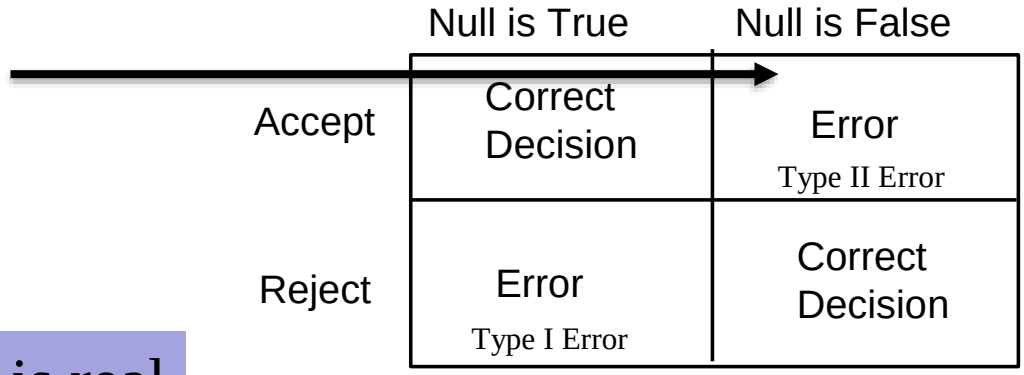
The predetermined risk of accepting HA .

We observe difference when it is not true --- false positive .

α Is called the size of the test

Possible Outcomes in Hypothesis Testing

Beta (β)



Difference observed is real
Failed to reject the Null

Type II Errors

- **A Type II error occurs when the sample does not appear to have been affected by the treatment when, in fact, the treatment does have an effect.**
- **In this case, the researcher will fail to reject the null hypothesis and falsely conclude that the treatment does not have an effect.**
- **Type II errors are commonly the result of a very small treatment effect. Although the treatment does have an effect, it is not large enough to show up in the research study.**

β error (Type II error) (accepting false H0)

The chance of error associated with the failure to reject the null hypothesis given that a difference actually does exist.

This is the case if we cannot detect a real difference.

(False negative) for a diagnostic test.

β error Determine the power of the study

Power: is the ability of a statistical test to detect a difference of a specified magnitude { known as a clinically important difference}, give that this difference exists in the population being compared .

That is , it is the probability that a statistical test will reject Ho given that Ho is false.

Power= 1-B

= probability to detect a real difference = true positive decision

1 - Power = β = probability to accept a false H0.

1 - α = the probability to reject false H0 (i.e. true negative decision)

Although it is possible to guard against a type I error simply by using a more strict (lower) level of α, preventing a type II error is not so easy. Because a type II error involves accepting a false null hypothesis, the ability of a statistical test to avoid a type II error depends on its ability to detect a null hypothesis that is false.

This ability is called the power of the test, and it is equal to 1 -β: it is the probability that a false null hypothesis will be rejected.
Conventionally, a study is required to have a power of 0.8 (or a β of 0.2) to be acceptable—in other words, a study that has a less than 80% chance of detecting a false null hypothesis is generally judged to be unacceptable.

Calculating β and determining the power of a test is complex. Nevertheless, it is clear that a test's power, or ability to detect a false null hypothesis, will increase as α increases (e.g., from .01 to .05).

This will make the critical values of t less extreme, thus increasing the size of the areas of rejection and making rejection of the null hypothesis more likely.

There will always be a trade-off between type I and type II errors: increasing α reduces the chance of a type II error, but it simultaneously increases the chance of a type I error.

Increasing the sample size is the most practical and important way of increasing the power of a statistical test.

Ideally, all studies that report acceptance of the null hypothesis should also report the power of the test used so that the risk of a type II error is made clear.

Unlike α and B , power is not the risk of a particular error . Instead , is the probability that a statistical test will reach a particular correct conclusion , the power of a statistical test is analogous to the sensitivity (true positive rate) of a diagnostic test.

Study is required to have a power of 0.8 to be acceptable

i.e. a study that has less than 80% chance of detecting

a false null hypothesis is unacceptable

Ways of Increasing Power:

- **Increase sample size**
- **Make alpha level less conservative**
- **Use one-tailed versus a two-tailed test**

A powerful test decreases the chances of making a Type II Error

POWER: ability to reduce type II error

Directional Tests

- **When a research study predicts a specific direction for the treatment effect (increase or decrease), it is possible to incorporate the directional prediction into the hypothesis test.**

- **The result is called a directional test or a one-tailed test. A directional test includes the directional prediction in the statement of the hypotheses and in the location of the critical region.**

- **For example, if the original population has a mean of μ = 80 and the treatment is predicted to increase the scores, then the null hypothesis would state that after treatment:**

  H0: μ ≤ 80 (there is no increase)

- **In this case, the entire critical region would be located in the right-hand tail of the distribution because large values for M would demonstrate that there is an increase and would tend to reject the null hypothesis.**

Meaning of " Statistically Significance" and " Not Statistically Significance" results

Statistical significance is not synonymous with biologic or clinical relevance .

Conversely ,the failure to demonstrate statistical significance does not rule out the existence of a clinically important difference between two population.

Any difference , however small , may be found statistically significant (unlikely to have occurred by random chance), if the sample size n is sufficiently large. However, a difference of small magnitude while statistically significant , may not be clinically important.

- **Presence of significance does not prove clinical or biological relevance of an effect.**

- **A lack of significance is not necessarily a lack of an effect**

- **Absence of evidence is not evidence of absence".**

Clinically Important Difference[effect size]

Definition: specifies the difference between two populations parameters that is the meaningful from a clinical perspective .

Two parameters that differ by less than the clinically important difference are assumed equivalent from a biological point of view

To compute the power of the test, one offers an alternative view about the "true" value of the population parameter, assuming that the null hypothesis is false. The effect size is the difference between the true value and the value specified in the null hypothesis.

The size of the difference between the sample mean and the hypothesized population mean increases(this is known as the effect size).

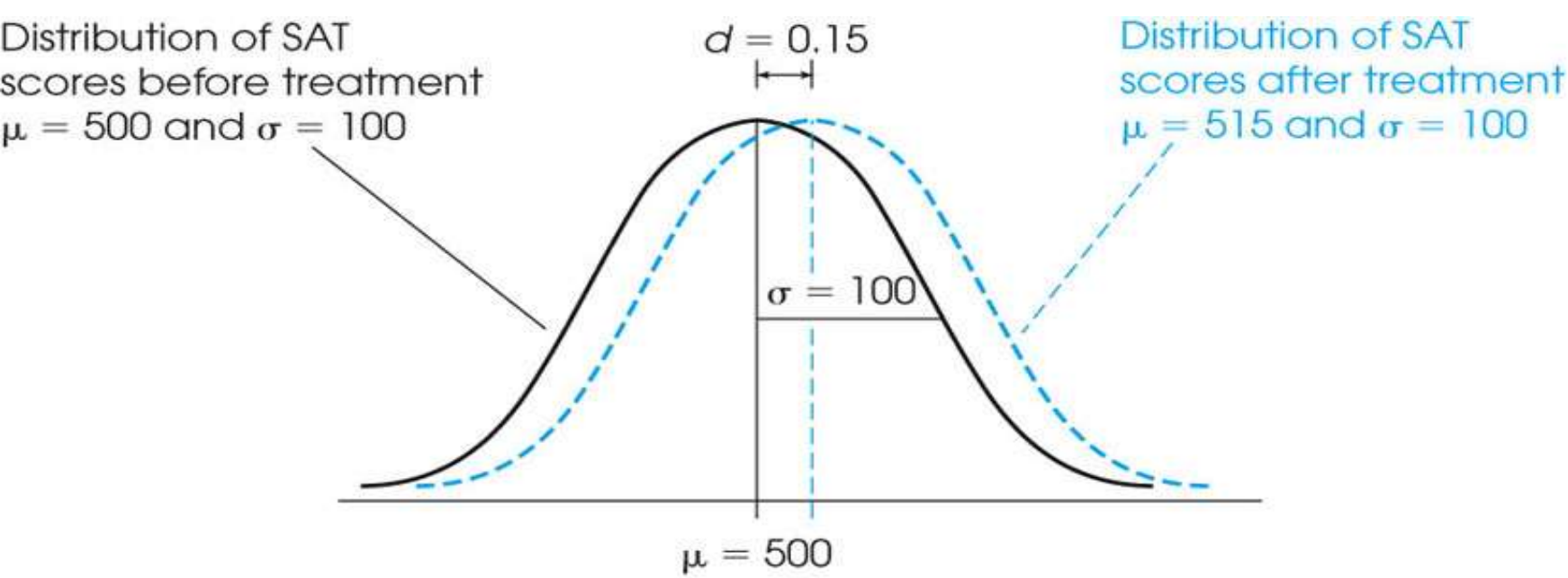
Effect size = True value - Hypothesized value

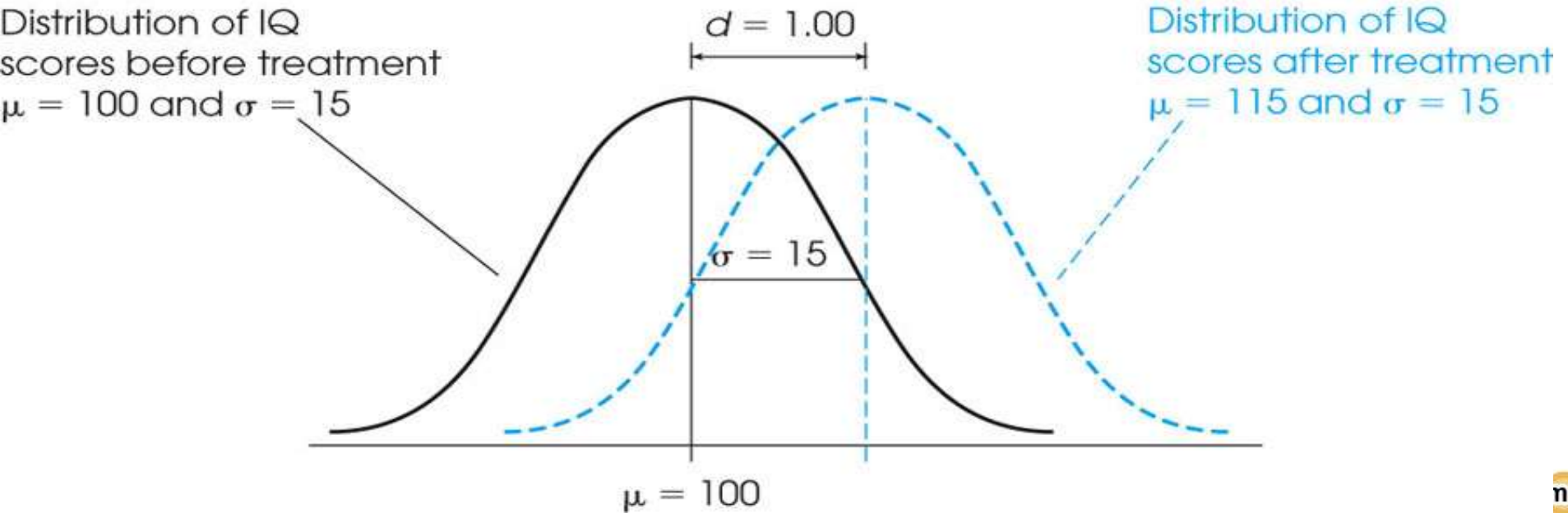
For example, suppose the null hypothesis states that a population mean is equal to 100. A researcher might ask: What is the probability of rejecting the null hypothesis if the true population mean is equal to 90? In this example, the effect size would be 90 - 100, which equals -10

In this example, a difference between a hypothesized
population mean IQ of 135 and a sample mean IQ of 100 would be detected much more easily (and hence the null hypothesis would be rejected more easily) than a difference between a hypothesized IQ of 135 and a sample mean IQ of 128.

Measuring Effect Size

- ✓ **Statistical significance alone does not imply a substantial effect; just one larger than chance**
- ✓ **Cohen's *d* is the most common technique for assessing effect size**
- ✓ **Cohen's *d* = Difference between the means divided by the population standard deviation.**
- ✓ ***d* > .8 means a large effect!**


Distribution of SAT
scores before treatment
μ = 500 and σ = 100
d = 0.15
Distribution of SAT
scores after treatment
μ = 515 and σ = 100
σ = 100
μ = 500


Distribution of IQ
scores before treatment
μ = 100 and σ = 15
d = 1.00
Distribution of IQ
scores after treatment
μ = 115 and σ = 15
σ = 15
μ = 100

***p*-value** = probability that our result (e.g. a difference between proportions or a RR) or more extreme values could be observed under the null hypothesis

The *P* – value Versus The Level α

A conclusion based on a statistical test is typically reported in conjunction with a P- value .

The P-value and α level , while similar in terms of the information they symbolize , have slightly different definitions

A- The P –value represent the actual probability of obtaining the particular sample out come (or one more extreme) from a population for which Ho is true .

That is , it is the probability of a type I error .

The P- value , therefore , varies from sample to sample .

B- The α - level represent the risk of incurring a type I error that the investigator is willing to tolerate .

It is chosen by the investigator and is independent of the data obtained from any given sample

Statistical analysis

Data were entered into SPSS 18 software and analyzed by descriptive statistics (i.e., mean, SD, frequency) and comparison means (i.e., one way ANOVA, t Test, X_2 test).

A P value less than 0.05 was considered statistically significant.

Higher knowledge scores toward Hepatitis B (P=0.03, P=0.0001, P=0.03, P=0.001, P= 0.02, P=0.02 and P=0.0001 respectively), but the relationship between their marital state, family history of hepatitis and history of hepatitis B vaccination and their knowledge toward Hepatitis B was not significant (P>0.05).

Small *p* values = low degree of compatibility between H_0 and the observed data:

→ **you reject H_0, the test is significant**

Large *p* values = high degree of compatibility between H_0 and the observed data:

→ **you don't reject H_0, the test is not significant**

We can never reduce to zero the probability that our result was not observed by chance alone

P-value is the magnitude of chance NOT magnitude of effect

- **P-value < 0.05 = Significant findings**
- **Small chance of being wrong in rejecting the null hypothesis**
- **If in fact there is no [effect], it is unlikely to get the**

[effect] = [magnitude of effect] or more extreme

- **Significance DOES NOT MEAN importance**
- **Any extra-large studies can give a very small P-value even if the [magnitude of effect] is very small.**

P-value > 0.05 = Non-significant findings

- **High chance of being wrong in rejecting the null hypothesis**

- **If in fact there is no [effect], the [effect] = [magnitude of effect] or more extreme can be occurred chance**
- **Non-significance DOES NOT MEAN no difference, equal, or no association**

- **Any small studies can give a very large P-value even if the [magnitude of effect] is very large**

Levels of significance

We need a cut-off !

0.01 **0.05** **0.10**

p value > 0.05 = H_0 non rejected (non significant)

p value ≤ 0.05 = H_0 rejected (significant)

BUT:

Give always the exact p-value rather than „significant“

vs. „non-significant“.

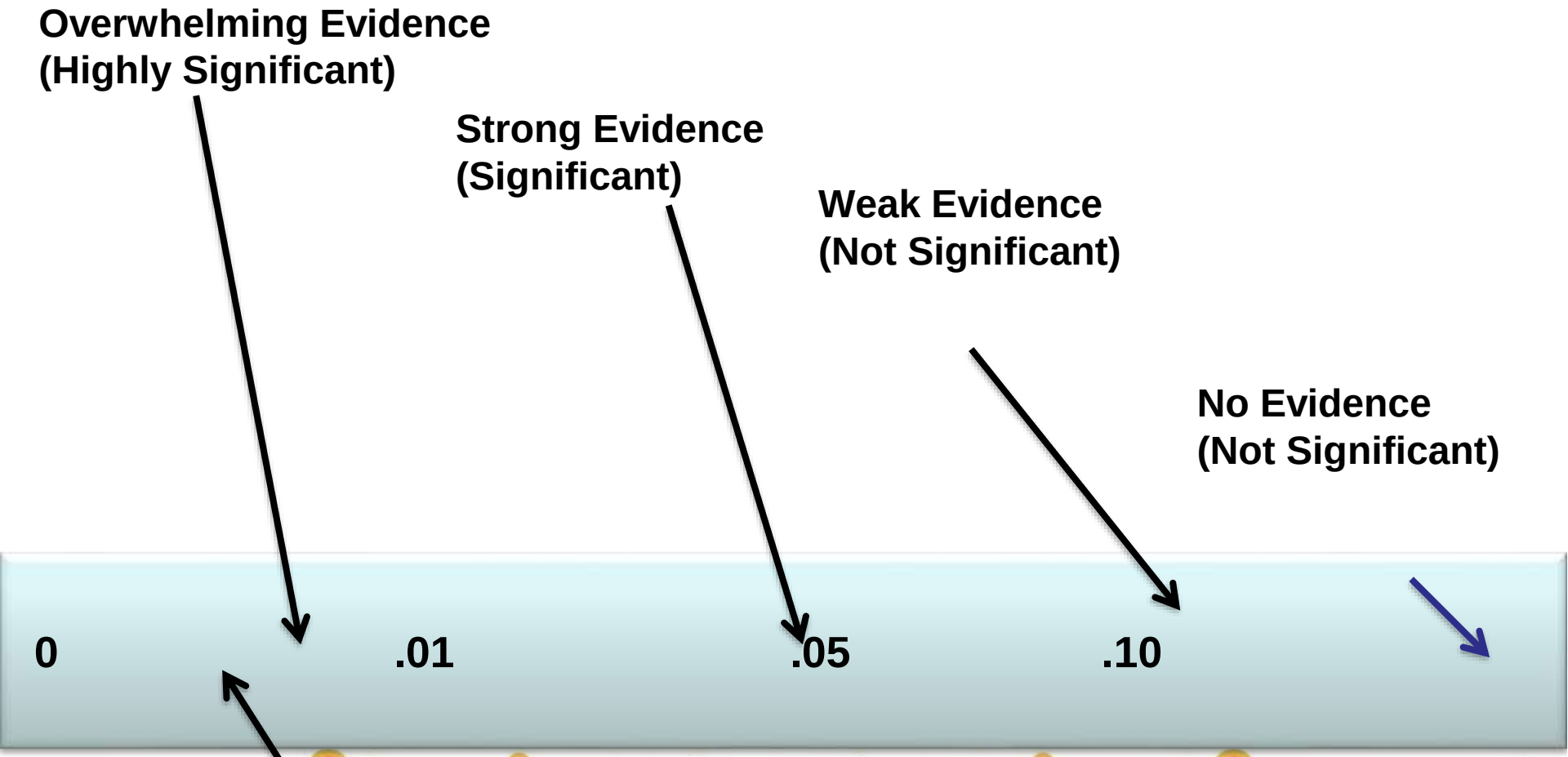
- ” The limit for statistical significance was set at p=0.05.”
- ” There was a strong relationship (p<0.001).”
- ”…, but it did not reach statistical significance (ns).”
- „ The relationship was statistically significant (p=0.0361)”

p=0.05 → Agreed convention

Not an absolute truth

"Surely, God loves the 0.06 nearly as much as the 0.05" (Rosnow and Rosenthal, 1991)

Interpreting the p-value…



As a rough and ready guide , we can think of P- value as indicating the strength of evidence like this:

Greater than 0.1	Little or no evidence of a difference or relationship
Between 0.05& 0.1	Weak evidence of a difference or relationship
Between 0.01 & 0.05	Evidence of a difference or relationship
Less than 0.01	Strong evidence of a difference or relationship
Less than 0.001	Very strong evidence of a difference or relationship

Criticism on significance testing

"Epidemiological application need more than a decision as to whether chance alone could have produced association."

→ Estimation of an effect measure (e.g. RR, OR) rather than significance testing.

The epidemiologist needs measurements rather than probabilities

χ^2 is a test of association

OR, RR are measures of association on a continuous scale
→infinite number of possible values

The best estimate = point estimate

Range of values allowing for random variability:

Confidence interval → precision of the point estimate

What we have to evaluate the study

χ^2 : A test of association. It depends on sample size.

p value : Probability that equal (or more extreme) results can be observed by chance alone

OR, RR: Direction & strength of association
if > 1 risk factor
if < 1 protective factor
(independently from sample size)

CI : Magnitude and precision of effect

P-value vs. 95%CI

- **Always report 95%CI with p-value, NOT report solely p-value**
- **Always interpret based on the lower or upper limit of the confidence interval, p-value can be an optional**
- **Never interpret p-value > 0.05 as an indication of no difference or no association, only the CI can provide this message**

Comments on p values and CIs

- **A huge effect in a small sample or a small effect in a large sample can result in identical *p* values.**

- **A statistical test will always give a significant result if the sample is big enough.**

- ***p* values and CIs do not provide any information on the possibility that the observed association is due to bias or confounding.**