Differences between MEANS and PERCENTAGES and PAIRED ALTERNATIVES

Luigi Greco, M.D., Ph.D. M.Sc. & Francesco Giallauria, M.D., Ph.D.

Department of Translational Medical Sciences

University of Naples Federico II

WHAT WE AIM TO PROPOSE

1. **To use few parameters to describe the data**
2. To evaluate the precision of the Mean
3. To consider Confidence Intervals
4. To formulate statistical hypothesis
5. To evaluate the Sample Size
6. To compare two samples

Numeric Descriptive Statistics

- Measures of central tendency of data are:
  - Mean
  - Median
  - Mode
- Measures of variability of data are :
  - Standard Deviation
  - Interquartile range

Sample Mean

- Most commonly used measure of central tendency
- Best applied in normally distributed continuous data
- Not applicable in categorical data
- Definition:
  - **"Sum of all the values in a sample, divided by the number of values"**

Sample Median

- Used to indicate the “average” in a skewed population
- Often reported with the mean
  - If the mean and the median are the same, sample is normally distributed.
- It is the middle value from an ordered listing of the values
  - If an odd number of values, it is the middle value
  - If even number of values, it is the average of the two middle values.
- Mid-value in interquartile range

Sample Mode

- Infrequently reported as a value in studies
- Is the most common value
- More frequently used to describe the distribution of data
  - Uni-modal, bi-modal, etc.

Interquartile range

- Is the range of data from the 25th percentile to the 75th percentile

- Common component of a box and whiskers plot
  - It is the box, and the line across the box is the median or middle value
  - Rarely, mean will also be displayed

WHAT WE AIM TO PROPOSE

1. To use few parameters to describe the data
2. **To evaluate the precision of the Mean**
3. To consider Confidence Intervals
4. To formulate statistical hypothesis
5. To evaluate the Sample Size
6. To compare two samples

Standard Error

- A fundamental goal of statistical analysis is to estimate a parameter of a population based on a sample

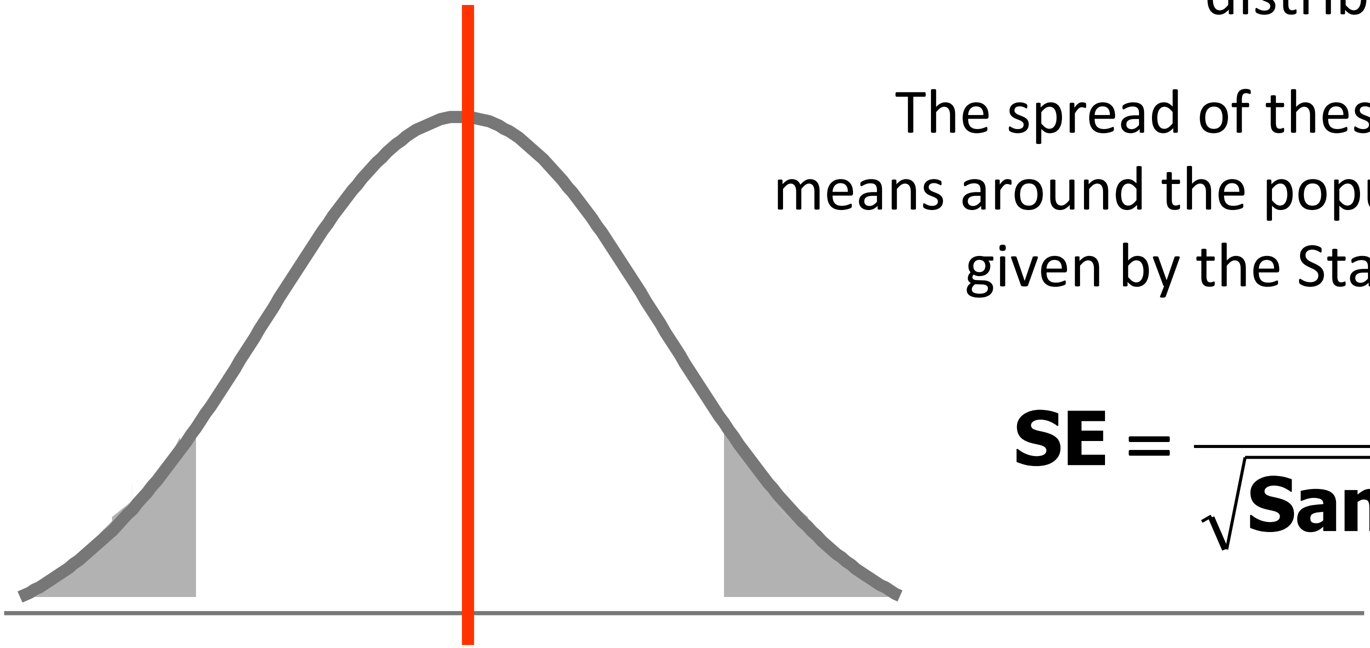
- The values of a specific variable from a sample are an estimate of the entire population of individuals who might have been eligible for the study

- The Standard Error is a measure of the precision of a sample in estimating the population parameter

- If we repeat our measurement in several samples from the same population, we will have several means, which are normally distributed

- The Standard Error estimate the precision of the sample mean

SAMPLES AND POPULATIONS



Repeatedly measuring small samples from the same population will give a normal distribution of means

The spread of these small sample means around the population mean is given by the Standard Error, SE

$$\mathbf{SE} = \frac{\mathbf{SD}}{\sqrt{\mathbf{Sample\ Size}}}$$

Standard Error

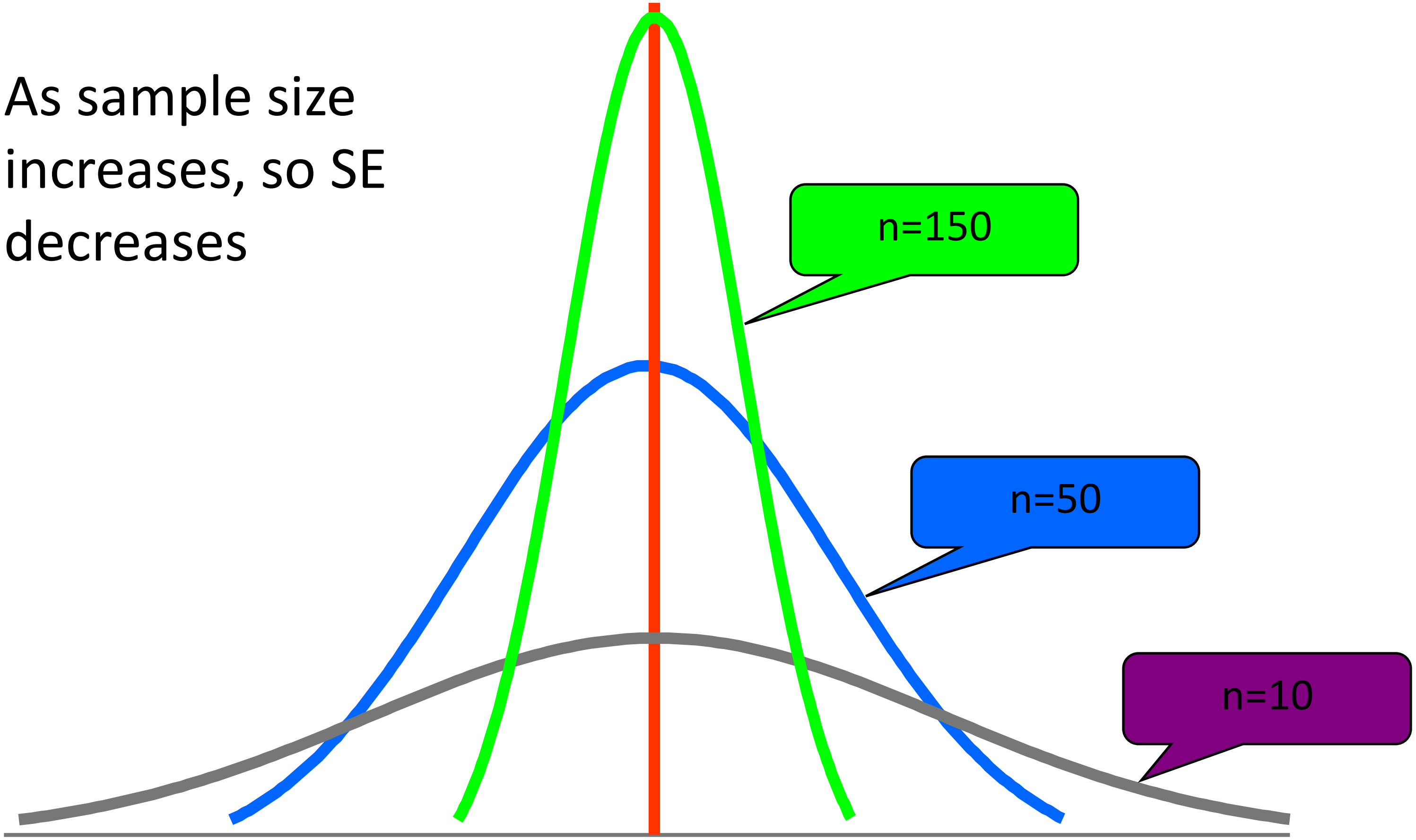
- Standard error of the mean
  - Standard deviation / square root of (sample size)
    - (if sample greater than 60)

- Standard error of the proportion
  - Square root of (proportion X 1 - proportion) / n)

- Important: dependent on sample size
  - Larger the sample, the smaller the standard error

STANDARD ERROR AND SAMPLE SIZE JUST AS THE SAMPLE STANDARD DEVIATION, BUT AT A SMALLER SCALE

As sample size increases, so SE decreases



Clarification

- **Standard Deviation** measures the variability or spread of the data in an individual sample

- **Standard Error** measures the precision of the estimate of a population parameter provided by the sample mean or proportion

Standard Error and ‘Confidence’

- Significance:
  - The Standard Error is the basis of confidence intervals : how much we can be confident that the estimated sample means represent the true population mean ?
  - A 95% confidence interval is defined by
    - **Sample Mean (or proportion) ± 1.96 X Standard Error**
  - Since Standard Error is inversely related to the sample size:
    - The larger the study (sample size), the smaller the confidence intervals and the greater the precision of the estimate

WHAT WE AIM TO PROPOSE

1. To use few parameters to describe the data
2. To evaluate the precision of the Mean
3. **To consider Confidence Intervals**
4. To formulate statistical hypothesis
5. To evaluate the Sample Size
6. To compare two samples

Confidence Intervals

- They show the range of the possible value of the sample mean if we repeat the sampling several times in the same population

- They actually say how robust is this sample mean

- May be used to assess a single point estimate such as mean or proportion

- Most commonly used in assessing the estimate of the difference between two groups

Confidence Intervals in Clinical Practice

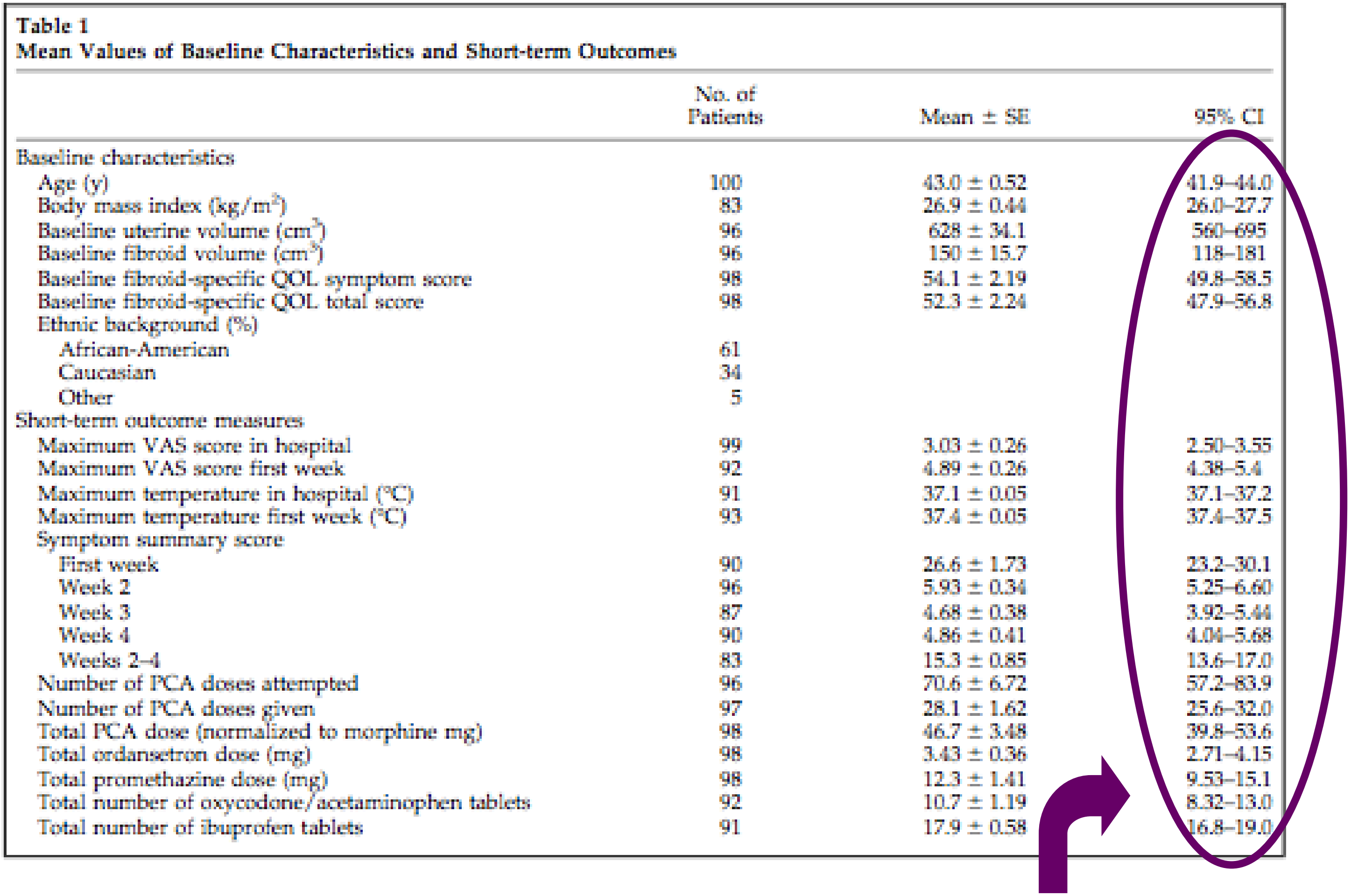


Table 1
Mean Values of Baseline Characteristics and Short-term Outcomes

	No. of Patients	Mean ± SE	95% CI
Baseline characteristics			
Age (y)	100	43.0 ± 0.52	41.9–44.0
Body mass index (kg/m^2)	83	26.9 ± 0.44	26.0–27.7
Baseline uterine volume (cm^3)	96	628 ± 34.1	560–695
Baseline fibroid volume (cm^3)	96	150 ± 15.7	118–181
Baseline fibroid-specific QOL symptom score	98	54.1 ± 2.19	49.8–58.5
Baseline fibroid-specific QOL total score	98	52.3 ± 2.24	47.9–56.8
Ethnic background (%)			
African-American	61		
Caucasian	34		
Other	5		
Short-term outcome measures			
Maximum VAS score in hospital	99	3.03 ± 0.26	2.50–3.55
Maximum VAS score first week	92	4.89 ± 0.26	4.38–5.4
Maximum temperature in hospital (°C)	91	37.1 ± 0.05	37.1–37.2
Maximum temperature first week (°C)	93	37.4 ± 0.05	37.4–37.5
Symptom summary score			
First week	90	26.6 ± 1.73	23.2–30.1
Week 2	96	5.93 ± 0.34	5.25–6.60
Week 3	87	4.68 ± 0.38	3.92–5.44
Week 4	90	4.86 ± 0.41	4.04–5.68
Weeks 2–4	83	15.3 ± 0.85	13.6–17.0
Number of PCA doses attempted	96	70.6 ± 6.72	57.2–83.9
Number of PCA doses given	97	28.1 ± 1.62	25.6–32.0
Total PCA dose (normalized to morphine mg)	98	46.7 ± 3.48	39.8–53.6
Total ordansetron dose (mg)	98	3.43 ± 0.36	2.71–4.15
Total promethazine dose (mg)	98	12.3 ± 1.41	9.53–15.1
Total number of oxycodone/acetaminophen tablets	92	10.7 ± 1.19	8.32–13.0
Total number of ibuprofen tablets	91	17.9 ± 0.58	16.8–19.0

Commonly reported in studies to provide an estimate of the precision of the mean

WHAT WE AIM TO PROPOSE

1. To use few parameters to describe the data
2. To evaluate the precision of the Mean
3. To consider Confidence Intervals
4. **To formulate statistical hypothesis**
5. To evaluate the Sample Size
6. To compare two samples

Formulation of Hypothesis

- Statistics **cannot** say *' The A population is absolutely diffrent from B',* but it suggests *' how much is not probable that A is equal to B'*
- **We may wish to reject the Null Hypothesis (H0) that 'A = B' that says that there is no difference between A and B, those observed are only due to sampling variability**

NULL HYPOTHESIS H0: A=B

- By convention, if the Null Hypothesis H0 has less than 5% chances to be true, we may discard it : we reject it with a 'p' probability less than 5% (< 0,05)
- *The Type I Error (alpha) is the area of the possible results that induce to reject the Null Hypothesis*
- *The Type II Error (beta) is the contrary: is the error to not rejecting the Null Hypothesis H0 whet H0 is false.*
- Its is then the inability to show a real difference between A and B and then accept the Experimental Hypothesis H1

P value: probability of the Null Hypothesis = only chance

- Statistics cannot produce the absolute truth: (i.e. smoke is the cause of cancer), but can estimate how it is not probable that there is no connection between smoke and cancer
- This is the NULL HYPOTHESIS: There is no link between smoke and cancer
- The probability that any observation is due to chance alone assuming that the null hypothesis is true *(Smokers have the same chance of cancer than non-smokers)*
  - Typically, an estimate that has a p value of 0.05 or less is considered to be "statistically significant" or unlikely to occur due to chance alone

  - The P value used is an arbitrary value that tell how small is the chance to observe the experimental hypothesis when it is not true

    - P value of 0.05 equals 1 in 20 chance
    - P value of 0.01 equals 1 in 100 chance
    - P value of 0.001 equals 1 in 1000 chance

P values and Confidence Intervals

- **P values provide less information than confidence intervals**
  - A p value provides only a probability that estimate is due to chance
  - A p value could be statistically significant but of limited clinical significance
    - A very large study might find that a difference of .1 on a VAS Scale of 0 to 10 is statistically significant but it may be of no clinical significance
    - A large study might find many "significant" findings during multivariable analyses.

"a large study dooms you to statistical significance"

Errors

- Type I error : Null Hypothesis
  - Claiming a difference between two samples when in fact there is none
    - Remember there is variability among samples- they might seem to come from different populations but they are not
  - Also called the α (alpha) error
  - Typically 0.05 is used = probability of the Null Hypothesis

Errors

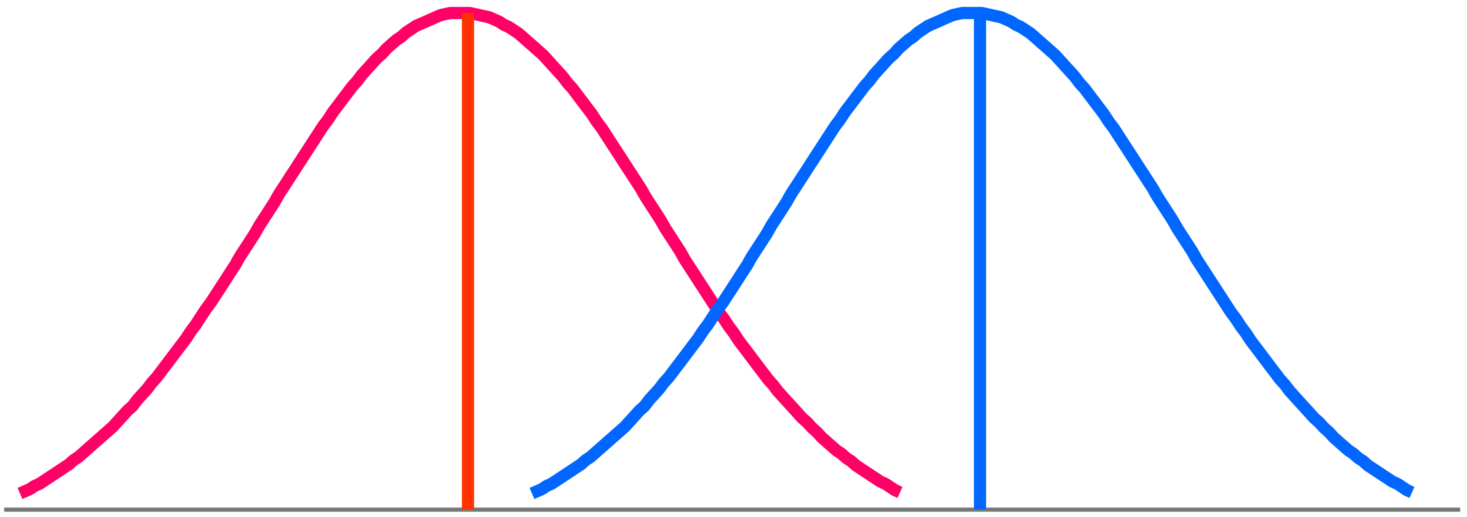
- Type II error : accept the null hypothesis when in fact it is not true *(smoke is related to cancer)*
  - – Claiming there is no difference between two samples when in fact there is
  - – Also called a β error
  - – The probability of not making a Type II error is 1 - β, which is called ***the power of the test***
  - – Hidden error because can’t be detected without a proper power analysis
  - – Type II error is very common when the sample size is not adequate for the design of the study

WHAT WE AIM TO PROPOSE

1. To use few parameters to describe the data
2. To evaluate the precision of the Mean
3. To consider Confidence Intervals
4. To formulate statistical hypothesis
5. **To evaluate the Sample Size**
6. To compare two samples

SAMPLE SIZE: POWER CALCULATIONS

Using the standard α=0.05 and β=0.20, and having estimates for the standard deviation and the difference in sample means, the smallest sample size needed to avoid a Type II error can be calculated with a formula



POWER CALCULATIONS

- Intended to estimate sample size required to prevent Type II errors
- For simplest study designs, can apply a standard formula
- Essential requirements:
  1. **A research hypothesis**
  2. **A measure (or estimate) of variability for the outcome measure**
  3. **The difference (between intervention and control groups) that would be considered clinically important**

Sample Size Calculation

- Also called "**power analysis**"
- When designing a study, one needs to determine how large a study is needed
- **Power** is the ability of a study to avoid a Type II error
- Sample size calculation yields the number of study subjects needed, given a certain desired power to detect a difference and a certain level of P value that will be considered significant

Please note:

- **Many studies are completed without proper estimate of appropriate study size**
- **This may lead to a "negative" study outcome in error**

Sample Size Calculation

- Depends on:
  - **Level of Type I error**: 0.05 typical
  - **Level of Type II error**: 0.20 typical
  - One sided *vs.* two sided: nearly always two
  - Inherent variability of population
    - Usually estimated from preliminary data
  - The difference that would be meaningful between the two assessment arms

Sample Size Calculation: Precision based sample size calculation

Suppose you want to be able to estimate your unknown parameter with a certain degree of precision. What you are essentially saying is that you want your confidence interval to be a certain width

In general a 95% confidence interval is given by the formula:

$$\boxed{\text{Estimate} \pm 2(\text{approx})^{1} \times \text{SE}}$$

where SE is the standard error of whatever you are estimating

This is because 95% confidence intervals are usually based on the normal distribution or a t-distribution - for a normal distribution the value is 1.96; for t-distributions the value is generally just over 2

The formula for any standard error always contains n, the sample size

Therefore, if you specify the width of the 95% confidence interval, you have a formula that you can solve to find n

Sample Size Calculation: Precision based sample size calculation - Example

Suppose you wish to carry out a trial of a new treatment for hypertension (high blood pressure) among men aged between 50 and 60. You randomly select 2n subjects

n of these receive the new treatment and n receive a the standard treatment, then you measure each subject’s systolic blood pressure

You will analyse your data by comparing the mean blood pressure in the two groups — i.e. carrying out an unpaired t-test and calculating a 95% confidence interval for the true difference in means

Sample Size Calculation: Precision based sample size calculation - Example

You would like your 95% confidence interval to have width 10 mmHg (i.e. you want to be 95% sure that the true difference in means is within ± 5 mmHg of your estimated difference in means)

- **How many subjects will you need to include in your study?**

We know that the 95% confidence interval for a difference in means is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm 2(\text{approx}) \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad *$$

Hence, we want $2 \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ to be equal to $5 \Rightarrow s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = s_p\sqrt{\frac{2}{n}} \approx 2.5$ (since we are aiming for groups of the same size)

* s_p = standard deviation of the sampling distribution

Sample Size Calculation: Precision based sample size calculation - Example

In order to work out our sample sizes we therefore need to know what s_p (standard deviation of the sampling distribution) is likely to be

This is either known from:

(a) previous experience (i.e. knowledge of the distribution of systolic blood pressure among men with hypertension in this age group)

(b) using other published papers on blood pressure studies in a similar group of people or

(c) carrying out a pilot study

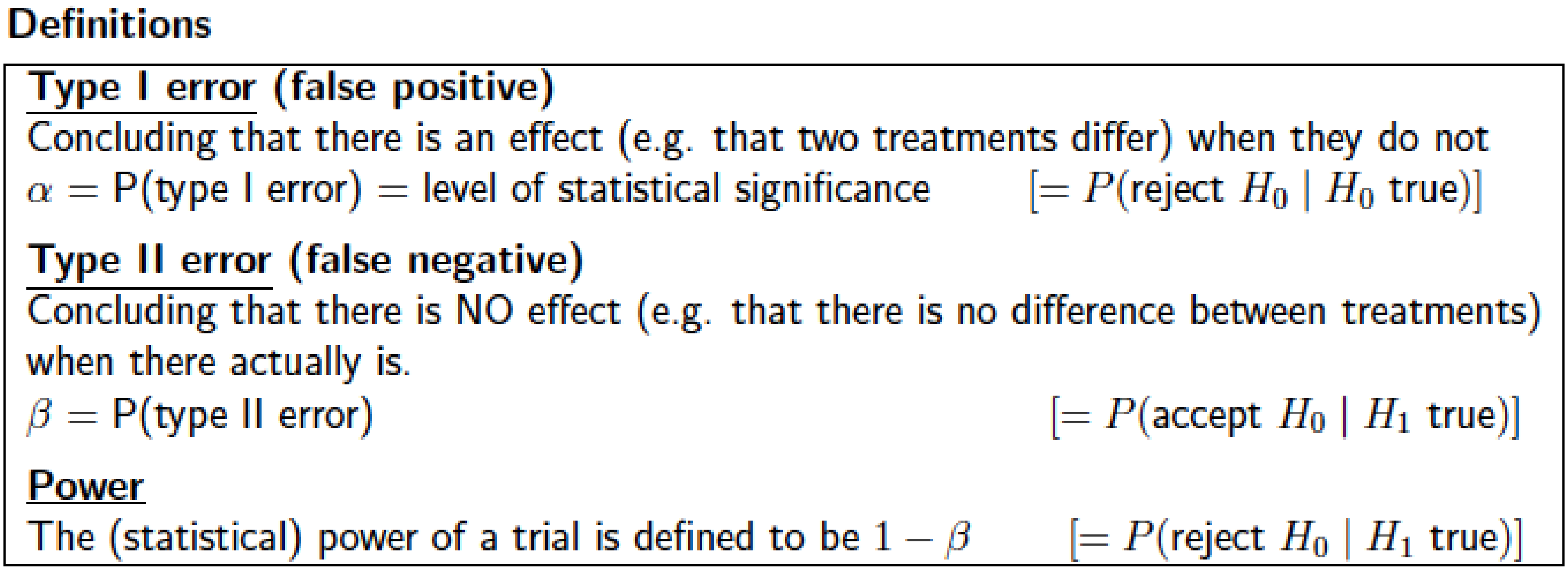
I have used option (b) to get a likely value for s_p of 20 mmHg!

This gives:

$$\boxed{2.5 = 20\sqrt{\frac{2}{n}} \Rightarrow \frac{n}{2} = \left(\frac{20}{2.5}\right)^2 \Rightarrow n = 128 \qquad \text{(in each group)}}$$

Power based Sample Size Calculation

We have seen above that precision-based sample size calculations relate to estimation. Power based sample size calculations, on the other hand, relate to hypothesis testing. In this handout, the formulae for power-based sample size calculations will not be derived, just presented



Definitions

Type I error (false positive)

Concluding that there is an effect (e.g. that two treatments differ) when they do not

α = P(type I error) = level of statistical significance $[= P(\text{reject } H_0 \mid H_0 \text{ true})]$

Type II error (false negative)

Concluding that there is NO effect (e.g. that there is no difference between treatments) when there actually is.

β = P(type II error) $[= P(\text{accept } H_0 \mid H_1 \text{ true})]$

Power

The (statistical) power of a trial is defined to be $1 - \beta$ $[= P(\text{reject } H_0 \mid H_1 \text{ true})]$

Sample Size Calculation: Precision based sample size calculation - Example

If you wanted your true difference in means to be within ±2.5 mmHg rather than ±5 mmHg of your estimate, this would become

$$\boxed{\frac{n}{2} = \left(\frac{20}{1.25}\right)^2 \Rightarrow n = 512}$$

i.e. if you want to increase your precision by a factor of 2, you have to increase your sample size by a factor of 4. In general, if you want to increase your precision by a factor k, you will need to increase your sample size by a factor k^2

This applies across the board — i.e. whether you are estimating a proportion, a mean, a difference in means, etc. etc.

Power Calculations: quantitative data - Example

Suppose you want to compare the mean in one group to the mean in another (i.e. carry out an unpaired t-test)

The number, n, required in each group is given by:

$$\boxed{n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2}}$$

Where:

α is the significance level (using a two-sided test) — i.e. your cut-off for regarding the result as statistically significant.

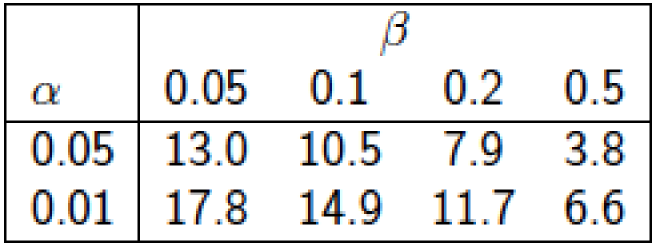
1 − β is the power of your test.

f(α, β) is a value calculated from α and β — see table below.

δ is the smallest difference in means that you regard as being important to be able to detect.

s is the standard deviation of whatever it is we're measuring — this will need to be estimated from previous studies.

$f(\alpha, \beta)$ for the most commonly used values for α and β



	β			
α	0.05	0.1	0.2	0.5
0.05	13.0	10.5	7.9	3.8
0.01	17.8	14.9	11.7	6.6

Power Calculations: quantitative data - Example

Suppose we want to be 90% sure of detecting a difference in mean blood pressure of 10 mmHg as significant at the 5% level (i.e. power = 0.9, β = 0.1, α = 0.05)
We have, from above, s = 20 mmHg
Using the table, we get f(α, β) = 10.5. This gives:

$$\boxed{n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2} = 10.5 \cdot \frac{2(20)^2}{10^2} = 84}$$

You would need 84 subjects in each group

Obviously, if you increase the power or want to use a lower value for α as your cut-off for statistical significance, you will need to increase the sample size

Power Calculations: categorical data

Suppose we are comparing a binary outcome in two groups of size n

Let **p1** = proportion of events (deaths/responses/recoveries etc.) in one group; **p2** = proportion of events in the other group

We need to choose a value for p1 − p2, the smallest practically important difference in proportions that we would like to detect (as significant)

We also need to have some estimate of the proportion of events expected. This can often be obtained from routinely collected data or previous studies

The number of subjects required for each group is given by:

$$\boxed{n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \cdot f(\alpha, \beta)}$$

Power Calculations: categorical data - example

A new treatment has been developed for patients who've had a heart attack. It is known that 10% of people who've suffered from a heart attack die within one year

It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect

Again we will use α = 0.05 and β = 0.1

We have p1 = proportion of deaths in placebo group = 0.1, p2 = proportion of deaths in treatment group = 0.05

This gives:

$$\boxed{n = \frac{0.1(0.9) + 0.05(0.95)}{(0.1 - 0.05)^2} \cdot 10.5 = 578}$$

$$\boxed{n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \cdot f(\alpha, \beta)}$$

STATISTICAL SIGNIFICANCE IS NOT NECESSARILY CLINICAL SIGNIFICANCE : very large samples produce significant results of no biological value

Sample Size	Population Mean	Sample 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

CONSIDER CLINICALLY SIGNIFICANT IMPROVEMENT

Large proportion of patients improving	Hugdahl & Ost (1981)
A change which is large in magnitude	Barlow (1981)
An improvement in patients’ everyday functioning	Kazdin & Wilson (1978)
Reduction in symptoms by 50% or more	Jansson & Ost (1982)
Elimination of the presenting problem	Kazdin & Wilson (1978)

One-sided *vs.* Two-sided

- Most tests should be framed as a two-sided test
  - When comparing two samples, we usually cannot be sure which is going to be better
    - You never know which directions study results will go!
  - **For routine medical research, use only two-sided tests**

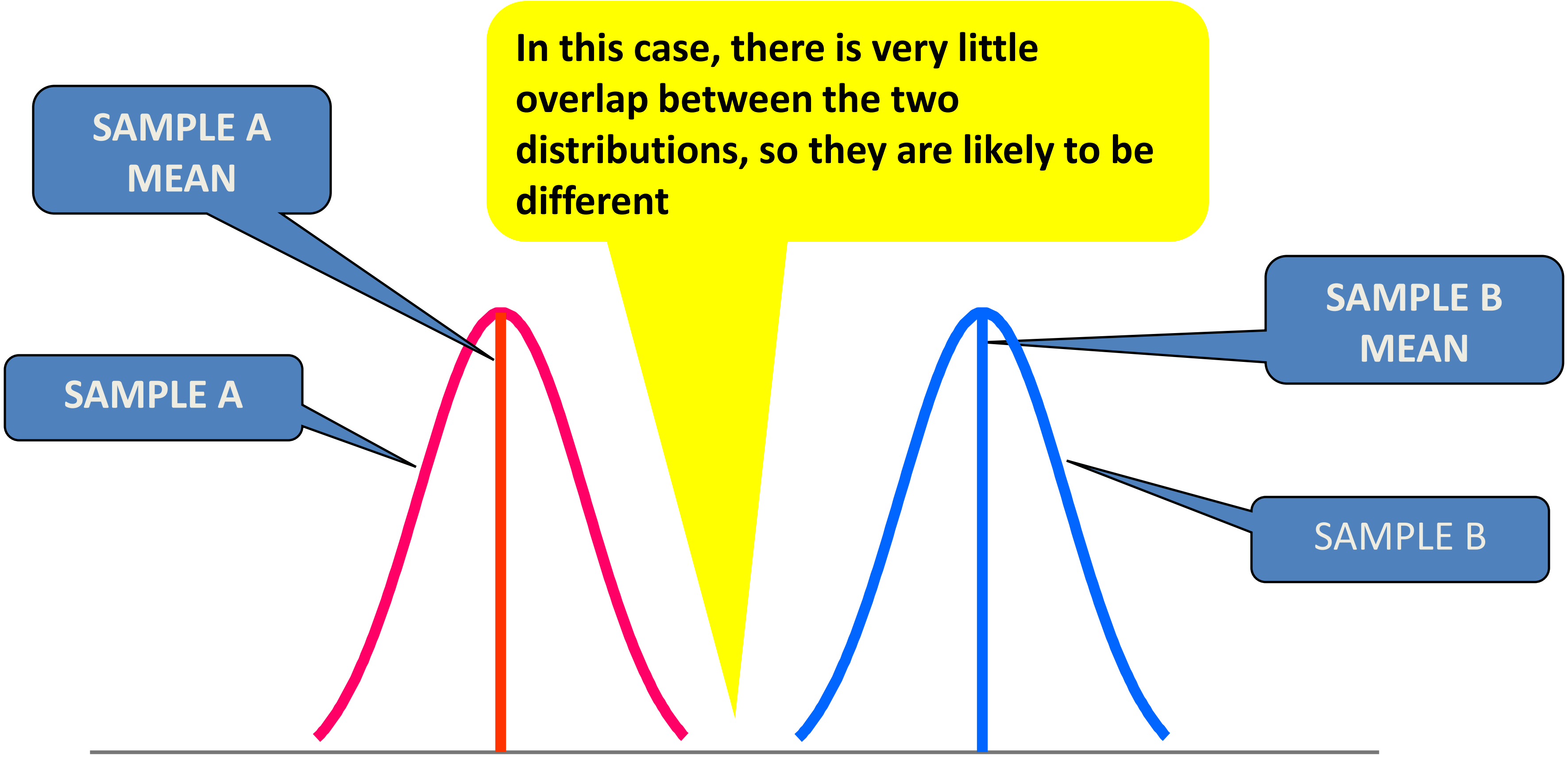
Statistical Tests

- **Parametric tests**
  - Continuous data normally distributed
- **Non-parametric tests**
  - Continuous data not normally distributed
  - Categorical or Ordinal data

WHAT WE AIM TO PROPOSE

1. To use few parameters to describe the data
2. To evaluate the precision of the Mean
3. To consider Confidence Intervals
4. To formulate statistical hypothesis
5. To evaluate the Sample Size
6. **To compare two samples**

COMPARING TWO SAMPLES

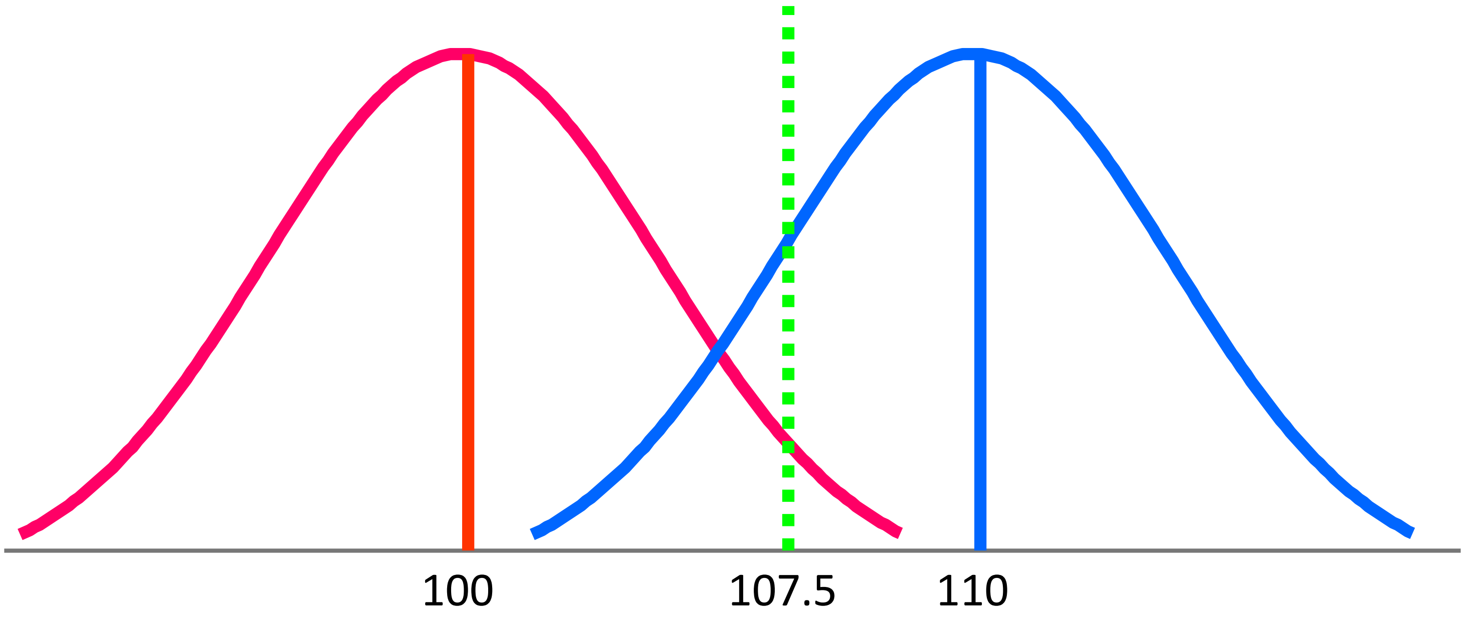


In this case, there is very little overlap between the two distributions, so they are likely to be different

COMPARING TWO SAMPLES

We are comparing the IQ Values in two groups of school children.
The red bell shows the distribution of the 'standard' sample and the blue bell the distribution of the test sample (*very gifted*)

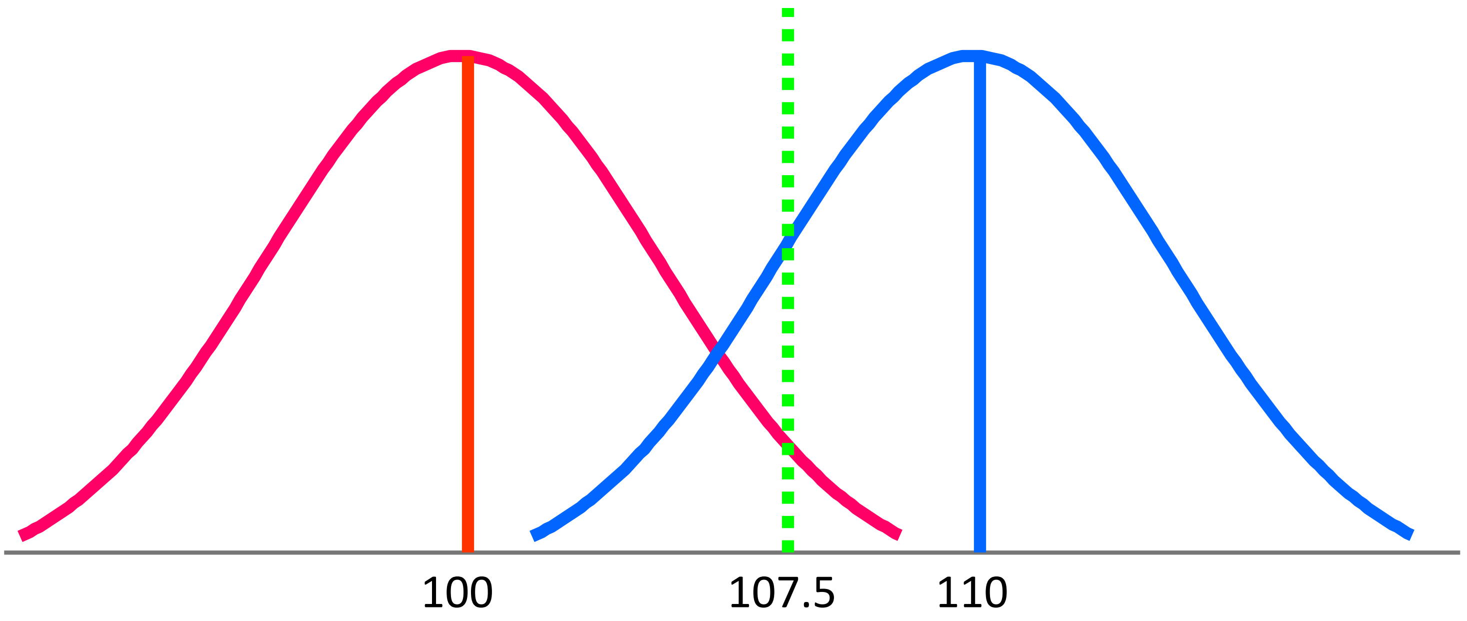
One child has a value of (IQ=107.5): we want to know if he actually came from a population with a mean IQ of 110 *(gifted)* or from the 'original' population with mean = 100



COMPARING TWO SAMPLES

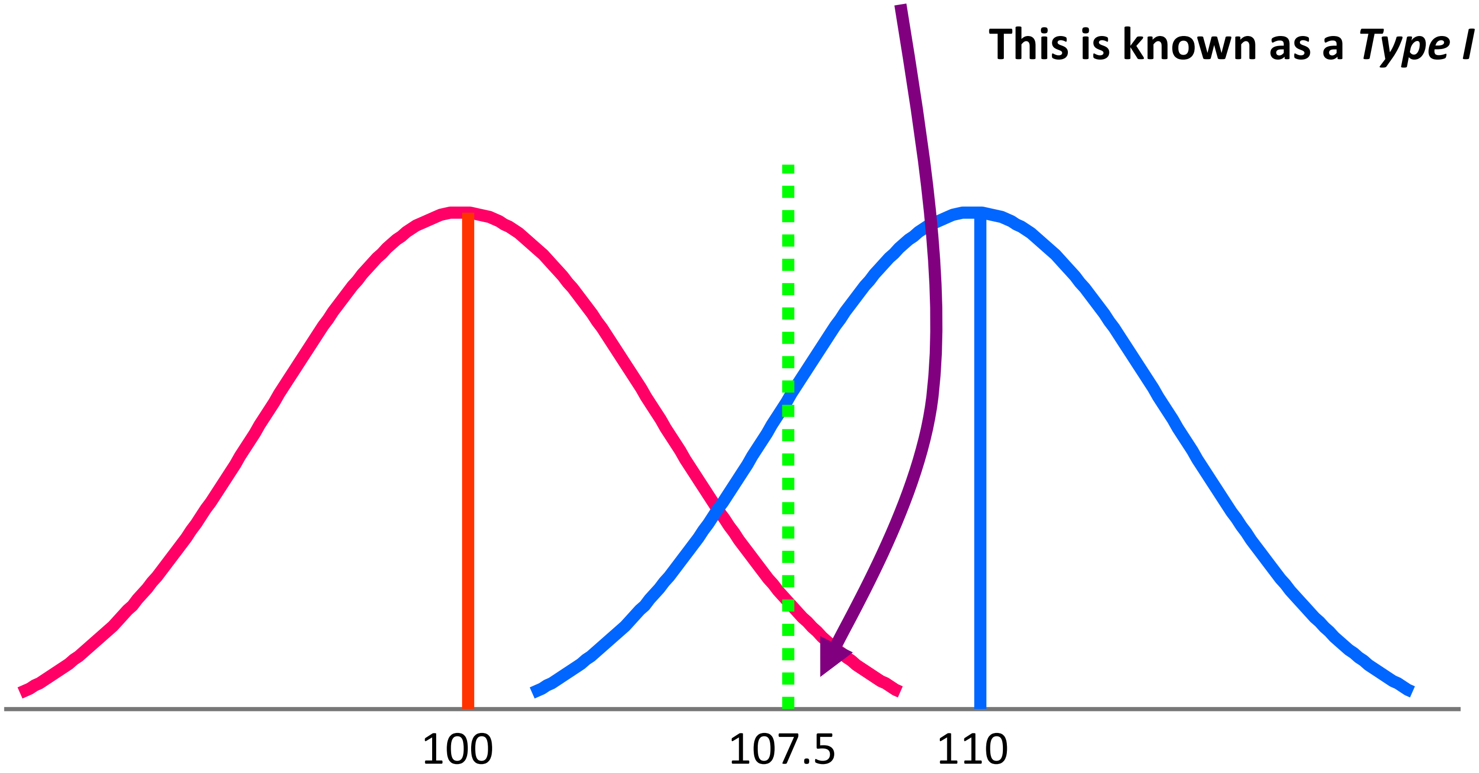
We start by assuming that our sample came from the original population

Our *null hypothesis* (to be tested) is that IQ=107.5 is not significantly different from IQ=100



COMPARING TWO SAMPLES

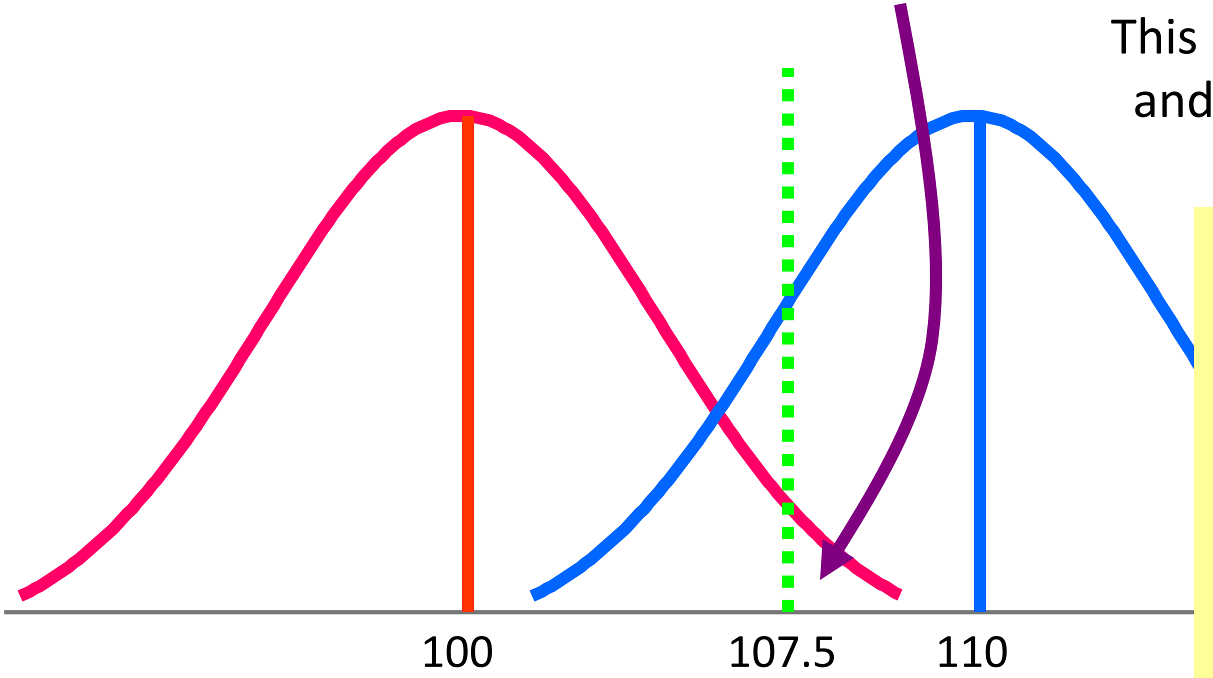
The α level represents the probability of finding a significant difference between the two means when none exists



This is known as a *Type I error*

COMPARING TWO SAMPLES

The area under the ‘standard population’ curve to the right of our sample IQ of 107.5 represents the likelihood of observing this sample mean of 107.5 by chance under the null hypothesis i.e. that the sample is from the ‘standard population’

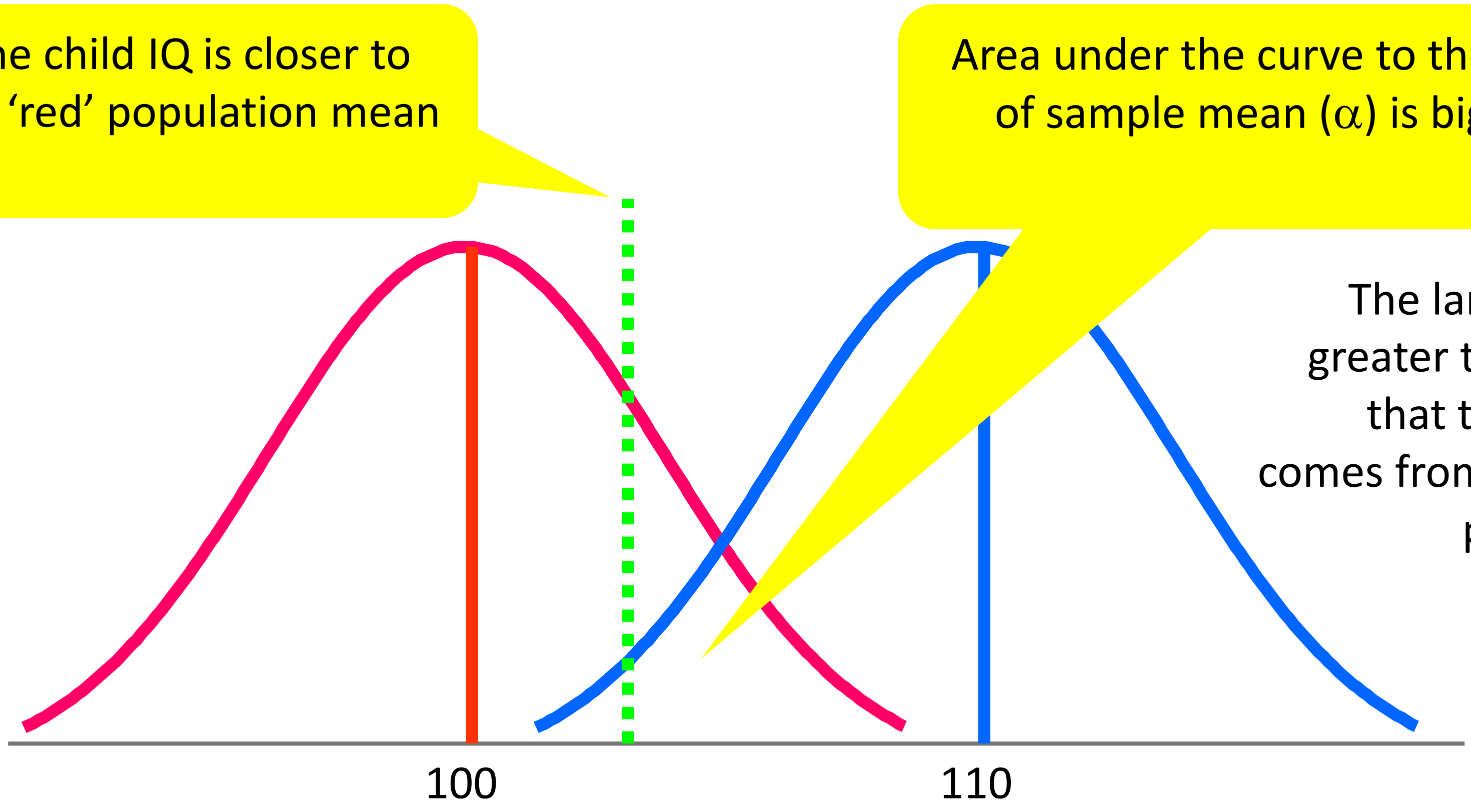


This is known as the α level and is normally set at 0.05

If the sample comes from the standard population, we expect to find a mean of 107.5 in 1 out of 20 estimates

COMPARING TWO SAMPLES

It is perhaps easier to conceptualise α by seeing what happens if we move the sample mean



The child IQ is closer to the 'red' population mean

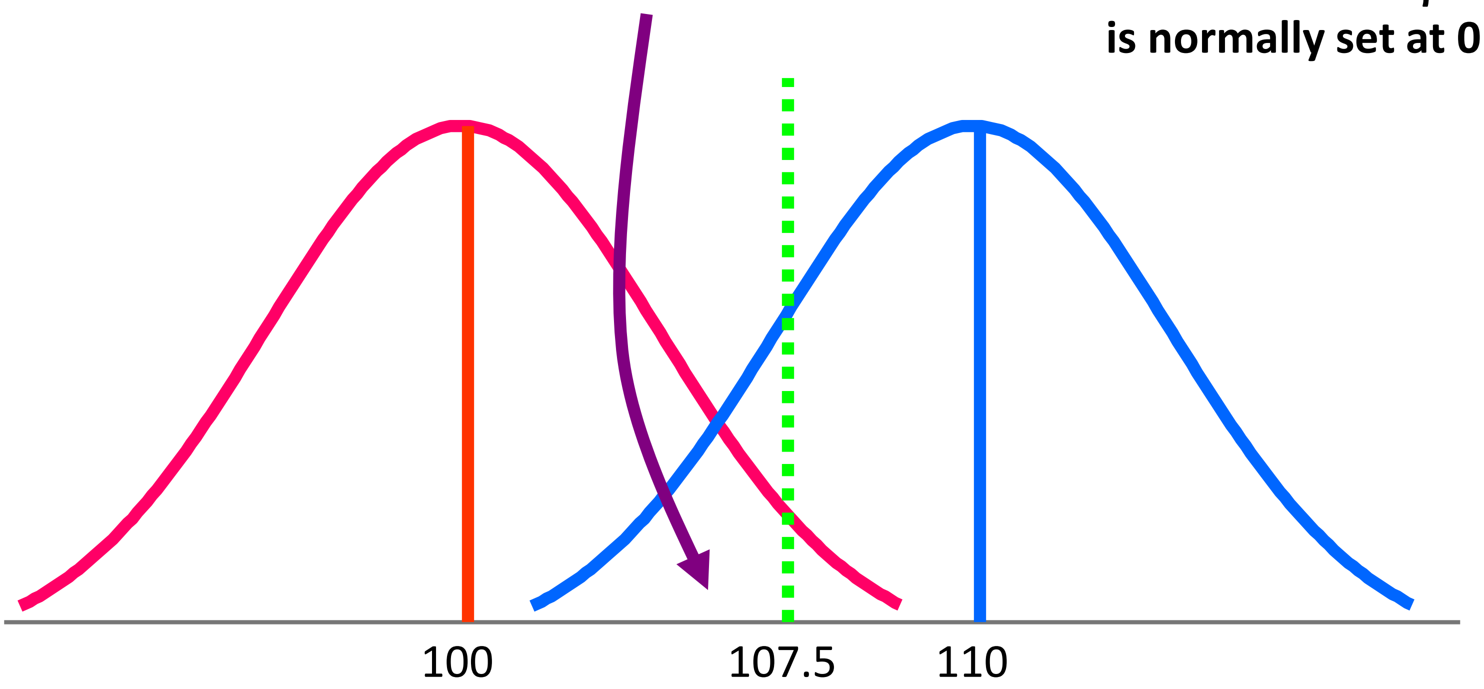
Area under the curve to the right of sample mean (α) is bigger

The larger α, the greater the chance that the sample comes from the 'Red' population

COMPARING TWO SAMPLES

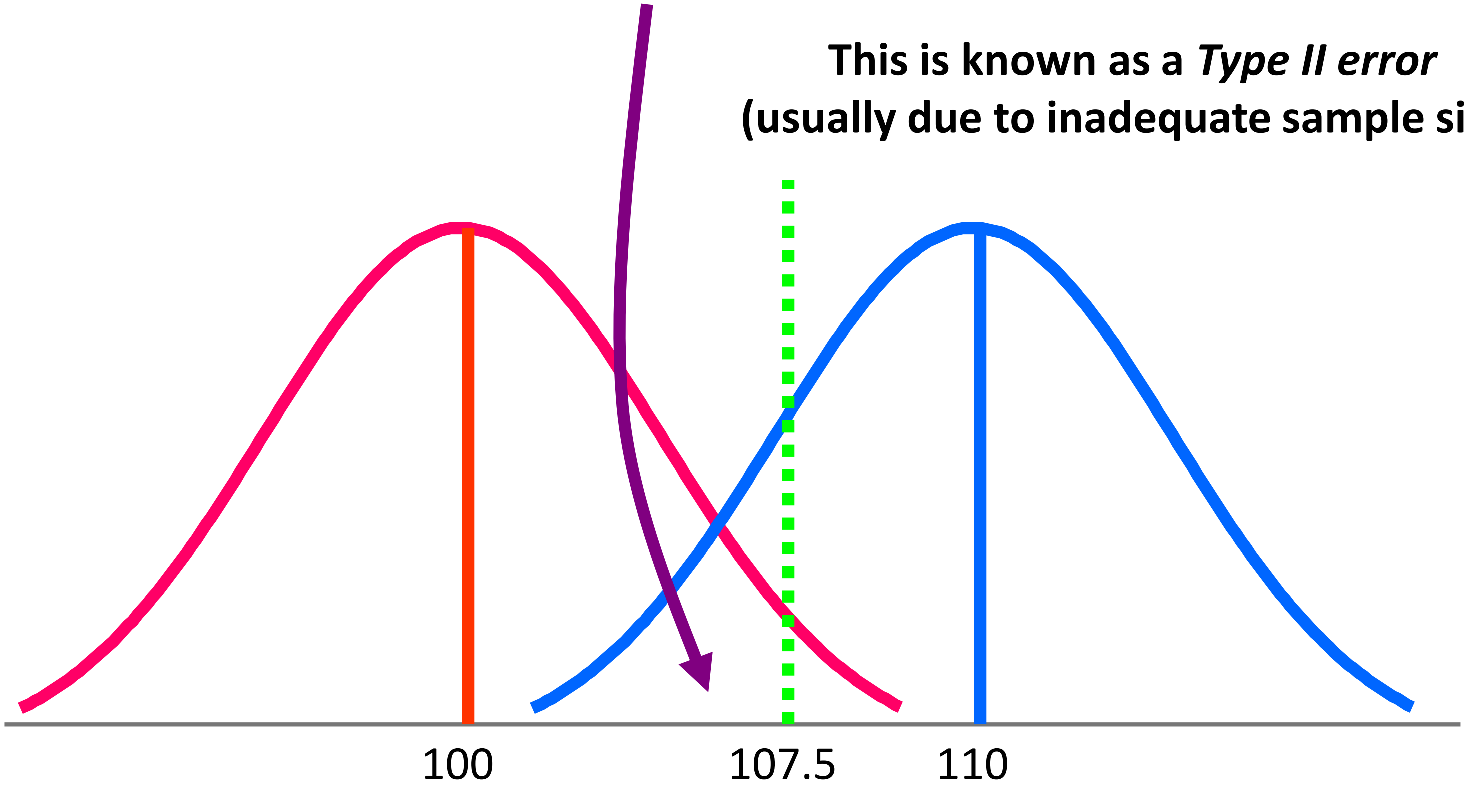
The area under the ‘other population’ curve (blue) to the left of our sample IQ of 107.5 represents the likelihood of observing this sample mean of 107.5 by chance under the alternative hypothesis (that the sample is from the ‘other population’)

This is known as the β level and is normally set at 0.20



COMPARING TWO SAMPLES

The β level represents the probability of not finding a significant difference between the two means when one exists

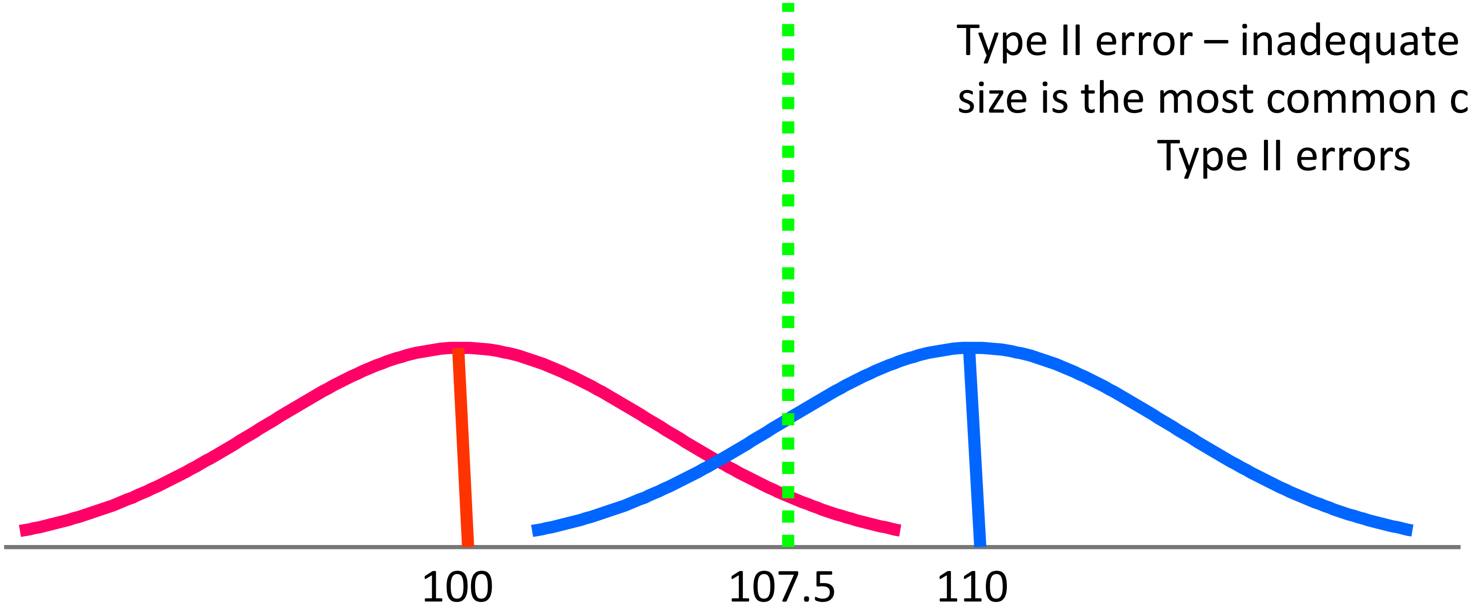


This is known as a *Type II error*
(usually due to inadequate sample size)

COMPARING TWO SAMPLES

Note that if the population sizes are reduced, the standard error (SE) increases, and so does β (hence also the probability of failing to find a significant difference between the two means)

This increases the likelihood of a Type II error – inadequate sample size is the most common cause of Type II errors

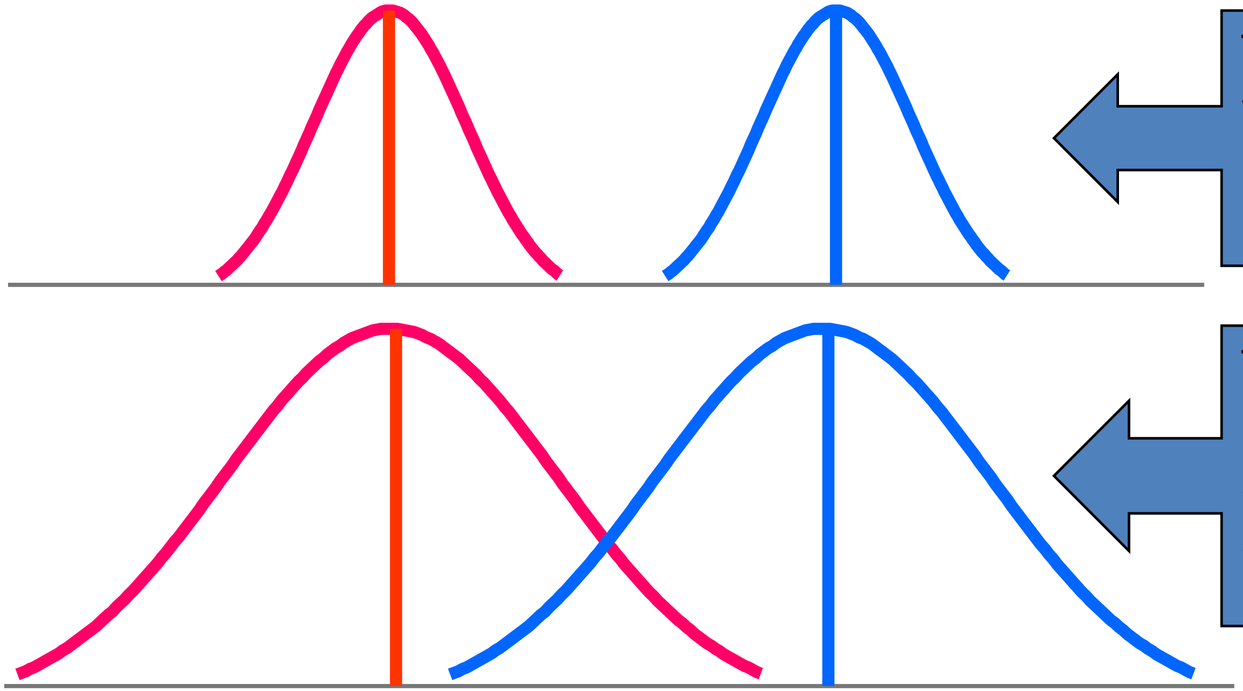


STATISTICAL ERRORS: SUMMARY

Type I (α)	● **'False positive'** ● **Find a significant difference even though one does not exist** ● **Usually set at 0.05 (5%) or 0.01 (1%)**
Type II (β)	● **'False negative'** ● **Fail to find a significant difference even though one exists** ● **Usually set at 0.20 (20%)** ● **Power = 1 – β (i.e. usually 80%)**

Remember that *Power* is related to sample size because a larger sample has a smaller SE thus there is less overlap between the curves

UNPAIRED OR INDEPENDENT-SAMPLE t-TEST: PRINCIPLE



The two distributions are widely separated so their means clearly different

The distributions overlap, so it is unclear whether the samples come from the same population

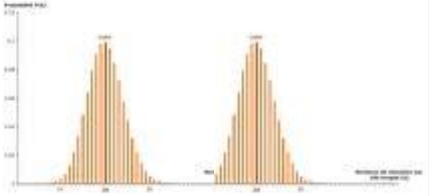
$$t = \frac{\textbf{Difference between means}}{\textbf{SE of the difference}}$$

In essence, the t-test gives a measure of the difference between the sample means in relation to the overall spread

Student t Test to compare two means

It is just the difference of the mean of A and the mean of B by their variance

$$t = \frac{(mA\text{-}mB)}{S}$$

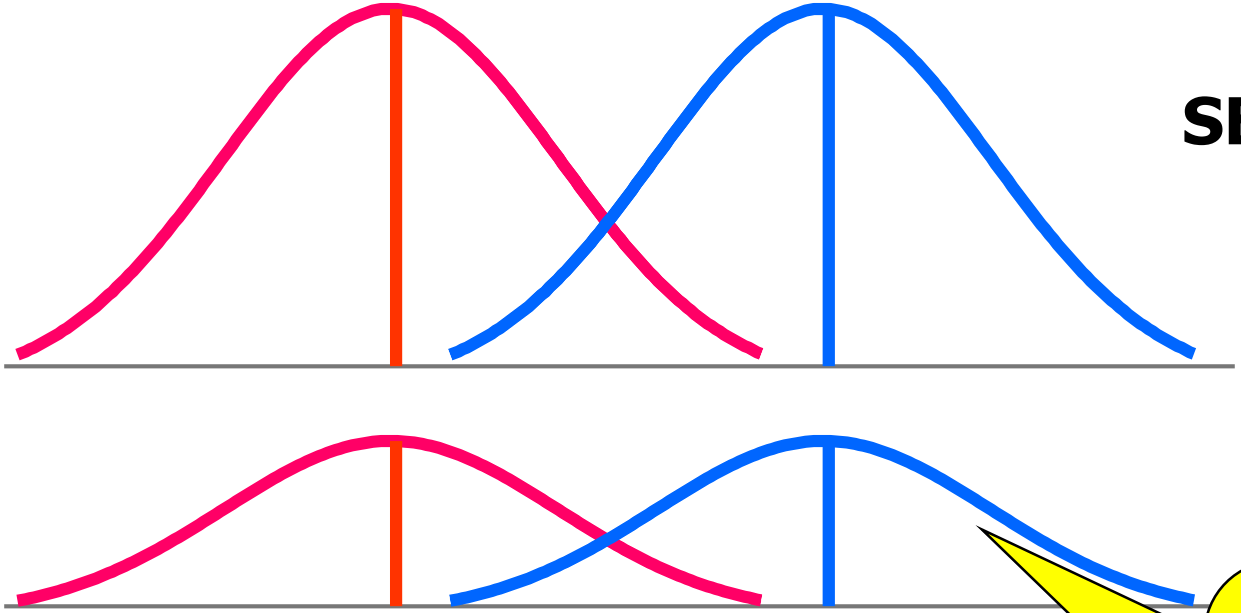


$$S = \sqrt{(\frac{\text{Deviance A+Deviance B)}}{nA+nB-2}}$$

For paired data we just estimate the mean of the difference in each pair A/B

$$t = d \cdot \sqrt{\frac{\text{DS diff.}^2}{n}}$$

UNPAIRED OF INDEPENDENT-SAMPLE t-TEST: PRINCIPLE



$$SE = \frac{SD}{\sqrt{\text{Sample Size}}}$$

$$t = \frac{\text{Difference between means}}{\text{SE of the difference}}$$

With smaller sample sizes, SE increases, as does the overlap between the two curves, so value of t decreases

Comparison of 2 Sample Means

- Student’s T test
  - Assumes normally distributed continuous data

$$\text{T value} = \frac{\text{difference between means}}{\text{standard error of difference}}$$

- T value then looked up in the Table to determine significance

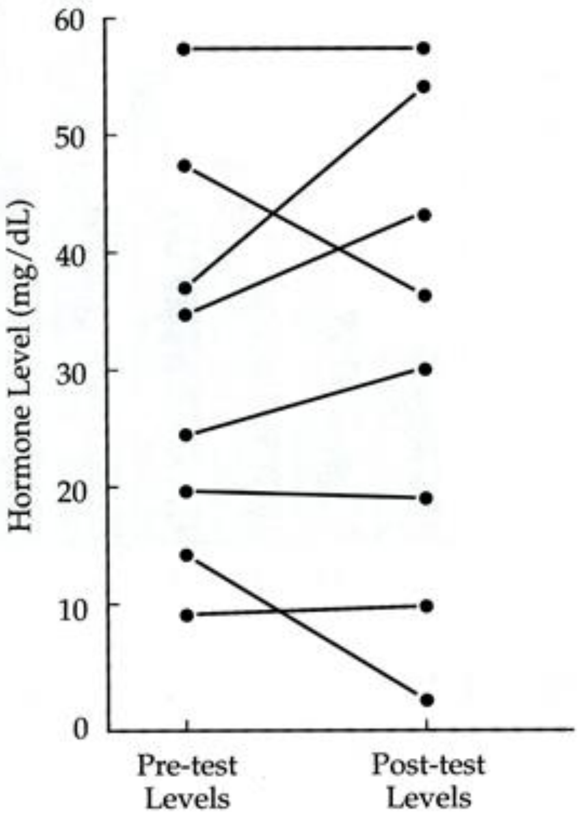
COMPARING TWO MEANS FROM THE SAME SAMPLE-THE PAIRED t TEST

Subject	A	B
1	10	11
2	0	3
3	60	65
4	27	31

- Assume that A and B represent measures on the same subject (i.e. at two time points)
- Note that the variation between subjects is much wider than that within subjects i.e. the variance in the columns swamps the variance in the rows
- Treating A and B as entirely separate, t=-0.17, p=0.89
- Treating the values as paired, t=3.81, p=0.03

Paired T Tests

- Uses the change before and after intervention in a single individual
- Reduces the degree of variability between the groups
- Given the same number of patients, has greater power to detect a difference between groups



MULTIPLE TESTS AND TYPE I ERRORS

- The risk of observing by chance a difference between two means (even if there isn't one) is α
- This risk is termed a Type I error
- By convention, α is set at 0.05
- For an individual test, this becomes the familiar $p<0.05$ (the probability of finding this difference by chance is <0.05 or less than 1 in 20)
- However, as the number of tests rises, the actual probability of finding a difference by chance rises markedly

Tests (N)	p
1	0.05
2	0.098
3	0.143
4	0.185
5	0.226
6	0.264
10	0.401
20	0.641

SUBGROUP ANALYSIS

- Papers sometimes report analyses of subgroups of their total dataset
- Criteria for subgroup analysis:
  - ✓ Must have large sample
  - ✓ Must have *a priori* hypothesis
  - ✓ Must adjust for baseline differences between subgroups
  - ✓ Must retest analyses in an independent sample

TORTURED DATA - SIGNS

Statistical methods do not protect from over enthusiastic interpretation of results!

Consider:

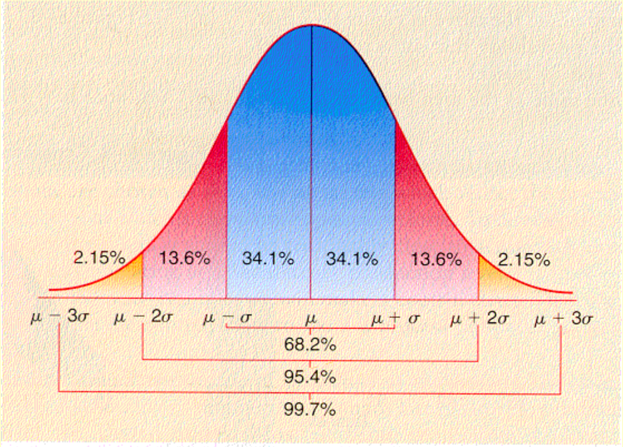
- How many statistical tests were performed, and was the effect of multiple comparisons dealt with appropriately?
- Are both P values and confidence intervals reported?
- Have the data been reported for all subgroups and at all follow-up points?
- What was the rationale for excluding various subjects from the analysis?

Mills JL. Data torturing. *NEJM* 329:1196-1199, 1993.

Test statistici non-parametrici

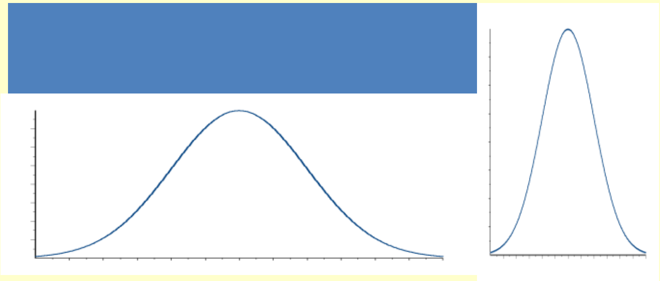
Il test t di Student e l'ANOVA sono basati su alcune assunzioni…

1. Variabili continue o almeno misurate in un intervallo (es. non conosco il valore assoluto, ma posso quantificare le differenze fra due valori)

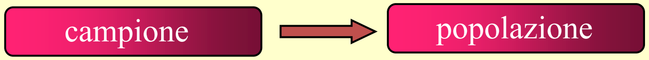


2. Indipendenza fra media e varianza (l'errore di misura deve essere indipendente dal valore misurato)

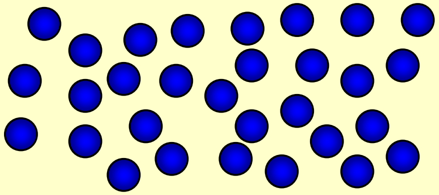
3. Variabili distribuite in modo (approssimativamente) normale



4. Omogeneità delle varianze

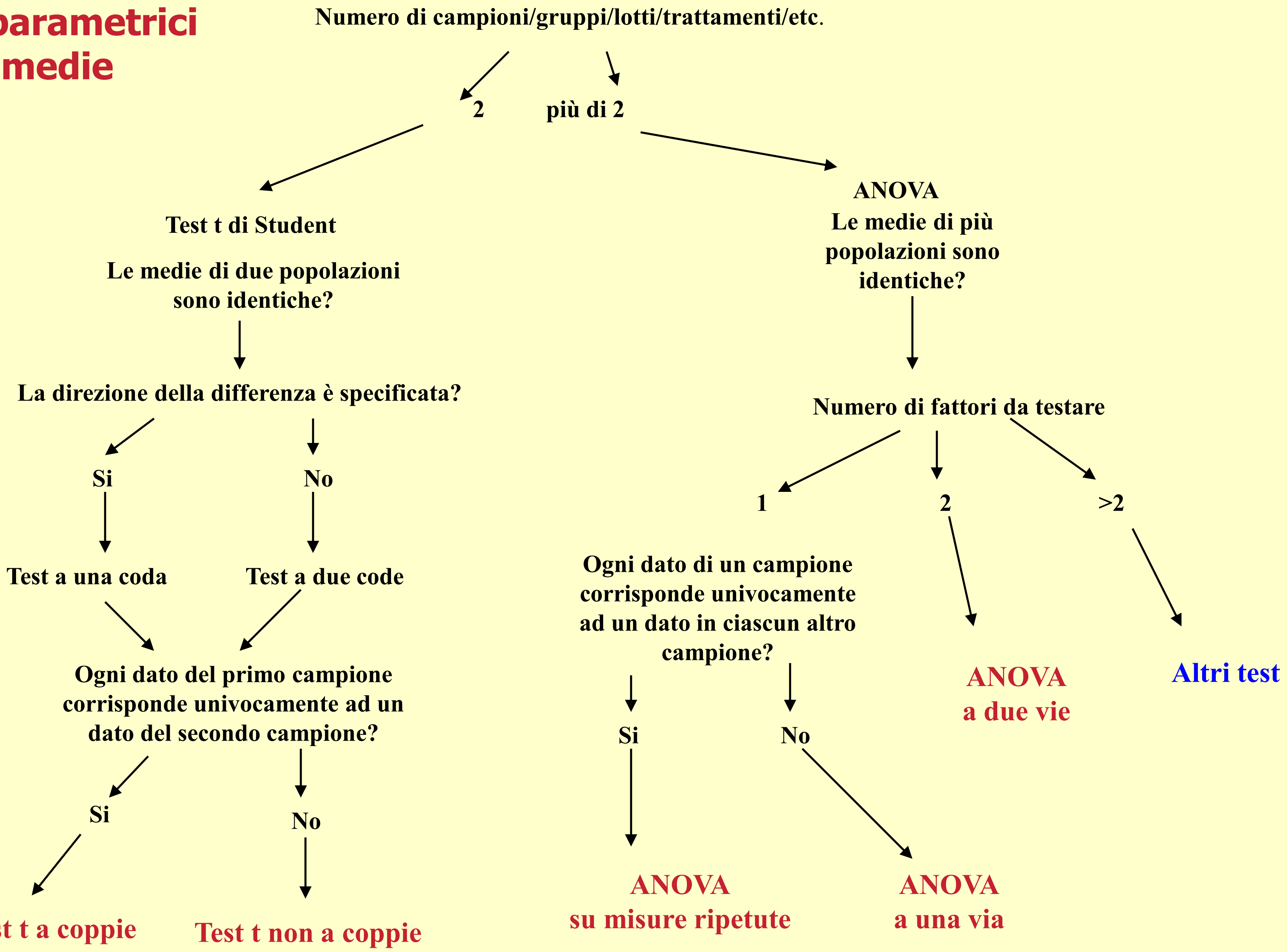

campione → popolazione

5. I risultati ottenuti con l'analisi di campioni si applicano alle popolazioni



6. Dimensione campione > 10 (meglio se ≥30)

Una chiave per i test parametrici sulle medie



Numero di campioni/gruppi/lotti/trattamenti/etc.

2

più di 2

Test t di Student

Le medie di due popolazioni sono identiche?

La direzione della differenza è specificata?

Si

No

Test a una coda

Test a due code

Ogni dato del primo campione corrisponde univocamente ad un dato del secondo campione?

Si

No

Test t a coppie

Test t non a coppie

ANOVA

Le medie di più popolazioni sono identiche?

Numero di fattori da testare

1

2

>2

Ogni dato di un campione corrisponde univocamente ad un dato in ciascun altro campione?

Si

No

ANOVA su misure ripetute

ANOVA a una via

ANOVA a due vie

Altri test

Se queste assunzioni (una o più sono violate)…

Assunzione	Altri test?	Rimedi?
1. Variabile non continua	Si	
2. Indipendenza media-varianza	No	Migliori metodi di misura
3. Distribuzione non normale	Si	Trasformazione dei dati
4. Varianze disomogenee	Si	
5. Campione ≠popolazione	Si	
6. n<10	Si	Raccogliere più dati

Test non-parametrici

- Questi test si impiegano quando almeno una delle assunzioni alla base del test t di Student o dell’ANOVA è violata.
- Sono chiamati “non-parametrici” perchè essi non implicano la stima di parametri statistici (media, deviazione standard, varianza, etc.).

Ne esistono almeno due grandi categorie:

1) Test di conformità (confronto fra valori osservati e valori attesi opportunamente calcolati)
2) Test equivalenti di test parametrici

Test di Wilcoxon

Due campioni non indipendenti, dati ordinali

Il test di Wilcoxon dovrebbe essere usato come alternativa non-parametrica al t di Student per campioni non indipendenti se una qualsiasi delle assunzioni necessarie per quest’ultimo è violata.

Test di Wilcoxon

Esperimento
Misura del tempo per cui si nutrono degli uccelli, come numero di minuti di attività nella mattina e nel pomeriggio

Uccello	Mattina	Pomeriggio	Differenza	Rango |differenza|	Rango con segno
1	23	46	17	4	4
2	28	51	23	7	7
3	37	29	-8	2	-2
4	24	49	25	8	8
6	27	46	19	5	5
6	27	39	22	6	6
7	31	30	-1	1	-1
8	28	41	13	3	3

H_0: non c'è differenza fra mattina e pomeriggio
H_1: esiste una differenza fra mattina e pomeriggio

Test di Wilcoxon: calcoli

Uccello	Mattina	Pomeriggio	Differenza	Rango |differenza|	Rango con segno
1	23	46	17	4	4
2	28	51	23	7	7
3	37	29	-8	2	-2
4	24	49	25	8	8
6	27	46	19	5	5
6	27	39	22	6	6
7	31	30	-1	1	-1
8	28	41	13	3	3

Somma dei ranghi positivi: T_+ = 4+6+8+7+5+3 = 33
Somma dei ranghi negativi: T_- = 2+1=3

Si rigetta H_0 se T_+ o $T_- \leq$ valore critico tabulare
In questo caso, poichè $T_{(.05,\ n=8)}$ = 3 , si rigetta H_0

Test U di Mann-Whitney

Due campioni indipendenti, dati ordinali

Il test U di Mann-Whitney dovrebbe essere usato come alternativa non-parametrica ad un test t di Student su campioni indipendenti, se una qualsiasi delle assunzioni necessarie è violata.

Test U di Mann-Whitney

Esperimento

Distanze al vicino più prossimo fra Nudibranchi in due quadrati campione

Quadrato 1	Quadrato 2
193	175
188	173
185	168
183	165
180	163
178	
170	

H_0: <u>non</u> c'è differenza fra i quadrati nella distanza al vicino più prossimo

H_1: c'è differenza fra i quadrati nella distanza al vicino più prossimo

Test U di Mann-Whitney: calcoli

Dati ordinati
193
188
185
183
180
178
175
173
170
168
165
163

Quadrato 1	Quadrato 2	Ranghi quadrato 1	Ranghi quadrato 2
193	175	1	7
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
178		6	
170		9	
$n_1 = 7$	$n_2 = 5$	$\Sigma R_1 = 30$	$\Sigma R_2 = 48$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \Sigma R_1 = 7 \cdot 5 + \frac{7 \cdot 8}{2} - 30 = 33$$

$$U' = n_1 n_2 - U = 7 \cdot 5 - 33 = 2$$

Se U o U’ $\geq U_{crit(.05,\ 7,\ 5)}$, si rigetta H_0

Poichè $U_{crit(.05,\ 7,\ 5)} = 30$ e U=33> 30, si rigetta H_0