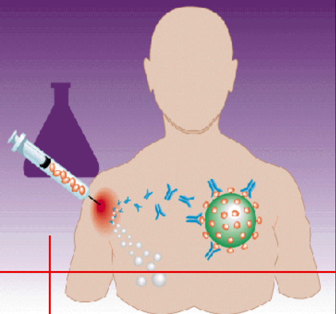
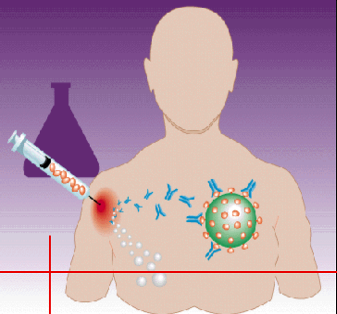


# Research Week 2016

# Basic Hypothesis Testing

Assoc. Prof. Dr Azmi Mohd Tamil
Dept of Community Health
Universiti Kebangsaan Malaysia

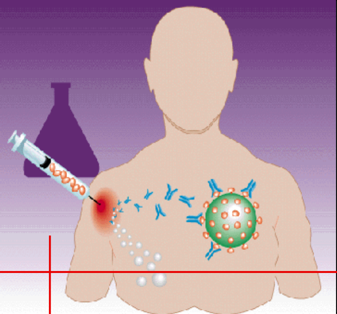
- Concept introduced by Jerzy Neyman & Egon Pearson in 1928.
- What does it mean to have a non-significant result in a significance test?
- Can we conclude that a hypothesis is true if we have failed to refute it?



- In many situations, hypothesis tests are used against a null hypothesis that is the straw man.
- For instance, when two drugs are being compared in a clinical trial, the null hypothesis to be tested is that the two drugs produce the same effect.
- However, if that were true, then the study would never have been run.
- The null hypothesis that the two treatments are the same is the straw man, meant to be knocked down by the results of the study.

# e.g. Drug to prevent recurrence of cancer

- Drug vs Placebo
- We expect if the drug is really effective, after 5 years the rate of recurrence of cancer is lower among treatment group (e.g. 0%) vs placebo group (e.g. 50%).



# Study with 8 samples

| | Relapse | Cured | |
|---|---|---|---|
| Treatment | 0 (0%) | 4 | 4 |
| Placebo | 2 (50%) | 2 | 4 |
| | 2 | 6 | 8 |

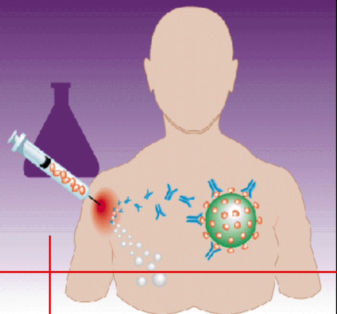
```
                    Chi-Squares      P-values
                    -----------      --------
 Uncorrected     :       2.67       0.1024704
 Mantel-Haenszel:        2.33       0.1266305
 Yates corrected:        0.67       0.4142162
Fisher exact: 1-tailed P-value: 0.2142857
              2-tailed P-value: 0.4285714

 An expected cell value is less than 5.
   Fisher exact results recommended.
```

**Null hypothesis**: There is no difference of relapse rate between the two treatment regimes.

**Result**: p>0.05

**Conclusion:** Null hypothesis not rejected.



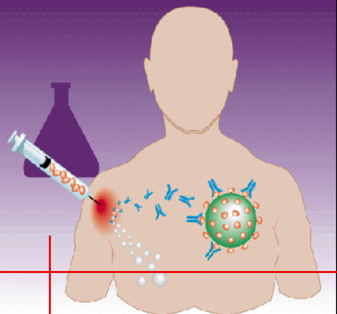
# Study with 16 samples

| | Relapse | Cured | |
|---|---|---|---|
| Treatment | 0 (0%) | 8 | 8 |
| Placebo | 4 (50%) | 4 | 8 |
| | 4 | 12 | 16 |

```
                        Chi-Squares    P-values
                        -----------    --------
 Uncorrected      :        5.33      0.0209213 ←——
 Mantel-Haenszel:          5.00      0.0253473 ←——
 Yates corrected:          3.00      0.0832645
Fisher exact: 1-tailed P-value: 0.0384615 ←—
              2-tailed P-value: 0.0769231

 An expected cell value is less than 5.
   Fisher exact results recommended.
```

- **Null hypothesis**: There is no difference of relapse rate between the two treatment regimes.
- **Result**: p>0.05
- **Conclusion:** Null hypothesis not rejected.
- But p value improving



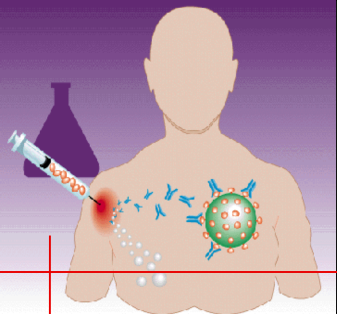
# Study with 32 samples

| | Relapse | Cured | |
|---|---|---|---|
| Treatment | 0 (0%) | 16 | 16 |
| Placebo | 8 (50%) | 8 | 16 |
| | 8 | 24 | 32 |

```
                  Chi-Squares    P-values
                  ___________    ___________
 Uncorrected     :     10.67     0.0010908 ←——
 Mantel-Haenszel:     10.33     0.0013065 ←——
 Yates corrected:      8.17     0.0042667 ←——
Fisher exact: 1-tailed P-value: 0.0012236 ←—
              2-tailed P-value: 0.0024472 ←—

 An expected cell value is less than 5.
   Fisher exact results recommended.
```

- **Null hypothesis**: There is no difference of relapse rate between the two treatment regimes.
- **Result**: $p<0.05$
- **Conclusion:** Null hypothesis rejected.
- Treatment has a significant effect on the outcome. The straw man is finally knocked down.



# Drug A versus Drug B

Hypothesis Testing

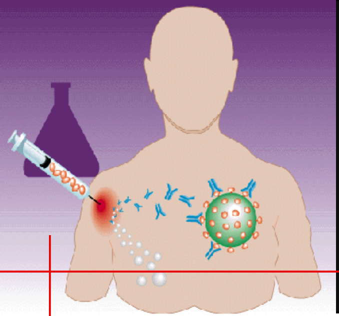
# Inferential Statistic

- When we conduct a study, we want to make an inference from the data collected. For example;

  “drug A is better than drug B in treating disease D"

# Is Drug A Better Than Drug B?

- Drug A  has a higher rate of cure than drug B. (Cured/Not Cured)
- If for controlling BP, the mean of BP drop for drug A is larger than drug B. (continuous data – mm Hg)



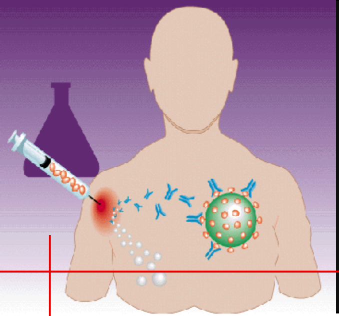
# Null Hypothesis or H0

- Null Hyphotesis;

  “no difference of effectiveness between drug A and drug B in treating disease D"

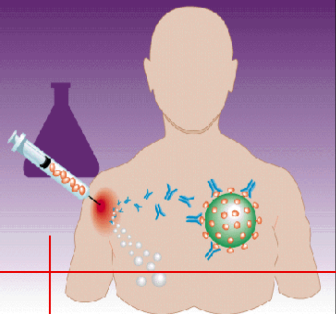
# Null Hypothesis

- **H0 is assumed TRUE unless data indicate otherwise:**
  - **The experiment is trying to reject the null hypothesis (the straw man)**
  - **Can reject, but cannot prove, a hypothesis**
    - **e.g. "*all swans are white*"**
      - » **One black swan suffices to reject**
        - » ***H0 "Not all swans are white"***
      - » **No number of white swans can prove the hypothesis – since the next swan could still be black.**



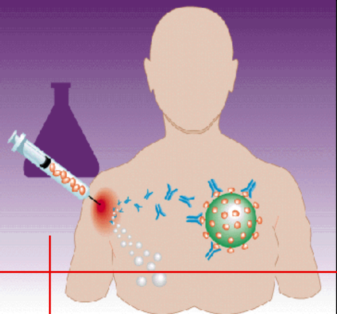
# Can reindeer fly?

- You believe reindeer can fly
- Null hypothesis: "reindeer cannot fly"
- Experimental design: to throw reindeer off the roof
- Implementation: they all go splat on the ground
- Evaluation: null hypothesis not rejected
    - This does not prove reindeer cannot fly: what you have shown is that
        - *"from this roof, on this day, under these weather conditions, these particular reindeer either could not, or chose not to, fly"*
- It is possible, in principle, to reject the null hypothesis
    - By exhibiting a flying reindeer!



# Significance

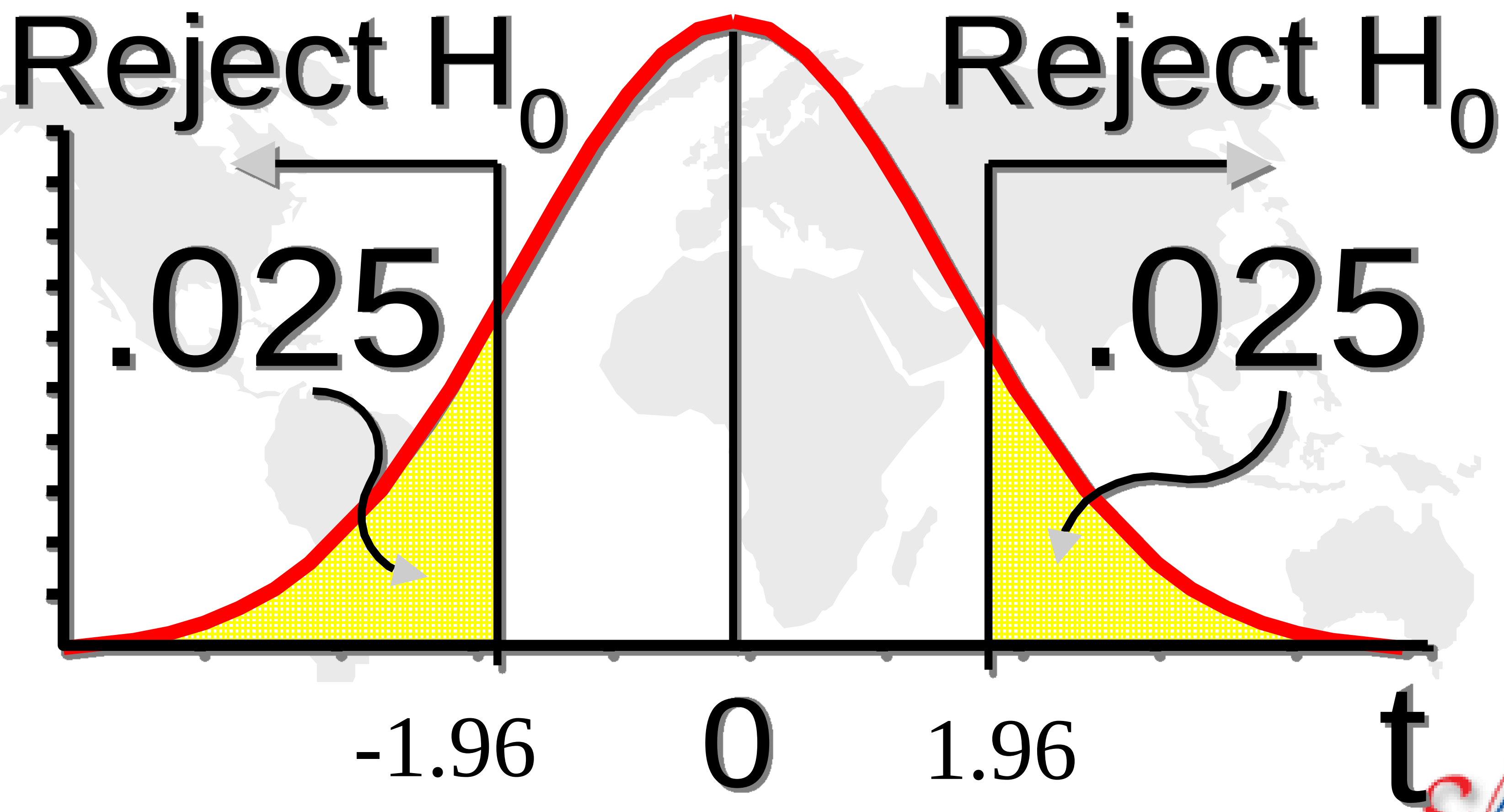
- Inferential statistics determine whether a significant difference of effectiveness exist between drug A and drug B.
- If there is a significant difference ($p<0.05$), then the null hypothesis **would be rejected.**
- Otherwise, if no significant difference ($p>0.05$), then the null hypothesis **would not be rejected**.
- The usual level of significance utilised to reject or not reject the null hypothesis are either 0.05 or 0.01. In the above example, it was set at 0.05.

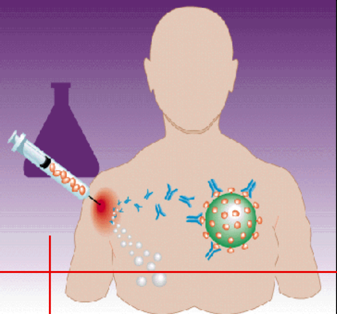


# Confidence interval

- Confidence interval = 1 - level of significance.
- If the level of significance is 0.05, then the confidence interval is 95%.
- CI = 1 – 0.05 = 0.95 = 95%
- If CI = 99%, then level of significance is 0.01.

# What is level of significance? Chance?





# Fisher's Use of p-Values

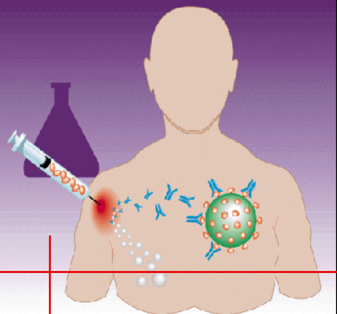
- R.A. Fisher referred to the probability to declare significance as "p-value".
- "It is a common practice to judge a result significant, if it is of such magnitude that it would be produced by chance not more frequently than once in 20 trials."
- 1/20=0.05. If p-value less than 0.05, then the probability of the effect detected were due to chance is less than 5%.
- We would be 95% confident that the effect detected is due to real effect, not due to chance.
- If $p < 0.001$? Then the probability that the effect detected were due to chance is less than 1 per 1,000 trials!

# Error

- Although we have determined the level of significance and confidence interval, there is still a chance of error.
- There are 2 types;
  - Type I Error
  - Type II Error

# Error

| | **REALITY** | |
|---|---|---|
| ***DECISION*** | Treatments are ***not different*** | Treatments are ***different*** |
| Conclude treatments are ***not different*** | **Correct Decision**<br><br>(Cell a) | **Type II error**<br>β error<br>(Cell b) |
| Conclude treatments are ***different*** | **Type I error**<br>α error<br>(Cell c) | **Correct Decision**<br><br>(Cell d) |

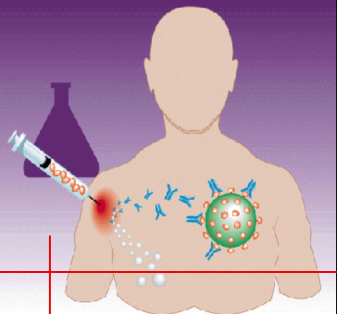


# Error

| Test of Significance | Correct Null Hypothesis (Ho not rejected) | Incorrect Null Hypothesis (Ho rejected) |
|---|---|---|
| Null Hypothesis Not Rejected | Correct Conclusion | Type II Error |
| Null Hypothesis Rejected | Type I Error | Correct Conclusion |

# Type I Error

- Type I Error – rejecting the null hypothesis although the null hypothesis is correct e.g.
- when we compare the mean/proportion of the 2 groups, the difference is small but the difference is found to be significant. Therefore the null hypothesis is rejected.
- It may occur due to inappropriate choice of alpha (level of significance).



# Example of a Type I Error

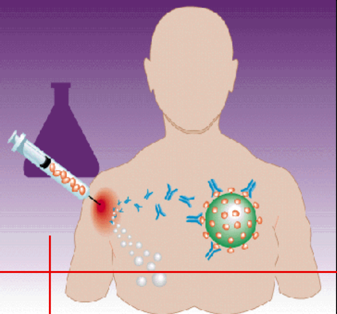
Multiple comparisons

- When we are comparing between 2 treatments A & B with a 5% significance level, the chance of a true negative in this test is 0.95. But when we perform A vs B and A vs C (in a three treatment study), then the probability that neither test will give a significant result when there is no real difference is 0.95 x 0.95 = 0.90; which means the type 1 error has increased to 10%.

| Number of comparisons | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Probability of false positive | 5% | 10% | 14% | 19% | 23% | 27% | 30% | 34% | 37% | 40% |

# Type II Error

- Type II Error – not rejecting the null hypothesis although the null hypothesis is wrong
- e.g. when we compare the mean/proportion of the 2 groups, the difference is big but the difference is not significant. Therefore the null hypothesis is not rejected.
- It may occur when the sample size is too small.



# Example of Type II Error

Data of a clinical trial on 30 patients on comparison of pain control between two modes of treatment.

**Type of treatment * Pain (2 hrs post-op) Crosstabulation**

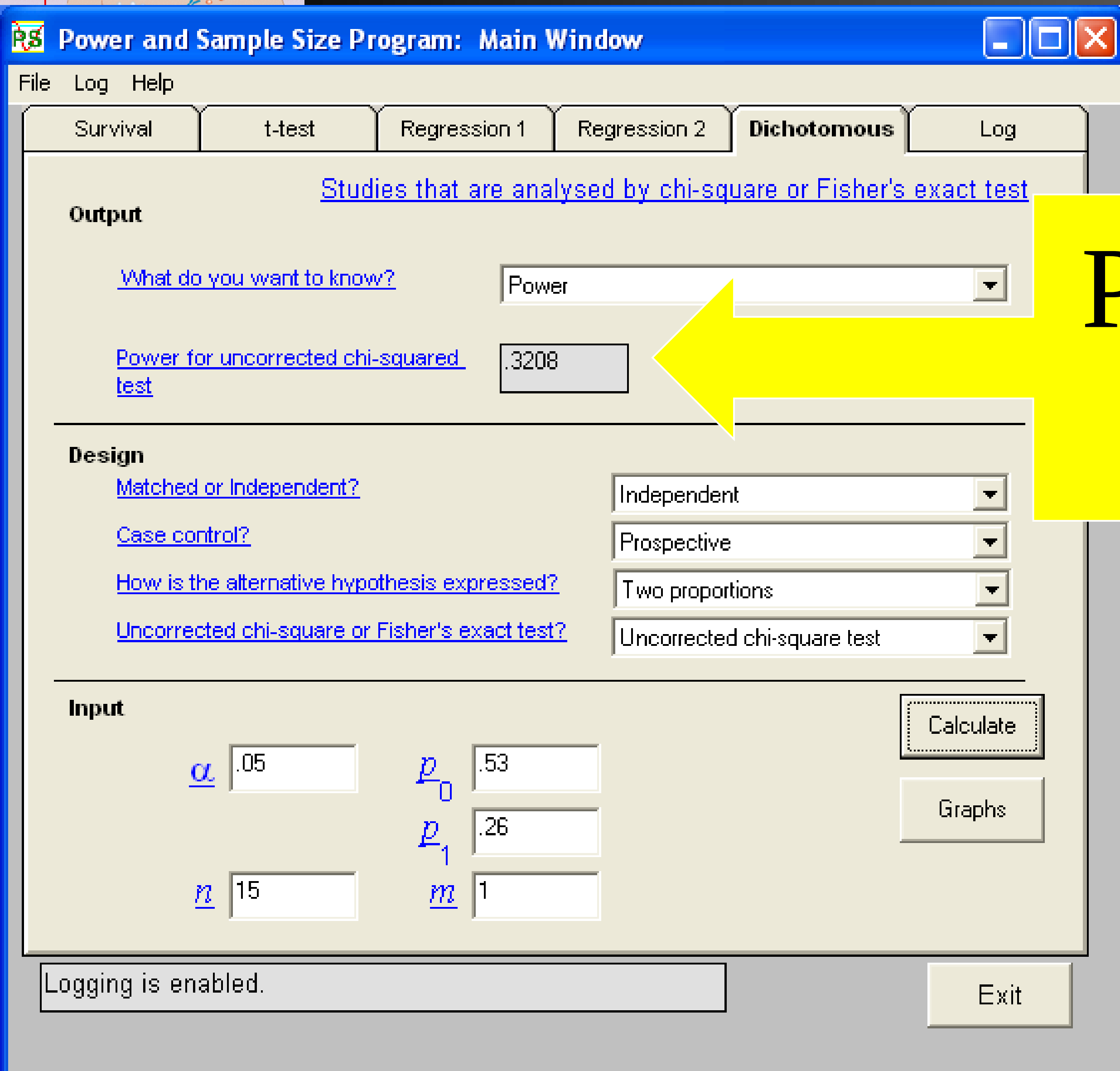
| | | | Pain (2 hrs post-op) | | |
|---|---|---|---|---|---|
| | | | No pain | In pain | Total |
| Type of treatment | Pethidine | Count | 8 | 7 | 15 |
| | | % within Type of treatment | 53.3% | 46.7% | 100.0% |
| | Cocktail | Count | 4 | 11 | 15 |
| | | % within Type of treatment | 26.7% | 73.3% | 100.0% |
| Total | | Count | 12 | 18 | 30 |
| | | % within Type of treatment | 40.0% | 60.0% | 100.0% |

Chi-square =2.222, p=0.136

p = 0.136. p bigger than 0.05. No significant difference and the null hypothesis was not rejected.

There was a large difference between the rates but were not significant. Type II Error?

# Not significant since power of the study is less than 80%.

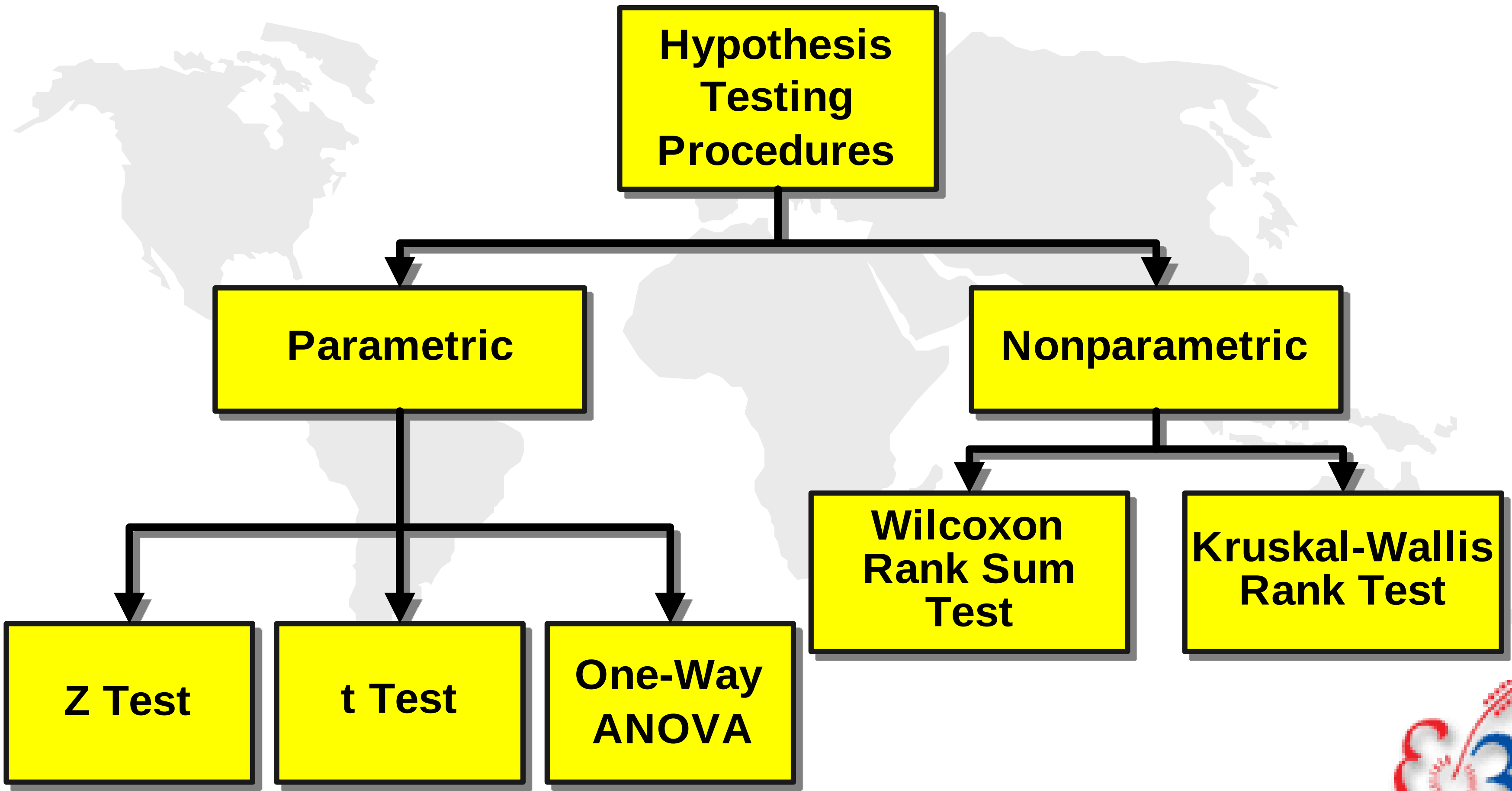


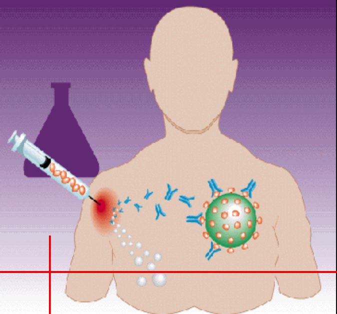

Power is only 32%!

# Check for the errors

- You can check for type II errors of your own data analysis by checking for the power of the respective analysis
- This can easily be done by utilising software such as Power & Sample Size (PS2) from the website of the Vanderbilt University

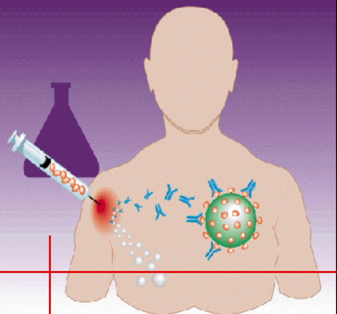
# Hypothesis Testing Procedures





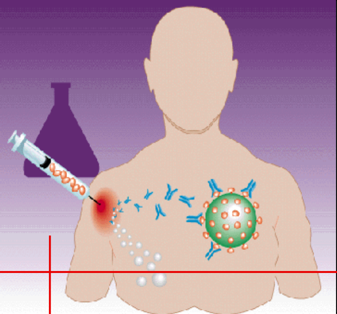
# Parametric Analysis – Quantitative

| | | | |
|---|---|---|---|
| Qualitative Dichotomus | Quantitative | Normally distributed data | Student's t Test |
| Qualitative Polinomial | Quantitative | Normally distributed data | ANOVA |
| Quantitative | Quantitative | Repeated measurement of the same individual & item (e.g. Hb level before & after treatment). Normally distributed data | Paired t Test |
| Quantitative - continous | Quantitative - continous | Normally distributed data | Pearson Correlation & Linear Regresssion |



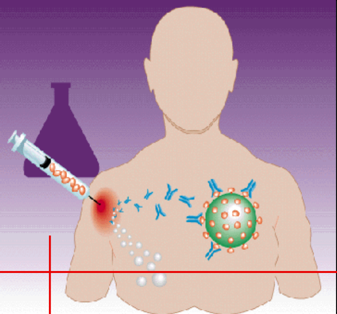
# non-parametric tests

| Variable 1 | Variable 2 | Criteria | Type of Test |
|---|---|---|---|
| Qualitative Dichotomus | Qualitative Dichotomus | Sample size < 20 or (< 40 but with at least one expected value < 5) | Fisher Test |
| Qualitative Dichotomus | Quantitative | Data not normally distributed | Wilcoxon Rank Sum Test or U Mann-Whitney Test |
| Qualitative Polinomial | Quantitative | Data not normally distributed | Kruskal-Wallis One Way ANOVA Test |
| Quantitative | Quantitative | Repeated measurement of the same individual & item | Wilcoxon Rank Sign Test |
| Quantitative - continous | Quantitative - continous | Data not normally distributed | Spearman/Kendall Rank Correlation |



# Statistical Tests - Qualitative

| Variable 1 | Variable 2 | Criteria | Type of Test |
|---|---|---|---|
| Qualitative | Qualitative | Sample size $\geq$ 20 dan no expected value < 5 | Chi Square Test ($X^2$) |
| Qualitative Dichotomus | Qualitative Dichotomus | Sample size > 30 | Proportionate Test |
| Qualitative Dichotomus | Qualitative Dichotomus | Sample size > 40 but with at least one expected value < 5 | $X^2$ Test with Yates Correction |
| Qualitative Dichotomus | Qualitative Dichotomus | Sample size < 20 or (< 40 but with at least one expected value < 5) | Fisher Test |



# Take Home Message

Use the tables to decide on what type of analysis to use.

| | | | |
|---|---|---|---|
| Qualitative Dichotomus | Quantitative | Normally distributed data | Student's t Test |
| Qualitative Polinomial | Quantitative | Normally distributed data | ANOVA |
| Quantitative | Quantitative | Repeated measurement of the same individual & item (e.g. Hb level before & after treatment). Normally distributed data | Paired t Test |
| Quantitative - continous | Quantitative - continous | Normally distributed data | Pearson Correlation & Linear Regresssion |