Materials Informatics

Lecture 3: Review of Probability and Statistics

Ulisses Braga Neto

Department of Electrical and Computer Engineering
Texas A&M University

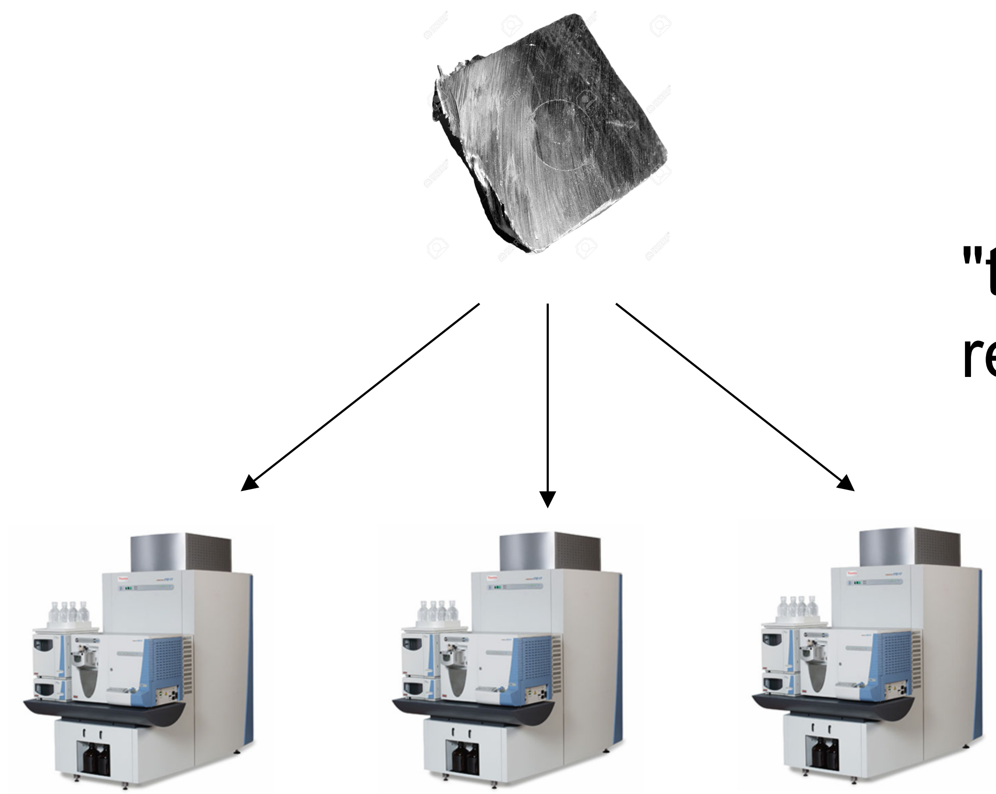
Variability/Reproducibility

- Results can be expected to have reproducibility only if variability in taken into account in experimental design.

- Even if the subject and assay are clearly specified, there are two sources of variability
  - Techical Variability : it comprises random changes from an experiment (microarray) to the next.
  - Populational Variability : it comprises random changes from an individual or sample to the next.

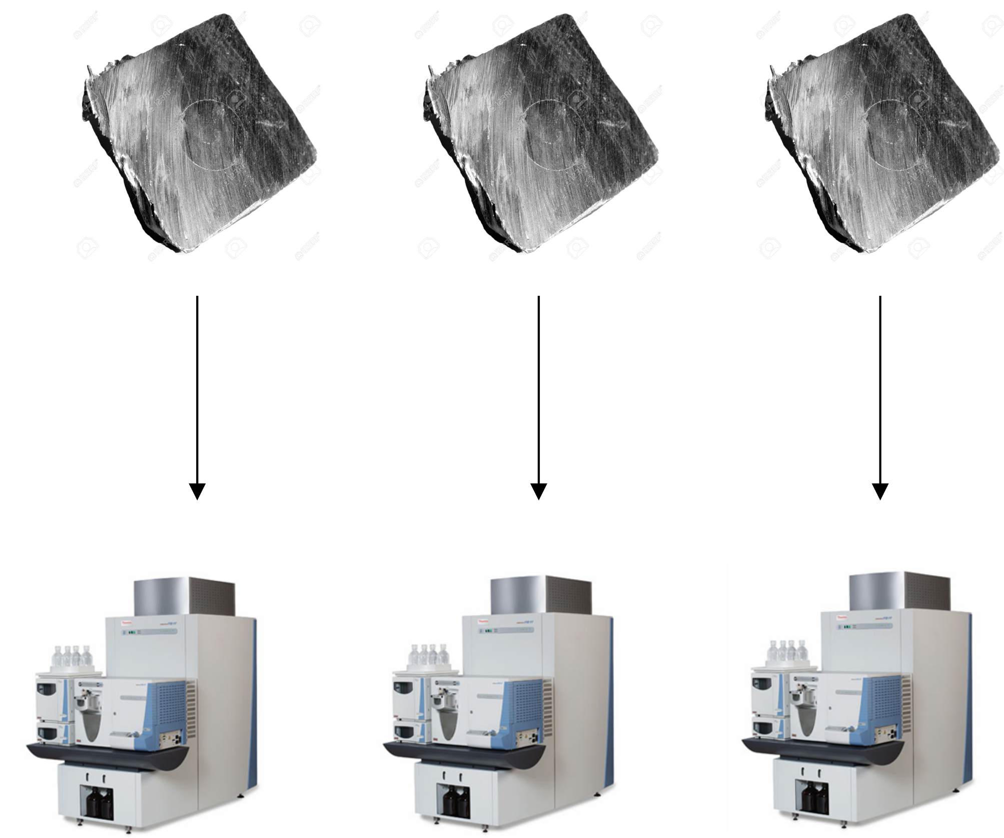
Replicates

- In order to address variability (and thus ensure reproducibility), one needs replicates



"technical replicates"

Replicates - II



”population replicates"

- As long as the assay is reliable, population replicates are more critical for reproducibility

Statistical Significance

- Suppose there are two conditions A and B (e.g., ordinary vs. superconductivity) under study, and one measurement Y (e.g. a QSPR).
- Suppose further that there are n replicate specimens, n/2 under condition A and n/2 under condition B (since A and B are represented by the same number of samples, this is called a balanced design).

Statistical Significance - II

- The question of interest is:
  "Based on the set of n replicates, can we conclude that Y is significantly different between A and B?"
- To examine this question, let us assume that there is indeed a difference between A and B. Suppose for example that in truth we have

$$Y = 100, \quad \text{under A}$$
$$Y = 200, \quad \text{under B.}$$

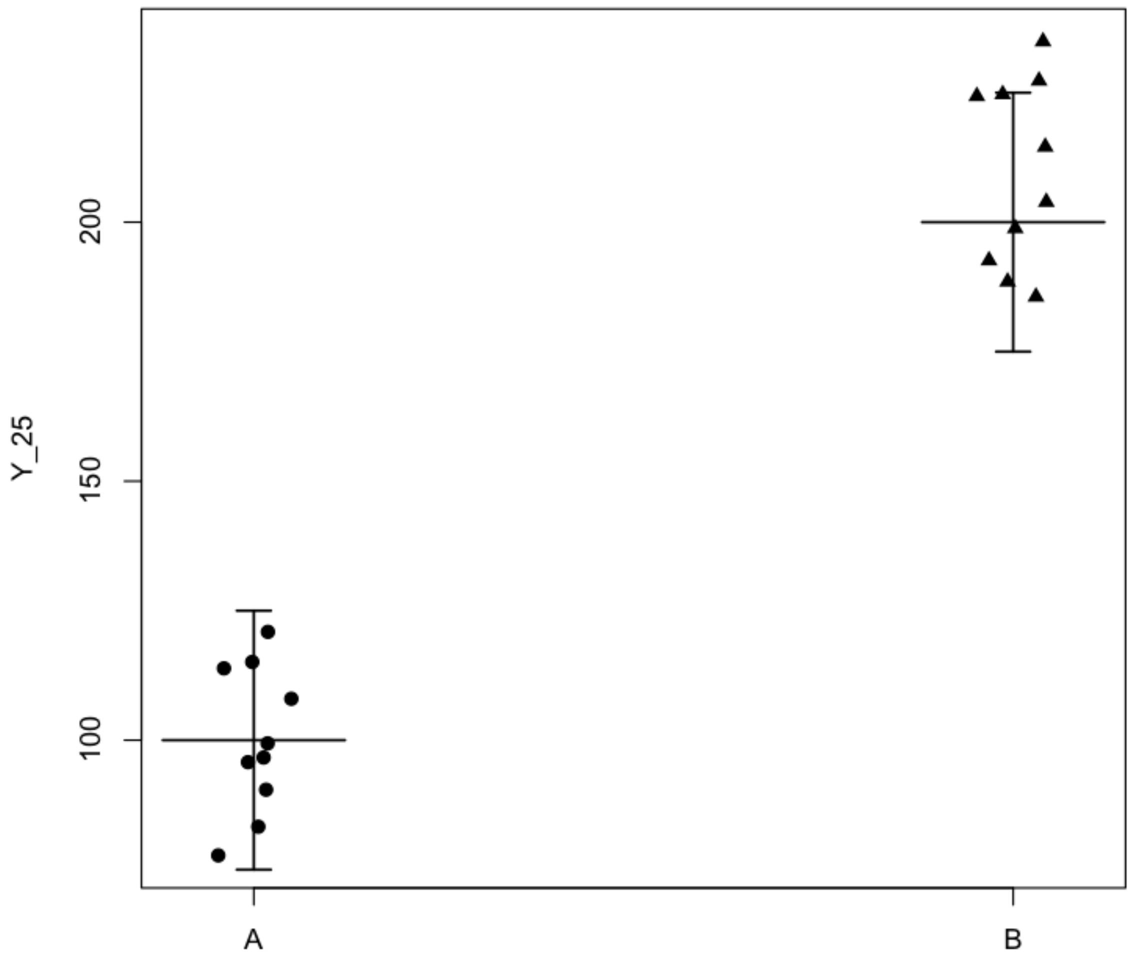
Statistical Significance - III

- The fold-change is 200/100 = 2.
- If there were no variability, then with n=2 (1 replicate for each condition) we would be able to conclude there was a difference.
- But clearly there will be some variability, both technical and populational. Let us model this by using Gaussian distributions

$$Y \sim N(100, \sigma^2) \quad \text{under A}$$
$$Y = N(200, \sigma^2) \quad \text{under B}$$

where σ^2 is the variance (assumed equal, for the moment) in A and B.

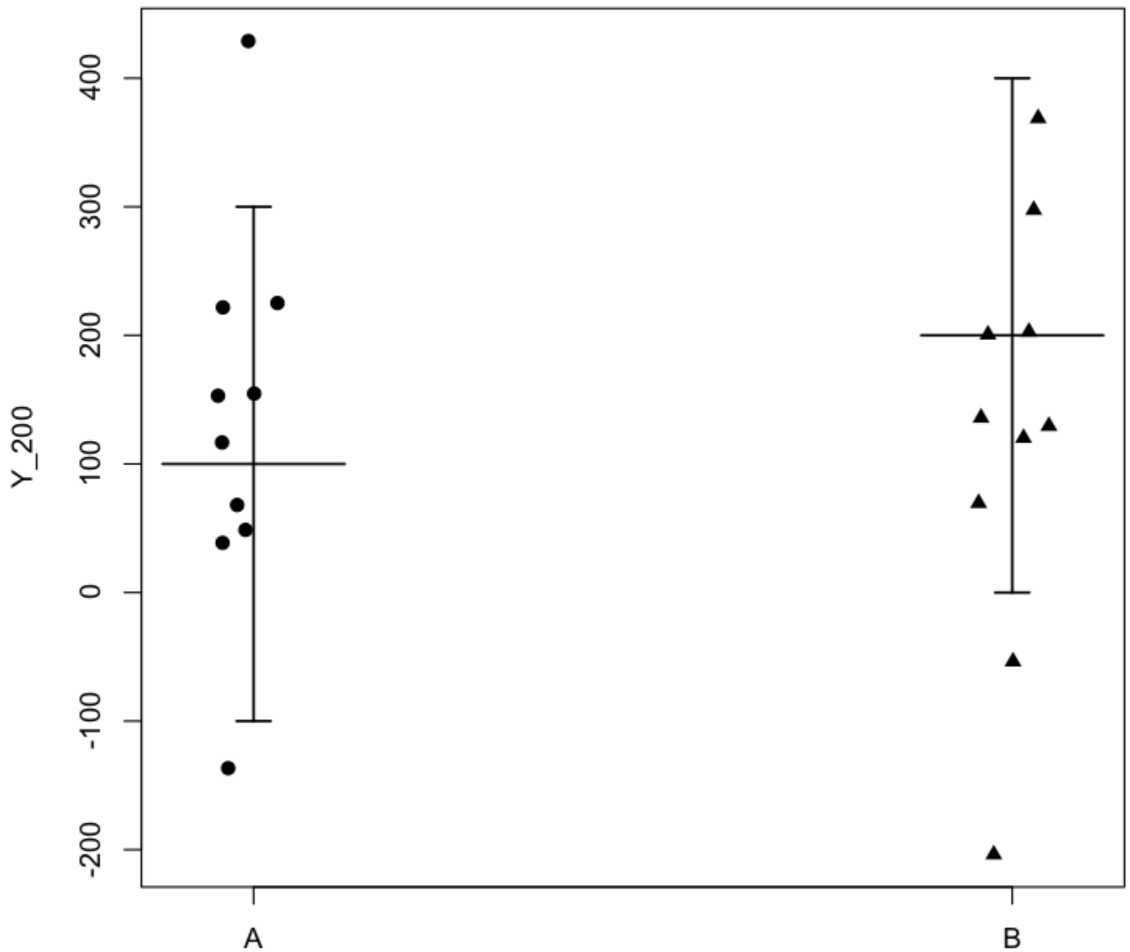
Statistical Significance - IV



- It seems clear here that there is indeed differential expression in Y between A and B.

Statistical Significance - V

- But what if variability were larger, e.g. $\sigma = 200$?



There does not seem to be a difference now.

Weakness of Fold Change

- The fold-change is the same in the two cases. How can one explain this discrepancy?
- The answer is that fold-change by itself is not a reproducible criterion for discrimination, because it does not take into account the inherent variability in the data.
- In practice, the fold-change would be estimated as the ratio between the sample means based on the data at hand, which would not change any of these conclusions.

Hypothesis Tests

- The matter can be formalized with the notion of hypothesis tests.
- Let μ_A and μ_B be the expression level of Y under conditions A and B, respectively. We would like to test the null hypothesis

$$H_0 : \mu_A = \mu_B$$

against the alternative hypothesis

$$H_1 : \mu_A \neq \mu_B$$

at a given significance level 100x(1-α)%.

Hypothesis Tests - II

- To do this, one computes a test statistic T, which is simply some function of the data.
- T is called a "statistic" because it is a function of the data. Because the data is random (it has variability), T is also a random variable.
- The null hypothesis is said to be rejected, and the alternative hypothesis accepted, if the observed T has an atypical value under the null hypothesis, i.e. T falls in the rejection region R of the test, which is defined such that

$$P_{H_0}(T \in R) = \alpha$$

The difference is called statistically significant.

The p-value

- The p-value is simply the probability, under the null hypothesis, of observing a more atypical (e.g. larger in magnitude) value of T than the one actually observed, T_0 :

$$p = P_{H_0}(|T| > |T_0|)$$

- Clearly, in this example, the larger $|T_0|$ is, the smaller p is.
- If $p < 1-\alpha$ then H_0 is rejected and the test is statistically significant.

Misconceptions about p-value

- The p-value is NOT the probability that the null hypothesis is true.
- Similarly, subtracting the p-value from 1 does NOT give the probability that the alternative hypothesis is true.
- The threshold $p<0.05$ is nothing other than a pure convention.
- The observed value of T itself can be useful (it measures the so-called effect size).

Error Rates

- An error rate is a probability of making a mistake in the hypothesis test.
- Error Type-I (α): Probability of rejecting H_0 when it is true; false positive.
- Error Type II (β): Probability of not rejecting H_0 when it is false; false negative.

	H_0 true	H_0 false
Reject H_0	α	$1 - \beta$
Not Reject H_0	$1 - \alpha$	β

Sensitivity/Specificity

- The sensitivity of a test, also called statistical power, is 1 - β. The more sensitive a test is, the smaller the differences it can detect.
- Sensitivity is a function of sample size. The larger it is, the more sensitive the test will be.
- The specificity of a test is simply the significance level α. The more specific a test is, the fewest false positives it will produce.
- Sensitivity and specificity are conflicting requirements. One can make one larger by decreasing the other. A superior test will be more sensitive at the same specificity.

Two-Sample t-test

- The most common test used in differential expression studies, by far, is the two-sample t-test. It assumes Gaussian data, but is quite robust against failure of this assumption.
- The version known as Welch's applies to conditions that have different variances.
- The test statistic in this case is

$$T = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\hat{\sigma}_A^2/n_A + \hat{\sigma}_B^2/n_B}}$$

Two-Sample t-test - II

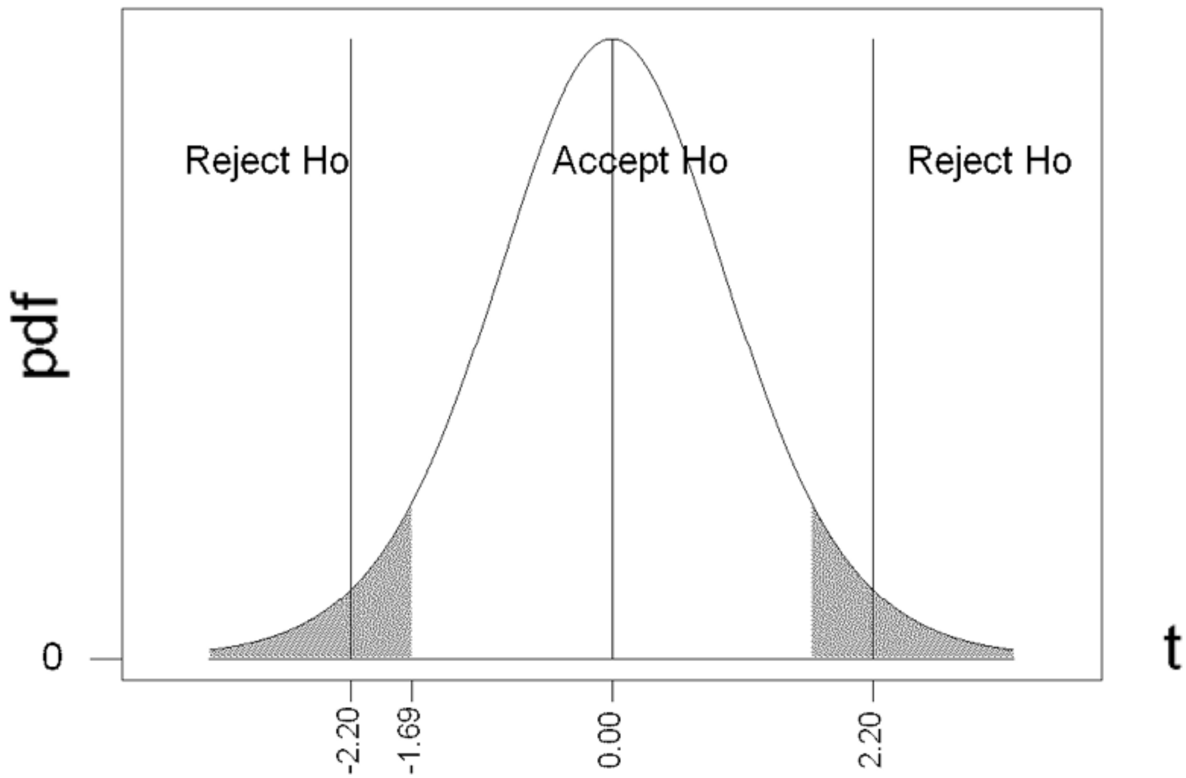
- In the previous formula:

$$\begin{aligned} \bar{x}_\cdot &= \text{sample mean} \\ n_\cdot &= \text{number of samples} \\ \hat{\sigma}^2_\cdot &= \text{sample variance} \end{aligned}$$

- Note that T takes into account the variance in the data, which the fold-change does not.
- The smaller the variances, and/or the larger the number of samples are, the larger the magnitude of T is, the smaller the p-value is, and the larger the evidence for rejecting H_0.

Two-Sample t-test - III

- For Welch's test, under the null hypothesis, T has an approximate Student's t distribution, with a noninteger number of degrees of freedom.



(From http://junior.apk.net/~pmathews/tot/abstract/MTBExecMacros.html)

Numerical Example

- Same simulated data as before, $\sigma = 25$

```
> t.test(Y_25~cond)
    Welch Two Sample t-test
data:  Y_25 by cond
t = -13.0557, df = 17.992, p-value =
1.292e-10
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -120.53782  -87.12028
sample estimates:
mean in group A mean in group B
       102.0391        205.8682
```

Numerical Example - II

- Same simulated data as before, σ = 200

```
> t.test(Y~cond)
    Welch Two Sample t-test
data:  Y_200 by cond
t = 0.0734, df = 17.806, p-value = 0.9423
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -142.2260  152.5159
sample estimates:
mean in group A mean in group B
       131.9043        126.7593
```

Nonparametric Test

- The t-test is founded on a parametric Gaussian assumption.
- A non parametric alternative can be obtained by considering not the numeric values of the measurement, but only their ranking.
- If most points in A are among the top and those in B are among the bottom of the rankings, intuitively there is discrimination.
- The resulting test is called the Wilcoxon rank-sum test.

Numerical Example

- Same simulated data as before, $\sigma = 25$

```
> wilcox.test(Y_25~cond)

	Wilcoxon rank sum test

data:  Y_25 by cond
W = 0, p-value = 1.083e-05
alternative hypothesis: true location shift
is not equal to 0
```

- Note that the p-value is larger than for the t-test; if the Gaussianity assumption is satisfied (it is for this data), then Welch's t-test is more powerful than Wilcoxon's test (and the equal-variance t-test is uniformly most powerful).

Numerical Example - II

- Same simulated data as before, $\sigma = 200$

```
> wilcox.test(Y_200~cond)

    Wilcoxon rank sum test

data:  Y by cond

W = 48, p-value = 0.9118

alternative hypothesis: true location shift
is not equal to 0
```

Multiple Testing Issue

- Suppose there are 1000 variables to be tested and none of them are significantly different between the conditions.
- If we apply the standard significance level of 0.05, and assuming the tests are independent, we expect to get 1000 x 0.05 = 50 false positives. (why?)
- This is clearly an acceptable situation. If I have 20 significant differences at the 95% confidence level, there is a good chance that all of them are false positives.
- This happens because of the multiple tests.

Bonferroni Correction

- One possibility to correct this situation is to enlarge the p-values by multiplying them by the number of tests. This is known as the Bonferroni correction.

- If one starts with 1000 variables, to get a corrected $p < 0.05$ one has to get $p < 0.00005$.

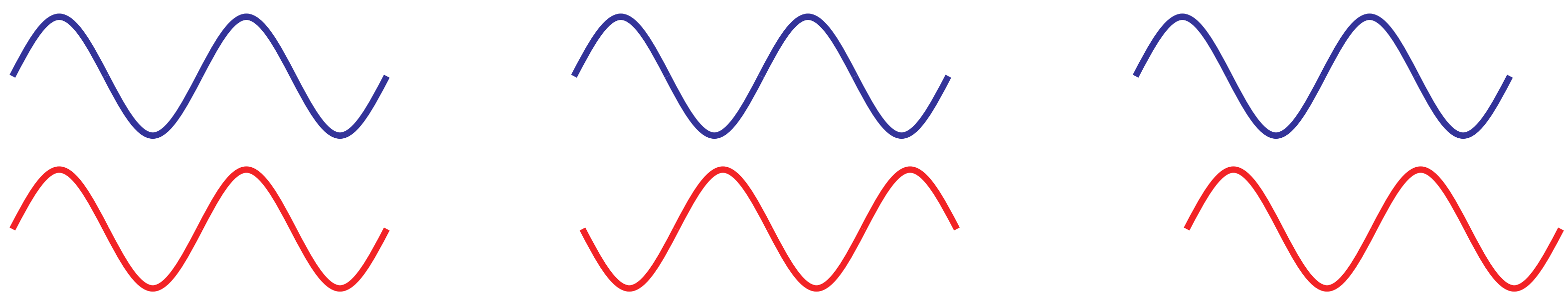
- However, it can be shown that the Bonferroni correction is conservative (it reduces the false positives too much, creating an excess of false negatives).

Ranking by Effect Size

- There are other less-conservatives alternatives to the Bonferroni correction, such as the "false discovery rate". But the sitation is a bit chaotic. No one knows for sure which correction method is the best.
- One alternative in practice is to ignore the p-values and simply rank the variables by effect sizes.

Correlation

- Correlation is a measure of the coincidence (or lack there of) of directionality of change across samples between two variables.



high positive correlation

high negative correlation

small correlation (in magnitude)

Correlation is NOT causality.

Sample Correlation Coefficient

- Correlation can be measured by Pearson's correlation coefficient

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- We have $-1 \leq \rho \geq 1$ and

$$\begin{aligned} \rho \rightarrow 1 &\Rightarrow \text{ large positive correlation} \\ \rho \rightarrow -1 &\Rightarrow \text{ large negative correlation} \\ \rho \approx 0 &\Rightarrow \text{ small correlation} \end{aligned}$$

Sample Correlation Coefficient

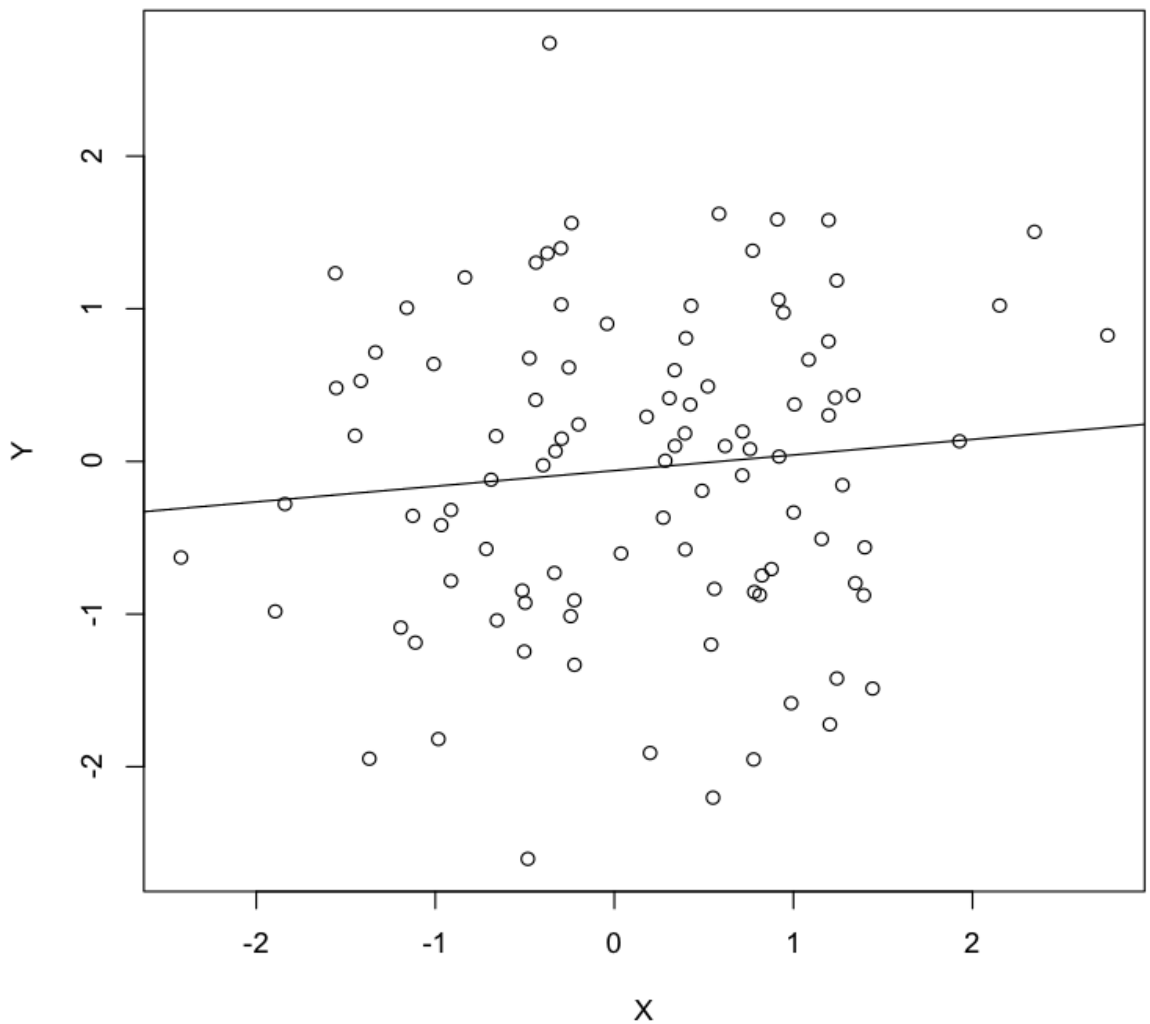
- Pearson's correlation coefficient measures linear association.
- It is quite sensitive to outliers, therefore plotting the data is always recommended.
- The coefficient of determination

  $$R^2 = \rho^2$$

  is a measure between 0 and 1 that measures the magnitude of association.
- There are other nonparametric correlation coefficients (as in the case of t-tests).

Numerical Example - II

- With cor = 0.1



Numerical Example - II

- With cor = 0.8