<https://github.com/boboppie/coursera-course-statistics_one>

**LECTURE 4 - Correlations**

**Segment 1:**

Correlation: measure and describe relationship between 2 variables

Correlations range from -1 to +1

* The magnitude of the correlation depends upon:
  + Reliability of X, Reliability of Y, Sampling
* The validity of the prediction
* The correlation coefficient is a sample statistic (for the group, not for individual)

**Segment 2: Calculating correlations**

Review Lecture 2:

1 variable (x)

Deviation score SD = (X -Mx)

Square of deviation score SD2 = (X – Mx)2

Sum the squared deviation scores SSx = E[(X-Mx)2]

2 variables (x,y)

Sum of cross products (SP): SPxy = E[(X-Mx) \* (Y – My)]

* **Pearson’s correlation coefficient (r):** is the measure of the linear correlation between two variables X and Y, giving a value -1 to +1. where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation.

**r = SPxy /√(SSx \*SSy)**

It also can be obtained from z-scores: Zx =(X-Mx)/SDx Zy =(X-My)/SDy

**r = E(zx\*zy)/N**

<https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient>

* Covariance:
  + Variance = MS = SS/N
  + Covariance = COV = SP/N (standardized value range -1 to 1)
* Correlation for descriptive purposes: use N
* Correlation for inferential purposes: use N-1

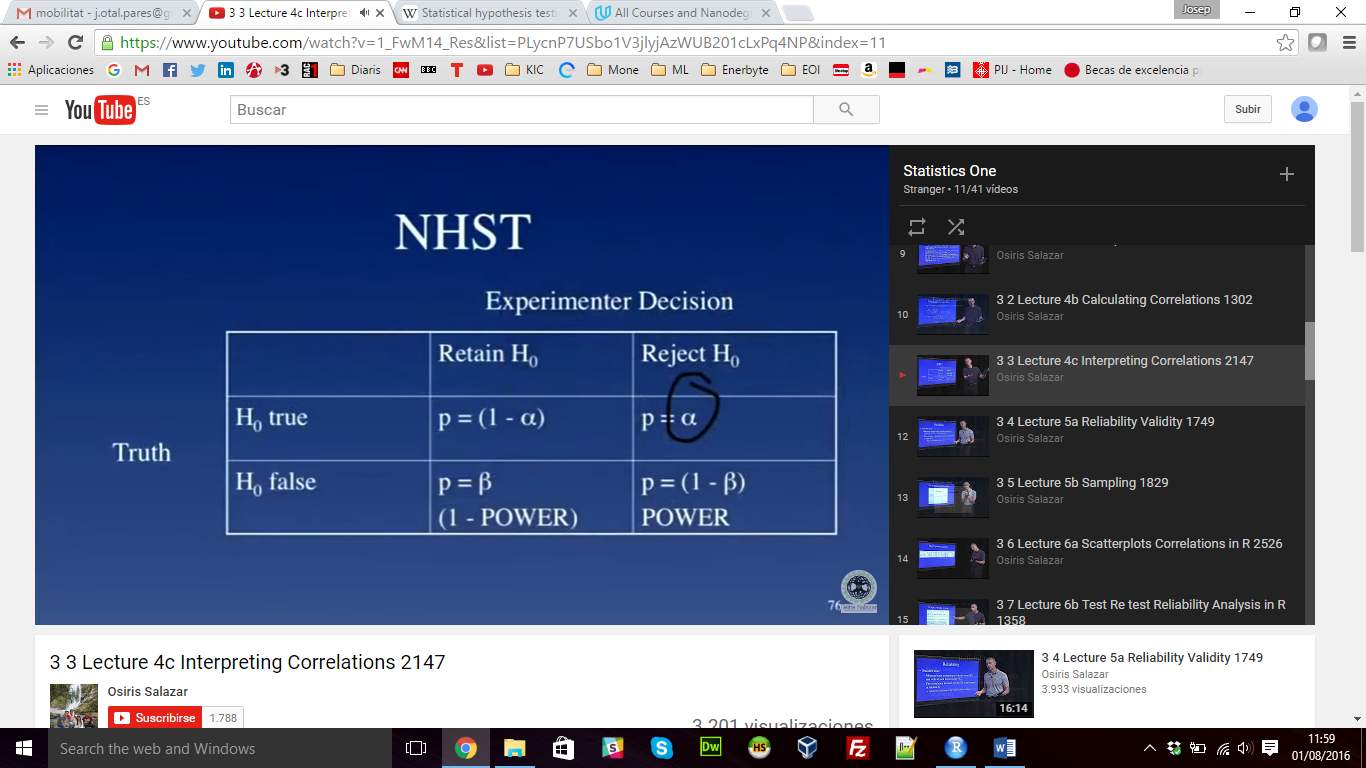
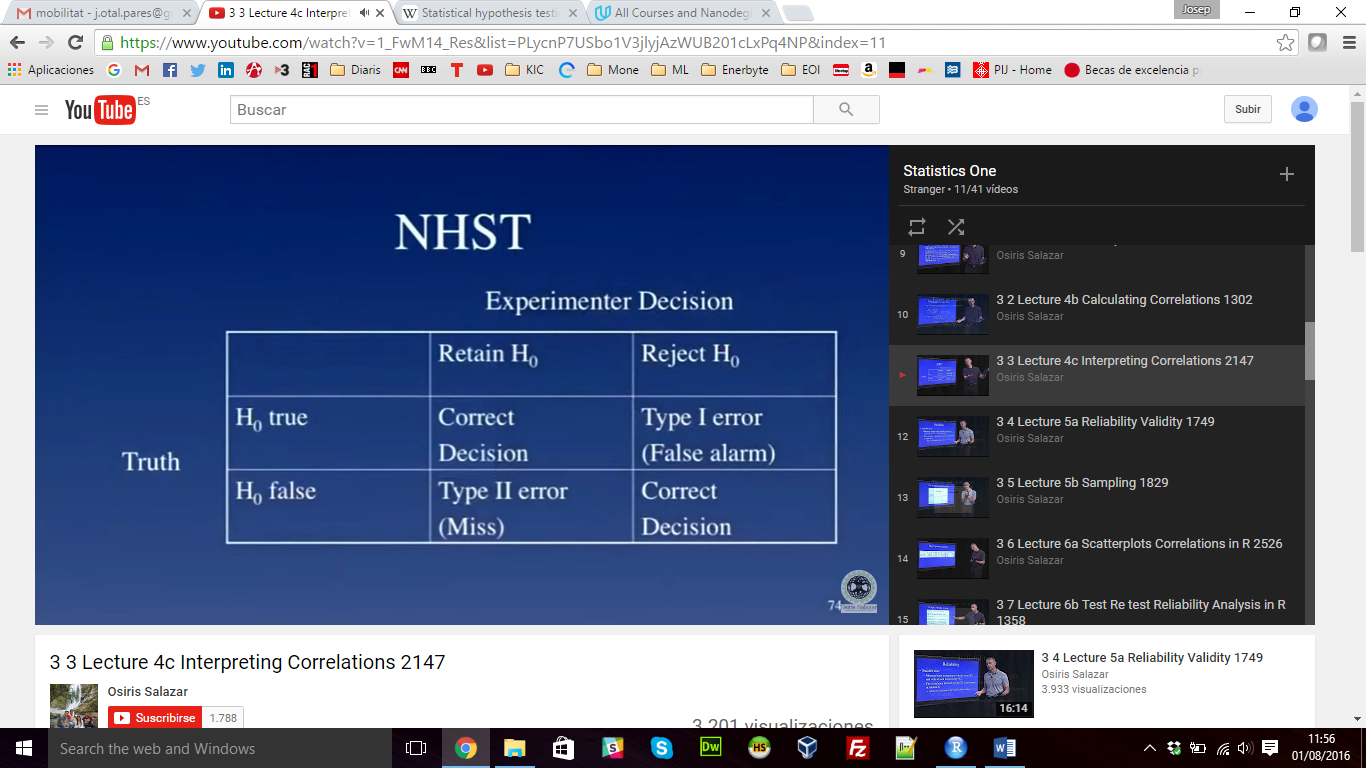
**Segment 3: Interpreting Correlations**

Validity: of correlation analysis depends on the assumptions

* Assumption 1: X and Y are normal distributions 🡪 descriptive stats
* Assumption 2: X and Y has linear relationship 🡪 scatter plots and plot the residuals
* Assumption 3: Homoskedasticity 🡪 scatter plots and plot the residuals
  + Distance between the dot and the regression line is the prediction error (or residual). Homoskedasticity states that these distances do not follow any pattern, so those distances are not a function of X (not related to the predicted variable)

Reliability: of a correlation

* Null Hypotesis Significance Testing (NHST)
  + H0 = Null hypothesis. Eg, no correlation r = 0
  + HA = alternative hypothesis r > 0



**p = P(D|H0)** **p-value** is the probability of the Data given the Null hypothesis is true

Is NOT the probality of theH0 is true