

IS/AAE374 Discussion Section
Problem Set 2, T-Test and Scatter Plots

Problem Set 2. What you need to know:

1) Representing distributions (discussed last week)

- a) Histograms
- b) Measures of Central Tendency (mean, median)
- c) Measures of Spread (variance, standard deviation)

2) Mean comparison between two groups

- a) T-tests are used to compare two groups or segments of the population. In the case of problem set 2, t-tests will allow us to check if the difference in means between two segments of the population is statistically significant.

i) Null hypothesis $H_0 : \mu_a = \mu_b$

ii) Alternative hypothesis $H_A : \mu_a \neq \mu_b$

iii) Formula: $t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$

- b) An example in Excel:

- i) Download statsareawesome.xml and go the sheet “enrollment2000” please!
- ii) Separately for low income and high income countries, estimate the mean and standard deviation for enrollment rates.
- iii) In a separate cell enter the function “=ttest(” and you should see “=ttest(array1,array2,tails,type)”
 - (1) Arrays: select low income data for array1 and high income for array2
 - (2) Tails: If you predicted that the mean of group 1 would be lower than that of group 2, pick 1 tail (enter 1). If you predicted that the mean of group 2 would be lower than that of group 1, pick 1 tail. If you didn’t predict which would be higher, use 2 tails. For this example, select 1 tail.
 - (3) Type 2: Select type 2 t-test (equal variances) – just trust me!
 - (4) You should have something like “=TTEST(C2:C61,C62:C91,1,2)” Congrats!
- iv) The output of this function represents the p-value associated to the t-test. In this case, a low p-value (<0.05) means that there is significant evidence to reject the null

hypothesis; i.e. that the difference in means is statistically significant.

v) T-value

(1) $|t| > 1.64$: you can reject the null hypothesis at the 90% confidence level

(2) $|t| > 1.96$: you can reject the null hypothesis at the 95% confidence level

3) Growth Rates

a) See handout #1. After reminiscing about how great section was that day, review growth rate information.

4) Scatter Plots

a) Two way graph & Trend line

b) Each point represents one data entry with one variable determining the position on the horizontal axis and the other variable the position on the vertical axis.

c) Scatter plots provide a visual depiction of the relationship between these two variables. Remember, causation does not imply correlation!

d) Again using statsareawesome.xml, make a scatter plot between infant mortality and child enrollment and include the trend line (under the Chart menu). What is the relationship between infant mortality and school enrollment? On a scale from 1 to 10 (10 being “stats are awesome!”), how much fun is this?

e) Covariance & Correlation

$$i) \rho = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

$$ii) -1 \leq \rho \leq 1$$