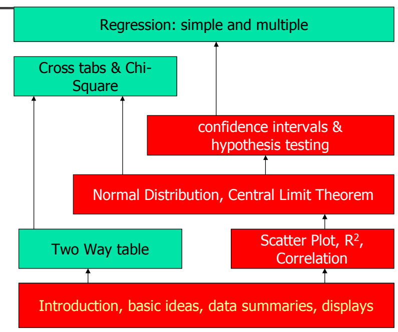
PUB – POS 316

Review (continue)

Navid Ghaffarzadegan
navidg@gmail.com
Last updated – May 1, 10

Course Road Map



---

9. t-test

- A typical question for hypothesis testing:
- You're an analyst for Ford. You want to find out if the average miles per gallon of Escorts is at least 32 mpg. Similar models have a standard deviation of 3.8 mpg. You take a sample of 60 Escorts & compute a sample mean of 30.7 mpg.
- At the 0.05 level, is there evidence that the miles per gallon is less than 32?

(source: Carnegie Mellon University, 90-711, Empirical Methods)

---

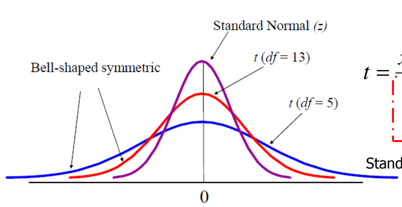
9. t-test

- A typical question for hypothesis testing:
- You're an analyst for Ford. You want to find out if the average miles per gallon of Escorts is at least 32 mpg. Similar models have a standard deviation of 3.8 mpg. You take a sample of 60 Escorts & compute a sample mean of 30.7 mpg.
- At the 0.05 level, is there evidence that the miles per gallon is at least 32?

(source: Carnegie Mellon University, 90-711, Empirical Methods)

---

9. t-test



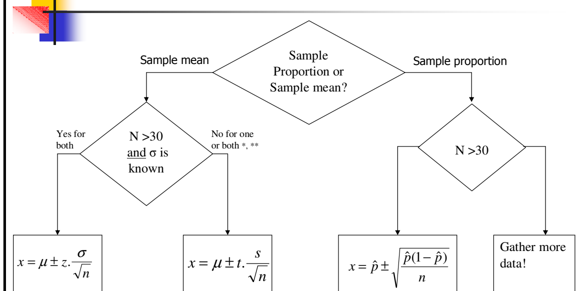
$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

Standard error (SE)

- •Degree of freedom = n-1 = (sample size-1)
- •Table

---

9. t-test (vs. z-test)

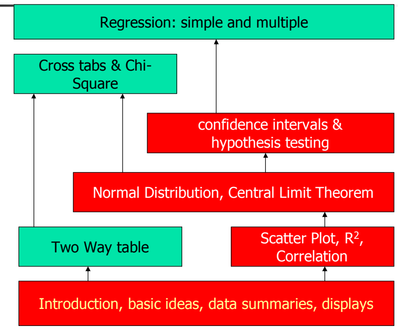


*: in case n is large, but you don't have because σ, you use $x = \mu \pm t.\frac{s}{\sqrt{n}}$, however your t will be very close to z.

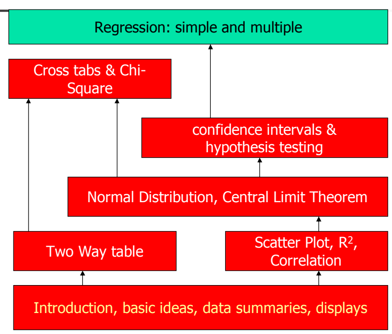
**: in case you have σ but your N is small, you need to use t, but you may follow: $x = \mu \pm t.\frac{\sigma}{\sqrt{n}}$

- **(9) t-test:** The average of monthly income per capita in our sample of study (n=36) is $3,500. The standard deviation in our sample is $1,200.
- a- state the 95% confidence interval for this finding.
- b- we would like to know if currently monthly income per capita in population is more than $3,200. State and test proper hypotheses.

Course Road Map



Course Road Map



Data in categories

- What is "data in categories?"
  - Two Variables are categorical:
    - X: Men or Women Y: Yes or No

	Gender	
Frequent of binge drinker	**Men**	**Women**
Yes	**1630**	**1684**
No	**5550**	**8232**

- How should we analyze this data?
- Joint Distribution: dist. of the whole data
- Conditional distribution, Marginal distribution

Data in categories

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

Conditional Distribution

	Gender	
Frequent of binge drinker	Men	Women
Yes	0.227019	0.1698265
No	0.772981	0.8301735

Systematic Investigation of two-way tables.

- How can we systematically compare two groups?
- Systematically:
  - With reporting the level of confidence.
  - Are we sure that the difference in two group is not just a matter of error in our study? (remember the issue of sampling vs. population?

Systematic Investigation of two-way tables.

- What do we expect to happen, if there is no systematic difference between male and female? (H_0)

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

	Gender		
Frequent of binge drinker	Men	Women	total
Yes			3314
No			13782
	7180	9916	17096

Systematic Investigation of two-way tables.

- If there were no difference between male and female the conditional distribution would have shown that.

- In another word, numbers should show:
  - proportion of male (out of total male) that are binge drinkers = proportion of female (out of total female) that are binge drinkers.

Systematic Investigation of two-way tables.

- What do we expect to happen, if there is no systematic difference between male and female? (H_0)

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

	Gender		
Frequent of binge drinker	Men	Women	total
Yes			3314
No			13782
	7180	9916	17096

Systematic Investigation of two-way tables.

- What do we expect to happen, if there is no systematic difference between male and female? (H_0)

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1391.818	1922.182	3314
No	5788.182	7993.818	13782
	7180	9916	17096

Systematic Investigation of two-way tables.

- What do we expect to happen, if there is no systematic difference between male and female? (H_0)

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

What we expect under the null hypothesis (No difference between male and female)

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1391.818	1922.182	3314
No	5788.182	7993.818	13782
	7180	9916	17096

Systematic Investigation of two-way tables.

- What do we expect to happen, if there is no systematic difference between male and female?

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
		9916	17096

Compare to see if we can reject the null hypothesis? Are we far enough from the null hypothesis?

	Gender		
Frequent of binge drinker	Men	Women	total
Yes	1391.818	1922.182	3314
No	5788.182	7993.818	13782
	7180	9916	17096

Chi-square, Chi-test

Frequent of binge drinker	Gender Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

Frequent of binge drinker	Gender Men	Women	total
Yes	1391.818	1922.182	3314
No	5788.182	7993.818	13782
	7180	9916	17096

- We can look at the difference between these numbers. Something like:
- (1630-1391)+(1684-1922)+(5550-5788)+(8232-7993)
- But again they cancel out! Can you guess what we should do?!
- This is what we look at:

$$X^2 = \frac{(1630-1391)^2}{1391} + \frac{(1684-1922)^2}{1922} + \frac{(5550-5788)^2}{5788} + \frac{(8232-7993)^2}{7933}$$

Chi-square, Chi-test

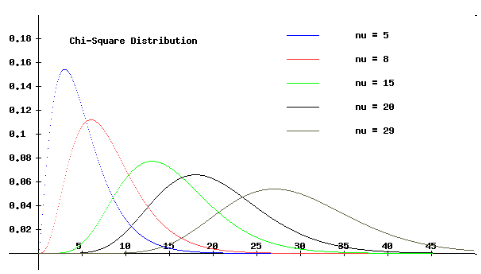
Frequent of binge drinker	Gender Men	Women	total
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

Frequent of binge drinker	Gender Men	Women	total
Yes	1391.818	1922.182	3314
No	5788.182	7993.818	13782
	7180	9916	17096

$$X^2 = \frac{(1630-1391)^2}{1391} + \frac{(1684-1922)^2}{1922} + \frac{(5550-5788)^2}{5788} + \frac{(8232-7993)^2}{7933}$$

- Now what should we do with this number?
- We know that as X^2 becomes larger, as we get far from the null hypothesis.
- In another word: P-value should decline.
- BUT X^2 does not follow z or t distributions!.. It follows X^2 (Chi-Square distribution)

Chi-square, Chi-test



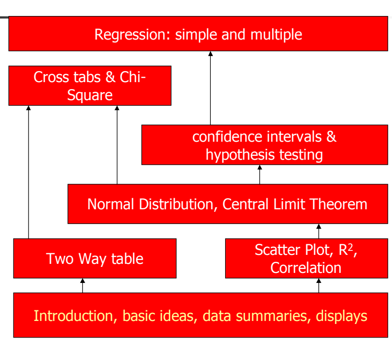
df = (number of rows – 1)*(number of columns – 1)

Data in categories

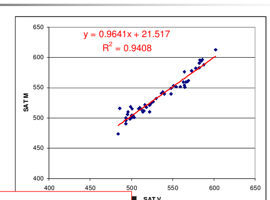
- **(1) Chi-Square:** We would like to compare Male Students' performance with Female students' performance in a stat course. The following table shows the number of people in each group that got a grade above and below B.
- (a) In a chi-square analysis of this table, what is the null hypothesis?
- (b) Fill in the table below with the numbers that would be expected under the null hypothesis.
- (c) The chi-square statistic for this analysis turns out to be 1.87. How was that computed?
- (d) How many degrees of freedom are present in this data?
- (e) Do you accept or reject the null hypothesis you state in (a)?
- (f) What can you say about the p-value of your conclusion in (e)?

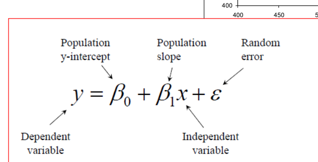
	Grades above B	Grades below B	Totals
Female	7	7	14
Male	8	20	28
Totals	15	27	42

Course Road Map



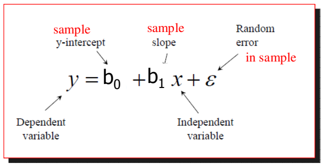
Least Square Regression





Least Square Regression

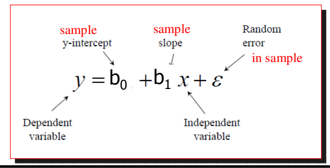
- What if we do not have the complete information about our population?



What does estimation of slope and intercept mean? (b estimation of β)

Tests for significance and CI

- What will happen for the slope and intercept if we conduct the study many times?
- The important question: <u>Are you confident enough that the slope is not zero? ($\beta_1 \neq 0$)</u>



Example

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.382981264					
R Square	0.146674648					
Adjusted R Square	0.12851879					
Standard Error	92.39064342					
Observations	49					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	68959.52299	68959.52299	8.078640185	0.006606171	
Residual	47	401193.4566	8536.030992			
Total	48	470152.9796				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363	5.983660873	2.84931E-07	267.6448515	538.7640727
X Variable 1	22.72341076	7.99474077	2.84229488	0.006606171	6.640067123	38.80675439

Check if the slope is significantly different from zero..
That's **the most important** thing

Analysis of Variance (ANOVA)

- <u>ANOVA:</u>

  <u>An</u>alysis **<u>o</u>f** **<u>Va</u>**riance

- As you have seen in this class, we are very interested to learned about variance (or standard deviation) in a data set. Remember?
- How can we explain why there is a variation in a data set?

Analysis of Variance (ANOVA)

What you need to remember:

- <u>F shows if your regression shows anything at all.</u> (or it is just a random pattern between your x and y).
- Excel reports F, compares it with F-table, reports p-value. Just we should be able to read it and know what it is about.

Example

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.382981264					
R Square	0.146674648					
Adjusted R Square	0.12851879					
Standard Error	92.39064342					
Observations	49					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	68959.52299	68959.52299	8.078640185	0.006606171	
Residual	47	401193.4566	8536.030992			
Total	48	470152.9796				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363	5.983660873	2.84931E-07	267.6448515	538.7640727
X Variable 1	22.72341076	7.99474077	2.84229488	0.006606171	6.640067123	38.80675439

Simple regression

- (2) We run a simple regression to see if there is any association between Graduate schools tuition (x-variable) and the number of applicant (y-variable). The sample size is 42. The regression estimates the intercept to be 2 and the slope to be - 1.56, with the standard errors of 1.5 and 0.73 respectively.
- (a): Based on this result can we state that more tuition results in fewer applicants?
- (b): Assume F-stat for this regression results in F=2.5 (p<0.01). How do you interpret this number?

Multiple regression

- Multiple regression:

Sample y-intercept

Sample coefficients

$$y = b_0 + b_1.x_1 + b_2.x_2 + b_3.x_3 + ... + b_n.x_n + \varepsilon$$

Dependent variable

Independent variables

Multiple regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.650391
R Square	0.423009
Adjusted R	0.384543
Standard E	77.64221
Observatio	49

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	3	198878.9	66292.97	10.99694	1.54E-05
Residual	45	271274.1	6028.312		
Total	48	470153			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*lpper 95.0%*
Intercept	651.2714	86.78105	7.504765	1.83E-09	476.4854	826.0574	476.4854	826.0574
X Variable	-112.376	24.20945	-4.64183	3E-05	-161.136	-63.6158	-161.136	-63.6158
X Variable	27.78912	7.370251	3.770444	0.000473	12.94467	42.63357	12.94467	42.63357
X Variable	-15.5316	8.306982	-1.86971	0.068039	-32.2628	1.199484	-32.2628	1.199484

Simple Reg: SATM=403+22.72*HighSchoolMath

Multiple Reg: SATM=545-94*Gender+22.12*HighSchoolMath

Multiple Reg: SATM=651-112*Gender+27.78*HighSchoolMath-15.53*HighSchool Science

- We can still add more variables to the right side!

Multiple regression

- (3) We would like to see how different factors influence employees' performance in different organizations. Based on the available data (n=85), we run a regression whereby employees' performance is our y-variable (dependent variable). Our x-variables (independent variables) are job satisfaction (JS), average salary (S), management performance (M), organizational conflicts (OC), and governance (G).
- We get the following table.

	Coefficients	**Standard Error (SE)**	**t**	**p**
Intercept	5	1.5	3.33	0.003
JS	1.1	0.3	3.67	0.000
S	0.3	0.2	1.50	0.13
M	2	0.35	5.71	0.000
OC	-1.5	0.5	-3.00	0.004
G	0.2	0.18	1.11	0.26

- (a): Interpret the table by 1) stating a function for employees' performance, and 2) listing the coefficients that are significant.
- (b): In the regression, G, represent organizational governance, whereby it is equal to 0 for public organization and is equal to 1 for private organizations. People believe that in private organizations, organizational performance is higher than in public organizations. Do you think based on the regression we have any support for this argument?
- (c) Look at the row for OC. Based on the coefficient and SE, re-calculate t and p-value.

Course Road Map