NOTE: this review emphasizes basic statistics concepts. It does not cover materials that WILL be on the exam, including: graphic presentation of data (Tufte); survey research/sampling approaches; economic analysis (e.g., multipliers, location quotients, interpretation of GINI coefficients); demography (e.g., life tables, age pyramids); case studies (e.g., analytical vs. statistical generalization). See syllabus and study guide for a more complete list.

Here is a SPSS regression output from the familiar world95.sav dataset:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	GDPcap1000, Females who read (%), Average female life expectancy	.	Enter

a. All requested variables entered.

b. Dependent Variable: Fertility: average number of kids

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.872[a]	.760	.751	.9410

a. Predictors: (Constant), GDPcap1000, Females who read (%), Average female life expectancy

R-square (adjusted) is omitted: would it be higher, lower, or the same as R Square?

(Rsquare adj is always lower or the same as R2.)
R2 = 1- (1 - R2)((n - 1) / (n - k - 1))

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	226.531	3 = k	75.510	85.283	.000[a]
	Residual	71.718	81 = n-k-1	.885		
	Total	298.249	84 = n-1			

a. Predictors: (Constant), GDPcap1000, Females who read (%), Average female life expectancy

b. Dependent Variable: Fertility: average number of kids

Write in the values for degrees of freedom (for Regression and Residual) – see blanks above.
Calculate F. *75.510 / .885 = 85.283*
Would it be significant at the 0.05 level? *__YES_____*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.394	.921		11.287	.000
	Females who read (%)	-.035	.006	**-.528**	-5.546	.000
	Average female life expectancy	-.059	.018	-.335	-3.185	.002
	GDPcap1000 (GDP per capita in $1000)	-.034	.027	-.085	-1.279	.204

a. Dependent Variable: Fertility: average number of kids

Calculate the t-score for per capita GDP. *= B/std error = -.034/.027 = -1.279* ***Would it be significant at the 0.05 level?*** *NO*

Which variable seems to have the most explanatory power? *_Females who read_____*
What statistic do you use to answer the above question? *Beta (standardized coefficient) = b stdev(x) / stdev(y)*

Can you write the regression equation?
Fertility rate = 10.394 - .035 Female Lit - .059 Female e_0 - .034 GDPpercap(1000s)

Is this the final regression model, or do you need to rerun the analysis with a revised set of variables?

NO – rerun the model with just the significant variables (exclude GDP)

Rank these US Census Geography categories in order (1 – 6) from smallest to largest:

__3_ **tract**

__2_ **block group**

_5__ **Division** (groups of states: there are 9 total, e.g., "East North Central")

_6__ **Region** (groups of divisions: there are 4 total, eg., "Midwest")

_1__ **Block**

__4_ **Metropolitan Area** (built around cities and their hinterlands)

In a survey of seven houses, the number of bedrooms was: 1, 2, 2, 3, 4, 5, 11

Determine the : mean __4__ median _3___ mode _2___

Difference of means test: using the world95.sav data set, a researcher uses SPSS to determine whether there is a statistically significant difference between fertility rates in nation-states where the predominant religion is Catholicism vs. all other religions. Below is the (abridged) SPSS output, with the Sig. levels (p-values) deleted. Is the difference of means significant at the .05 level? No, since |t|<1.96

Group Statistics

	catholic	N	Mean	Std. Deviation	Std. Error Mean
Fertility: average number of kids	1	41	3.138	1.6865	.2634
	0	65	3.819	2.0061	.2488

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Fertility: average number of kids	Equal variances assumed	-1.806	104	.074	-.6807	.3769	-1.4280	.0666
	Equal variances not assumed	-1.879	95.645	.063	-.6807	.3623	-1.4000	.0386

Evaluation - a fictional research project

41. A research team is studying the link between the built environment and the likelihood that children will walk or bike to school (as opposed to being driven by car or bus). The research team wants to know if there is any empirical proof to the recent argument that postwar American suburban environments create long and car-friendly but walking-unfriendly distances between residences and public schools **[A]**.

The team examines typical postwar, car-dependent sprawling suburbs with low population density **[B]**. They combine direct observation and survey work to estimate the mode of travel to school (walk, bike, car, bus, public transit). As a point of comparison, the research team also conducts this same survey in twelve older (i.e., prewar) communities with higher density, mixed use settlement patterns **[C]**. After analyzing the data, the team finds that only 15 percent of school children in the suburban communities either walk or bike to school **[D]**. By contrast, 47 percent of school kids walk or bike to school in the older, higher density communities **[E]**. The research team argues that the difference between the two sets of numbers is due to the differences in the built environment.

After the report is released, a rather skeptical Institute for the Preservation of the Suburban Way of Life **[F]** challenges the study's conclusion. The institute claims that the difference in the mode of travel values is due to differences in income and other socio-economic characteristics between newer suburban and older urban environments, not due to differences in the built environment **[G]**.

Identify each of the following five concepts in the scenario (and write the letter on the corresponding line):

__E__ counterfactual (what would have happened without the "intervention" of suburbia)

_A___ program theory

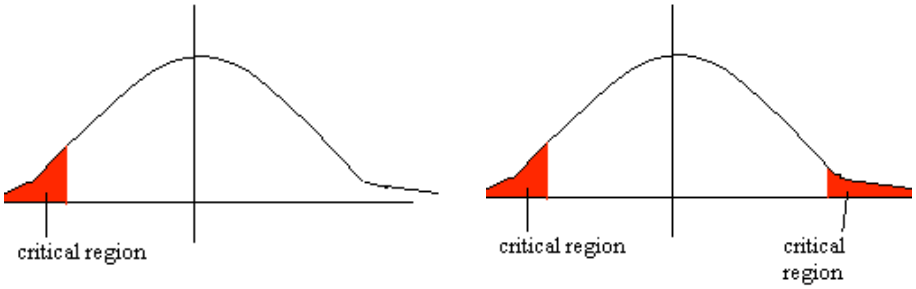
_B___ experimental group

__G__ rival explanation

_C___ control group

[Note: there are five terms and seven letters [A-G], so not all letters are used.] ALSO: we assume here that the "experiment" is surburbia, but that is not so obvious. One could conceivably reverse the control and the experiment.

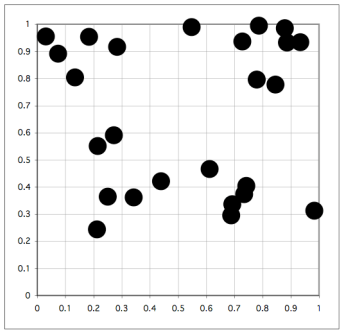
One-Tailed or Two-Tailed Test?



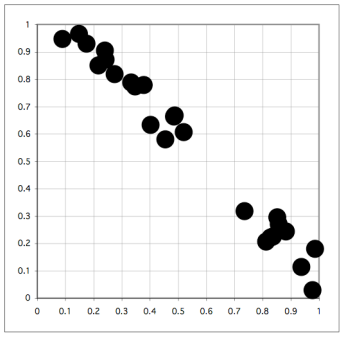
Difference of means test…

If you hypothesize that men earn more than women, then you would use a **one**- or two-tailed test?
If you hypothesize that the earnings of left-handers and right-handers are different (without anticipating direction of difference), would you use a one- or **two-tailed** test? (circle the correct answer)

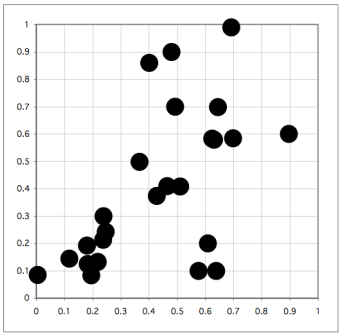
images from : http://www.mathsrevision.net/alevel/pages.php?page=64



A



B



C

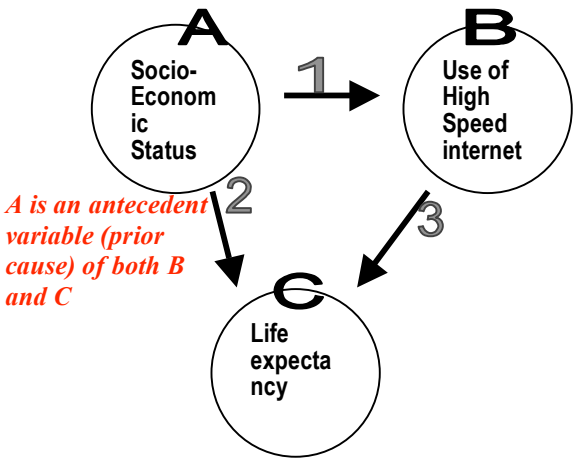
Match the x-y scatterplot to the correct correlation value (r) and statistical significance of the relationship (probability value):

R (correlation)	*Scatterplot (A,B, or C?)*
-0.9877	***B***
-0.1988	***A***
+0.5416	***C***

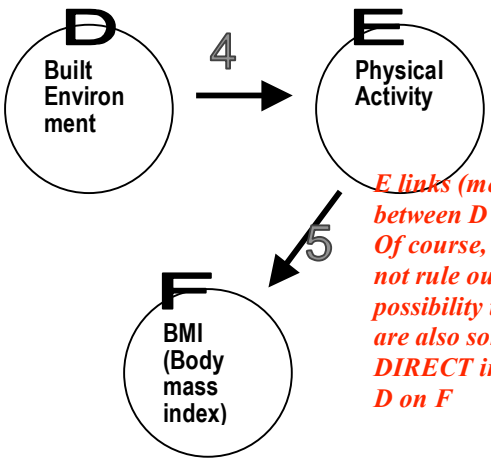
p-value	*Scatterplot (A,B, or C?)*
0.0000	***B***
0.0689	***C***
0.5356	***A***

Note: you can't determine p-values from the graph per se, but you do know that (given equal sample size), stronger linear bivariate relationships (e.g., higher |r|) are more statistically significant (i.e., lower p-values).

Find the spurious relationship (1-5) _3____ and the intervening variable (A-F) _E_____



A is an antecedent variable (prior cause) of both B and C



E links (mediates) between D and F. Of course, this does not rule out the possibility that there are also some DIRECT impacts of D on F