Using SPSS for Basic Analyses

This chapter introduces procedures for performing three kinds of data analysis with SPSS: t-test, correlation, contingency table analysis.

The t-test

Statistical concept

The t-test is used to compare mean values of groups within a sample when the measured variable is interval or ratio level. Two kinds of t-tests are used to compare means in two types of designs: between-group designs and repeated-measure designs. Between-group designs include two different groups of subjects, such as a group of men and a group of women. Repeated-measure designs include one group of subjects from whom some sort of measurement is obtained at two points in time. For example, a group of people could be given a pre-test on the Beck Depression Inventory, participate in some psychotherapy, then be given the BDI again to see if the therapy worked. When the design has more than two groups, or more than two measurements over time, the t-test cannot be used (analysis of variance is used instead). The between-subjects design t-test is termed "independent t-test" or "uncorrelated t-test." The repeated-design t-test is termed "dependent t-test" or "correlated t-test." We will only discuss the between groups case in this chapter.

The t-test follows the logic of all analyses that involve comparisons of means: The degree of difference between the means is compared to the amount of variability within the sample. The t value ("Student's t," named after a modest statistician who called himself "student," not after you) is the ratio of these two values, with some other adjustments. Therefore, as the mean difference increases, t increases, and as the variability within the sample increases, t decreases. The within-sample variability is the "noise" in the data and is referred to as error variance. If everyone in a group responded precisely the same way, there would be zero error in that group; if the same happened in the other group, overall error variance would be zero. Under this circumstance, even the smallest mean difference would be important. However, this situation never occurs because living organisms are very complex and never all act exactly the same way. Hence, we are always laboring to do experiments in which the mean difference is large compared to the error variance.

The null hypothesis for a t-test is: H_0: $u_1 = u_2$

where u_1 and u_2 are the population means of the two groups.

Your task is to reject this hypothesis, that is, support the alternate hypothesis that the two means are different. When t gets large enough (the ratio of the mean difference to the error variance), we can safely say that the null hypothesis could not be true, and reject it. Recall that we need to be certain with a 5% (alpha=.05) level of Type I Error before we can do that.

Besides the mean difference and the error variance, the sample size also affects the size of t. Larger samples provide us with more confidence that whatever difference we think we see is in fact present and not a fluke of sampling. Therefore, larger samples produce a larger t, and they also affect the size of t necessary to reject the null hypothesis. For example, if the sample size (both groups combined) were 12, t must be at least 2.23 to reject the null; but if it were 102, t need only be 1.98. For this reason, it is always better to obtain larger samples if resources permit.

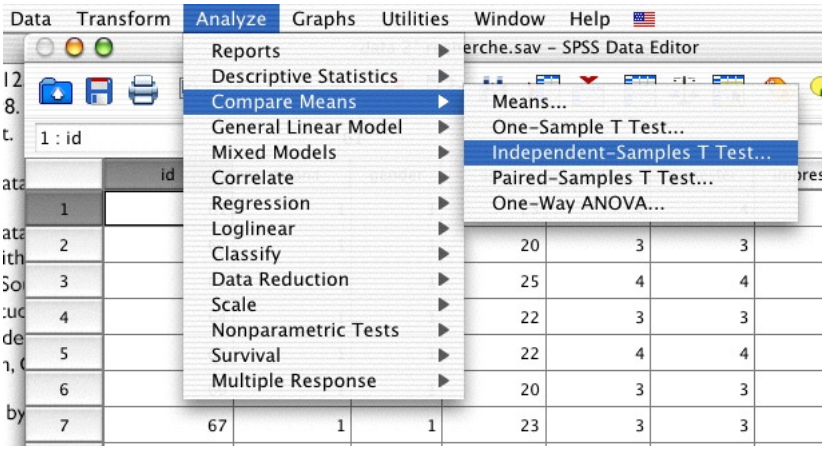
The dataset

The dataset used in this chapter is from Cécile Morvan's Advanced Methods project, with additional data collected a year later. The results were presented at the 2003 Southeastern Psychology Association convention in New Orleans as "title." The study examined cultural differences in the 16PF personality test, a measure of academic behaviors and values, and a measure of adjustment to life in America. French, Caribbean, and American students served as participants in the study.

T-test by SPSS

T-tests are very easy to perform in SPSS. Cécile's study measured physical symptoms often associated with poor adjustment, including frequency of headaches. A five-point scale ranging from "never" to "very frequently" was used to assess the symptoms.

Do men and women differ in frequency of headaches? Because the five-point scale is considered to be an interval level scale and only two genders were used, the appropriate analysis is an independent t-test. To find this analysis in the Analysis menu:



The procedure:

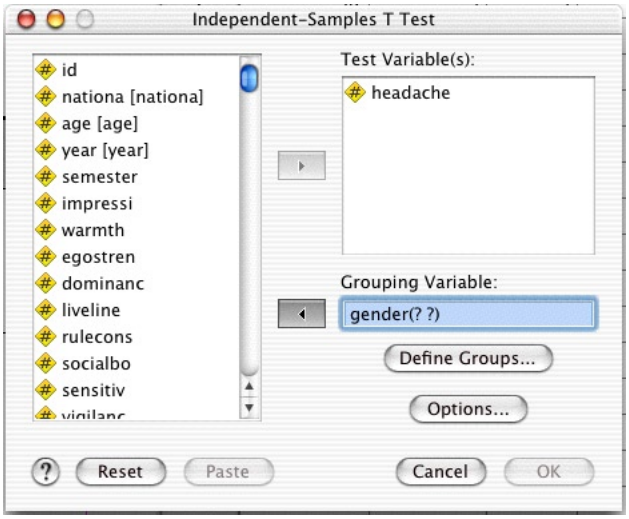
1. Choose the dependent variable and move it into the Test Variables field.

2. Choose the independent variable and move it into the Grouping Variable field

3. Click on Define Groups...

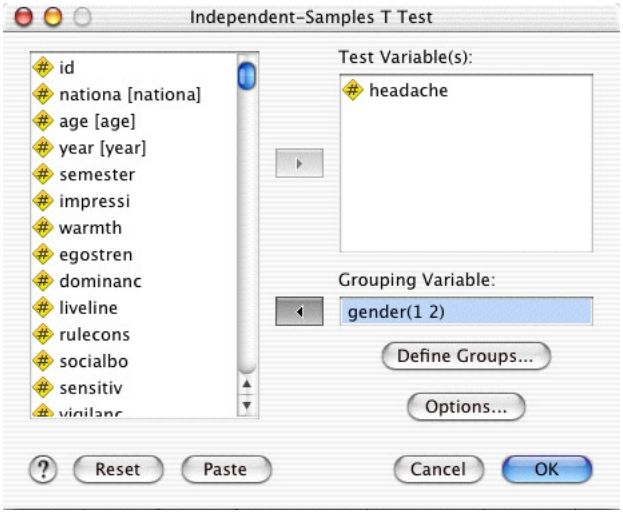
4. Indicate the value of the IV that will be one of the groups in the analysis (it doesn't matter which one). Do the same with the other. Click Continue.

5. Click OK to run the analysis.

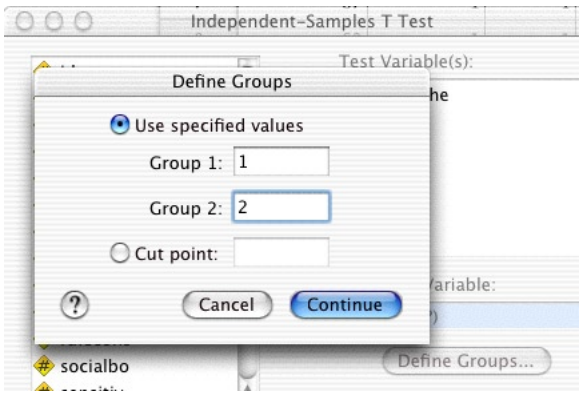


Syntax:

```
T-TEST
  GROUPS=gender(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=headache
  /CRITERIA=CIN(.95) .
```



The results come as two tables. The first table presents descriptive statistics for the two groups. The second presents the t-test results. For the purposes of this class, use the two row of the second table, "Equal variances assumed." The t value, degrees of freedom, and p values are the most important parts of this table. The t value, in this case 3.306, is very impressive. Ignore the negative sign: it only indicates that Group 2 (females) was higher than Group 1. Degrees of freedom (df) reflects the sample size (df = N-2) and is explained in a statistics class. The p value indicates the probability of Type I Error (rejecting the null when it is actually true) for the analysis. Recall that we will tolerate no more than a .05 probability of Type I Error, and since .001 is even lower than this, we can be confident that a Type I Error has not occurred. Altogether, the result would be called "statistically significant," meaning that we will reject the null hypothesis. If we reported this result in a paper, we might write:



➔T-Test

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
HEADACHE	male	45	2.33	.929	.139
	female	44	3.09	1.217	.183

> Females reported suffering headaches more frequently than males, $t(87)=3.31$, $p<.05$, *M*s = 3.1 and 2.3, respectively.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HEADACHE	Equal variances assumed	3.960	.050	-3.306	87	.001	-.76	.229	-1.213	-.302
	Equal variances not assumed			-3.296	80.468	.001	-.76	.230	-1.215	-.300

Just because the results are statistically significant doesn't mean that they are psychologically or practi-

cally significant. The finding reported here is quite strong, but whether or not it is large enough to justify, say, changing the way the counseling center deals with male and female students is a different issue.

Correlation

Statistical concept

Correlation indicates the degree of relationship between two variables, that is, the extent to which they covary. As one variable goes up, the other goes up, and vice-versa. In the 19th Century, Karl Pearson invented a way to represent this relationship in a number, the Pearson Product-Moment Correlation Coefficient, r. Values of r range from zero to 1, where zero indicates no relationship and 1 indicates a perfect relationship. If the relationship is inverse (one variable goes up as the other goes down), the value of r becomes negative.

More example, we expect a positive relationship between intelligence (IQ) and SAT scores. As IQ rises, SAT rises. Is the relationship perfect (r=1)? No.

The correlation coefficient is not a linear representation of the strength of the relationship, so it can be deceptive. A correlation of r=.50 seems to be twice as strong as r=.25, but for mathematical reasons the real strength of the relationship is the correlation squared. $.50^2 = .25$, while $.25^2 = .06$. So this tells us that a .50 correlation is about 4 times as strong as a .25 correlation. It also shows that a .50 correlation is not half way to perfect, but only a quarter of the way. The halfway point is r=.707. The correlation coefficient squared is called the coefficient of determination, and we can say that r=.50 means "25% of the variability in SAT is accounted for by IQ." "Accounted for" means "explained by." If you know a person's IQ, you can explain or predict 25% of his/her SAT scores. (I am just making these numbers up.) This means that 75% of the variability in SAT is "unaccounted for," meaning still a mystery.

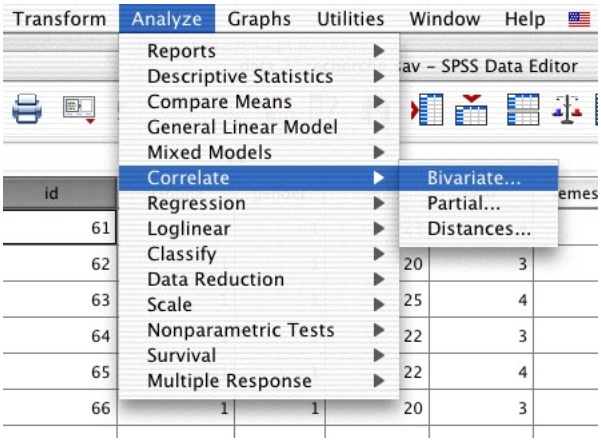
The null hypothesis for correlation is: $H_0: \rho = 0$

where ρ is the Greek symbol rho.

Rho is used to indicate that we are inferring from the sample to the population, ultimately. The researcher's task is to find a correlation large enough to be able to claim that rho cannot possibly be zero; reject the null.

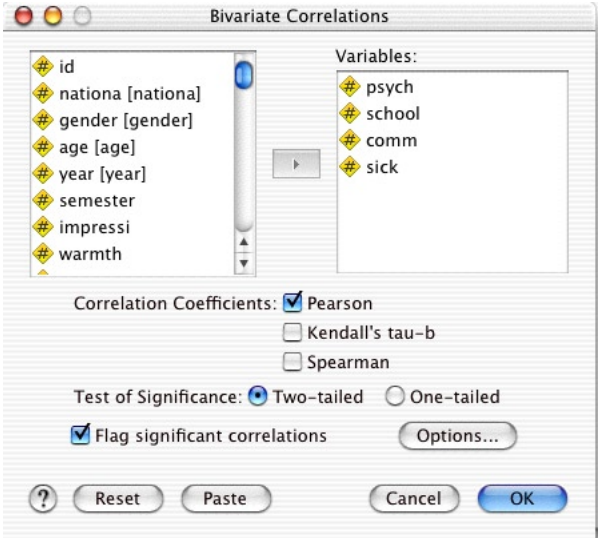
Correlation in SPSS

Access the correlation procedure from this menu:



Bivariate means "two variables." In this course, we will only use bivariate correlations, that is, correlations that represent the degree of relationship between two variables at a time.

To run the analysis, simply enter the variables for which you would like to compute correlations in the Variables field. Correlations for all possible combinations of variables will be calculated. Normally, leave the other options that appear on this screen as they are.



Syntax:

```
CORRELATIONS
  /VARIABLES=psych school comm sick
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

The output of the correlation procedure is a correlation matrix.

➔Correlations

Correlations

		PSYCH	SCHOOL	COMM	SICK
PSYCH	Pearson Correlation	1	.605**	.279**	.190
	Sig. (2-tailed)	.	.000	.009	.074
	N	89	89	88	89
SCHOOL	Pearson Correlation	.605**	1	.191	.060
	Sig. (2-tailed)	.000	.	.075	.577
	N	89	89	88	89
COMM	Pearson Correlation	.279**	.191	1	.269*
	Sig. (2-tailed)	.009	.075	.	.011
	N	88	88	88	88
SICK	Pearson Correlation	.190	.060	.269*	1
	Sig. (2-tailed)	.074	.577	.011	.
	N	89	89	88	89

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The correlation matrix is symmetrical around the diagonal, so the correlations above the diagonal are the same as the correlations below it. The relationship of a variable to itself is perfect, r=1 (on the diagonal).

The first number in each cell of the table is the correlation coefficient, the number in the middle is the p value, and the number at the bottom is the number of subjects for which the correlation was calculated (which could vary if some subjects have missing data for just some variables).

In this part of Cécile's dataset, several types of adjustment variables are examined.

PSYCH: Psychological distress (depressed, anxious, lonely...)
SCHOOL: Problems with school (poor grades...)
COMM: Problems communicating (people can't understand what I mean...)
SOCIAL: Problems with social life (can't make friends...)

A large correlation emerged for PSYCH and SCHOOL, r = .605. The p value is quite low, in fact it is too small to print (p = .000) so it must be .0005 or smaller. 89 subjects were used in the analysis. Given this outcome, we can reject the null hypothesis that the correlation is zero and say that it is "statistically significant." As psychological distress increases, school problems increase. In writing:

> A positive relationship was found between psychological distress and school problems, $r(87) = .60$, $p < .05$.

Note that the value 87 is the degrees of freedom, not the sample size (df = N-2).

Does this finding mean that psychological distress causes school problems? Not necessarily. Correlation does not mean causality. Maybe students who are flunking out become distressed. Or maybe a failed relationship causes both psychological and school problems. Interpreting the correlation requires both theoretical conceptions and additional data.

Contingency Table Analysis

Statistical Concept

We often need to look at the relationship between nominal variables. Nominal

variables are those in which was simply name or categorize something, such as categorizing subjects as male or female, or as young or old. After we categorize subjects, we can count how many fall into each category: how many males were in the sample, how many were young, etc. If we have two category variables, we might need to know if there is a relationship between the variables. In this case, a relationship means that the frequencies on one variable differ as a function of the other. For example, the frequency of old vs. young subjects may differ as a function of gender: perhaps there are more older males and more younger females in the samples. Put another way, knowing where a subject is categorized on one variable tells us something about where he/she is categorized on another: knowing the subject is a female suggests that she is probably in the younger category.

Two variables such as gender and age form a 2x2 table. The cells contain counts or frequencies. The bottom row an the right column are marginal totals. This table includes the cell values observed in a hypothetical sample:

	Gender		
Age	*Male*	*Female*	*Total*
Young	*20*	*40*	*60*
Old	*40*	*20*	*60*
Total	*60*	*60*	*120*

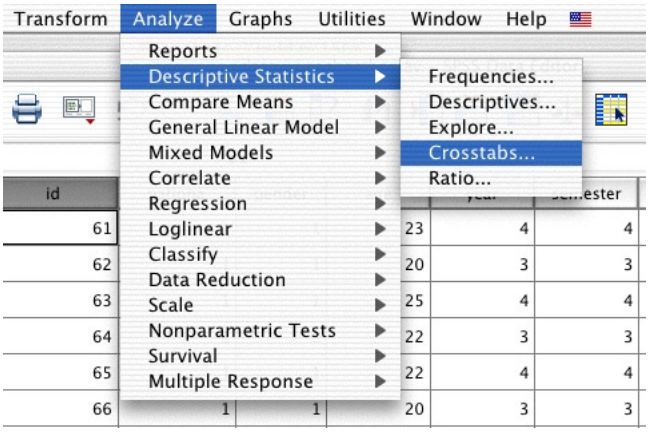
The first step, conceptually, in this analysis is to determine what the table would look like under the null hypothesis: H_0: No relationship between Age and Gender. If there were no relationship, then all the cells would be the same. Of the 60 young subjects, half (30) would be male and half would be female. Turned sideways, of the 60 male subjects, half (30) would be young and half would be old. Under the null hypothesis table, knowing that a subject is young would give no clue as to gender. This is the null hypothesis table, usually termed the expected frequencies table:

	Gender		
Age	*Male*	*Female*	*Total*
Young	*30*	*30*	*60*
Old	*30*	*30*	*60*
Total	*60*	*60*	*120*

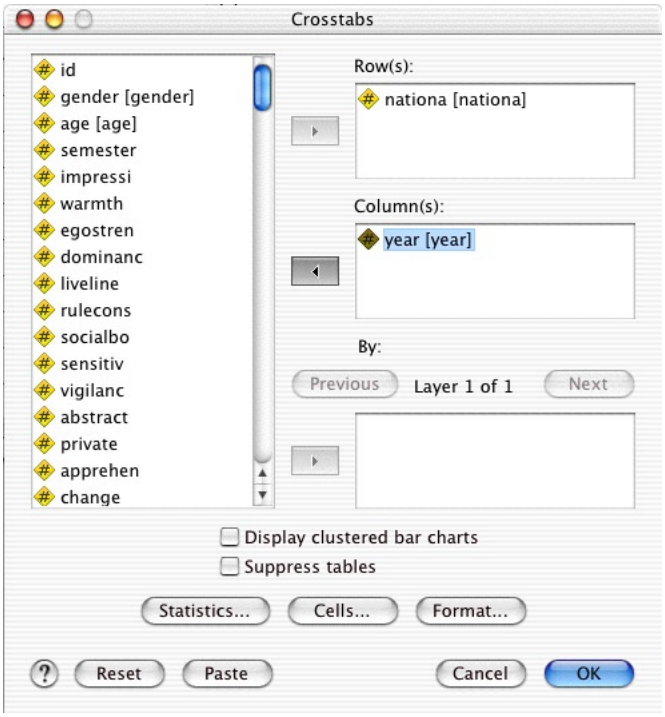
The second step is to determine the extent to which the observed table deviates from the expected table. Cell by cell, you can calculate how different the tables are. The null table expects 30 young males, but the observed table has 20, for a deviance of 10. These deviances are combined in a formula to produce a value called chi-square, χ^2. As the amount of deviation increases, χ^2 increases. At some point, the value of χ^2 gets sufficiently large that we can say with some certainty that the deviation is so large that the null hypothesis is probably not true. In this example, the value of χ^2 is 13.3 and it is statistically significant. Hence, there is a relationship between gender and age. As age increases in the sample, gender becomes male.

Crosstabs in SPSS

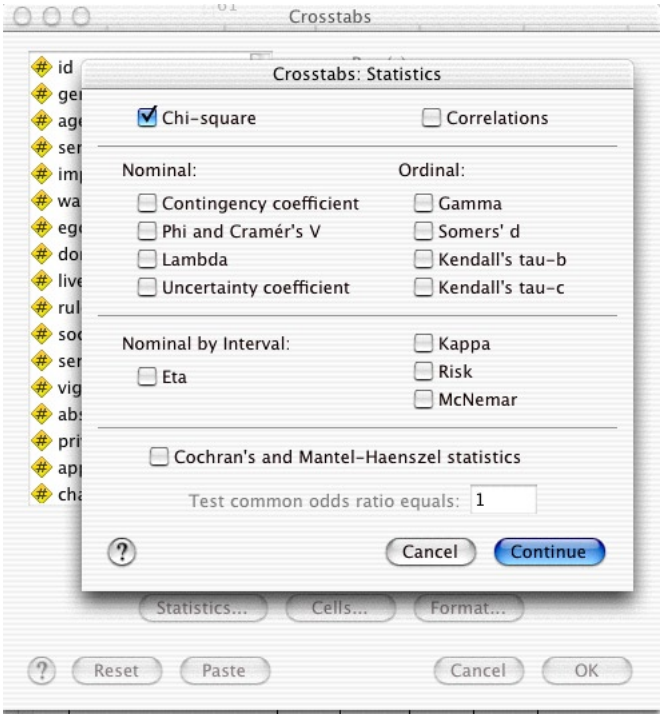
The SPSS procedure for contingency table analysis is called crosstabs.



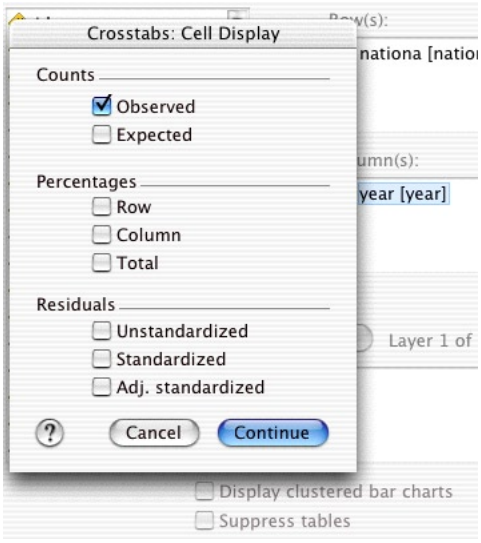
Specify the row and column variables by moving them into the appropriate fields. Which variable goes where is not important.



Tell SPSS to compute the Chi-square value by clicking on the Statistics button and selecting Chi-square:



Click on Cells... to tell SPSS what kind of cell data to output, and whether or not to include percentages along with counts.



In this example, the relationship between nationality (American, French, Caribbean) and year in school (Freshman, etc.) in Cécile's sample was examined. The question is: are the three nationalities distributed over year in school the same way?

Syntax:

```
CROSSTABS
  /TABLES=nationa  BY year
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT .
```

The output, showing only observed frequencies:

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
nationa * year	90	100.0%	0	.0%	90	100.0%

nationa * year Crosstabulation

Count

		year				
		freshman	sophomore	junior	senior	Total
nationa	French			18	12	30
	American	9	13	5	3	30
	Carribean	4	14	11	1	30
Total		13	27	34	16	90

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	43.286[a]	6	.000
Likelihood Ratio	54.925	6	.000
Linear-by-Linear Association	20.245	1	.000
N of Valid Cases	90		

a. 3 cells (25.0%) have expected count less than 5. The minimum expected count is 4.33.

The bottom table presents the chi-square test. It can be seen that the chi-square value is high and significant. In writing:

> A relationship was found between Nationality and Year in school, $\chi^2(6) = 43.3$, $p < .05$. All French students were Juniors and Seniors, while Americans tended to be Freshmen and Sophomores and Caribbeans tended to be Sophomores and Juniors.

The relationship between Nationality and Year in this example is not a simple relationship as we found in the correlation between psychological distress and school problems. For example, we could have found that as Nationality proceeded from American to Caribbean to French, students became progressively more advanced, from Freshman Americans to Senior French. But this did not happen. The form of the relationship is essentially: Year in school differs as a function of Nationality.