TABLE PERCENTAGES AND ASSOCIATION BETWEEN VARIABLES

Let us consider the following hypothesized association from Handout #9 (p.2):

RELIGIOUS AFFILIATION ==========> PRESIDENTIAL VOTE [*individuals*]
(Protestant vs. Catholic) (Dem. vs. Rep.)

Suppose we collect appropriate data and run the crosstabulation, and that it looks like this:

TABLE 1A. VOTE BY RELIGION (Absolute Frequencies)

	RELIGION		
PRES VOTE	Protestant	Catholic	***Total***
Dem.	300	300	600
Rep.	350	50	400
Total	650	350	1000

Is there an association between these variables and, if so, what is its direction and how strong is it? [We will discuss this in class.] It is harder to "see" any association (or lack of it) in this table than in the hypothetical WHETHER/NOT VOTE by LEVEL OF INTEREST tables on pp. 2-3 of Handout #10, because this table has *non-uniform marginal frequencies* — that is, we do not have equal numbers of Protestants and Catholics nor of Democratic and Republican voters. This means that it is not immediately evident what a crosstabulation displaying a zero association between RELIGION and VOTE would look like. In the VOTE by INTEREST example with uniform marginal frequencies, we saw than a zero association meant that cases would uniformly distributed (or "evenly spread") over the four (interior) cells of the table. But we cannot have such a simple pattern here because the rows and columns must add up to the specified (non-uniform) marginal frequencies.

Notice that the sample as a whole (i.e., the set of 1000 cases) is divided 60% to 40% between Democratic and Republican voters. In the event RELIGION had no association with (and no apparent influence on) VOTE, we would expect that Protestants and Catholics would vote Democratic vs. Republican, not necessarily in a 50%-50% proportion, but *in the same proportion as the population as a whole* (and as each other); hence $0.6 \times 650 = 390$ Protestants would vote Democratic and $0.4 \times 650 = 260$ would vote Republican; likewise $0.6 \times 350 = 210$ Catholics would vote Democratic and $0.4 \times 350 = 140$ would vote Republican.

Notice also that the population as a whole is divided 65% to 35% between Protestants and Catholics. In the event RELIGION had no association with (and no apparent influence on) VOTE, we would expect that the Democratic and Republican voters would be Protestants and Catholics, not in necessarily in a 50%-50% proportion, but again *in the same proportion as the population as a*

whole (and as each other); hence 0.65 × 600 = 390 Democratic voters would be Protestants 0.35 × 600 = 210 would be Catholic; likewise 0.65 × 400 = 260 Republican voters would be Protestants and 0.35 × 400 = 140 would be Catholics. Note that these two sets of calculations both produce the same *expected frequencies* shown in Table 1B below.

TABLE 1B. ZERO ASSOCIATION BETWEEN VOTE AND RELIGION

	RELIGION		
PRES. VOTE	Protestant	Catholic	***Total***
Dem.	.6×650 =.65×600 = 390	.6×350 =.35×600 = 210	600
Rep.	.4×650 =.65×400 = 260	.4×350 =.35×400 = 140	400
Total	650	350	1000

Given this table displaying expected frequencies in the absence of association, we can see that Table 1A shows that there are in fact *more cases* in the Dem-Cath and the Rep-Prot cells, and conversely *fewer cases* in the Dem-Prot and Rep-Cath cells, than would be the case if there were zero association. So we can conclude that (i) there is an association between the variables and (ii) its direction is this: Catholics vote Democratic more than Protestants do and Protestants vote Republican more than Catholics do.

How strong is this association between RELIGION and VOTE? This is equivalent to asking where Table 1A stands in relation to Table 1B showing zero association and a table showing maximum association between the variables. In the VOTE by INTEREST example that introduced Handout #10, a maximum association was exemplified by a table in which everyone with high interest votes and no one with low interest votes. By the same token, it might seem that if there were a maximum association between RELIGION and VOTE (in the specified direction), every Catholic would vote Democratic and every Protestant would vote Republican. But the latter stipulation cannot be fulfilled, since there are 650 Protestants but only 400 Republican voters, so *at most* 400 Protestant can vote Republican. Table 1C shows the maximum possible association between in the variables the same direction exhibited in Table 1A.

TABLE 1C. MAXIMUM ASSOCIATION BETWEEN VOTE AND RELIGION

	RELIGION		
PRES. VOTE	Protestant	Catholic	***Total***
Dem.	250	350	600
Rep.	400	0	400
Total	650	350	1000

If there were no association, there would be 210 cases in the Dem-Cath cells (Table 1B); if there were maximum association, there would be 350 cases in the Dem-Cath cells (Table 1C). In fact, there are 300 cases in the Dem-Cath cells (Table 1A), so in this sense Table 1A is somewhat closer to Table 1C than to Table 1B. Thus we might expect a measure of association to have a value somewhat closer to 1 than to 0, i.e., somewhat greater than 0.5.

Let's also consider what a table displaying a maximum association between the variables but in the opposite direction would look like. The VOTE by INTEREST example suggests that this would mean every Catholic votes Republican and every Protestant votes Democratic. But again the latter stipulation cannot be fulfilled, since there are 650 Protestants but only 600 Democratic voters, so *at most* 600 Protestants can vote Democratic. Table 1D shows maximum possible association between in the variables in the opposite direction from that exhibited in Tables 1A and 1C.

TABLE 1D. MAXIMUM ASSOCIATION BETWEEN VOTE AND RELIGION (OPPOSITE DIRECTION)

	RELIGION		
PRES. VOTE	Protestant	Catholic	***Total***
Dem.	600	0	600
Rep.	50	350	400
Total	650	350	1000

If all this seems a bit confusing, you will glad to learn that there is another more intuitive and transparent way to "see" an association in a crosstabulation. This is accomplished by converting the *absolute frequencies* (or *case counts*) we have been working with *into the appropriate kind of adjusted relative frequencies* (or *valid percents*). In particular, the existence, direction, and strength of the association between RELIGION and VOTE becomes immediately apparently when we convert Table 1A into the following variant.

TABLE 1E. VOTE BY RELIGION (Column Percentages)

	RELIGION	
PRES. VOTE	Protestant	Catholic
Dem.	46%	86%
Rep.	54%	14%
Total	100% (n = 650)	100% (n = 350)

What we have done here is to replace each absolute frequency with its *column percentage*. For example, the 46% in the Dem-Prot. cell tells us that *300 is 46% of column total of 650* — substantively that *46% of all Protestants vote Democratic*. More generally, each set of column percentages shows relative frequencies with respect to the dependent (row) variable for a given value of the independent (column) variable. If column percentages are about the same across all columns, we infer that the independent variable has little or no apparent influence on, or association with, the dependent variable. If column percentages differ substantially from column to column, we infer that the independent has substantial apparent influence on, or association with, the dependent variable, and the direction and strength of that association is revealed by the nature of the column to column differences.

Especially in a 2×2 table like Table 1E, the apparent influence of the independent variable on the dependent variable, or the association between them, can be summarized by the *percentage difference* between columns — in this case, by saying that *Catholics are 40 percentage points more likely to vote Democratic than Protestants are* (or, equivalently, that *Protestants are 40 percentage points more likely to vote Republican than Catholics are*). (Calculating column percentages for Table 1C shows that Catholics could be *at most* 68 percentage points more likely to vote Democratic than Protestants are; calculating column percentages for Table 1D shows that Protestants could be *at most* 92 percentage points more likely to vote Democratic than Catholics are.)

One potential (and, unfortunately, often actual) source of confusion concerning table percentages is that, given a "two-dimensional" (cross) tabulation, there are two — indeed, actually three — sets of totals on which percentages may be based. Table 1E shows one of these, i.e., column percentages.

Column percentages are based on the total number of (valid) cases in each column. Therefore column percentages add up to 100% in each column. Column percentages answer this question: *of all cases that have a particular value with respect to the column variable, what percent of them have a particular value with respect to the row variable*.

Row percentages are based on the total number of (valid) cases in each row. Therefore row percentages add up to 100% in each row. Row percentage answer this question: *of all cases that have a particular value with respect to the row variable, what percent of them have a particular value with respect to the column variable*.

TABLE 1F. VOTE BY RELIGION (Row Percentages)

RELIGION

PRES. VOTE	Protestant	Catholic	***Total***
Dem.	50%	50%	100% (n = 600)
Rep.	88%	12%	100% (n = 400)

Total percentages are based on the total number of (valid) cases in the whole table, and therefore add up to 100% in the whole table. Table percentages answer this question: *of all cases in the table, what percent of them have a particular combination of values with respect to the row and column variables*.

TABLE 1G. VOTE BY RELIGION (Total Percentages)

	RELIGION		
PRES. VOTE	Protestant	Catholic	***Total***
Dem.	30%	30%	60%
Rep.	35%	5%	40%
Total	65%	35%	100% (n = 1000)

Review Table 1E-1G with these definitions in mind. Normally, a table title does not explicitly say "Column [etc.] Percentages." However, there should be a "total" row at the bottom of the columns and/or a "total" column at the end of each row that shows percentages adding up to 100% (perhaps with rounding error) in one or other directions or overall, thereby making it clear what type of percentages the table is displaying. For reasons discussed below, such a table should also show *the number of cases constituting each 100%* (as each of Tables 1E, 1F, and 1G does).

Most commonly a crosstabulation is constructed to address a question of this type: *what impact (or influence) does (variation in) the independent variable have on the distribution of values with respect to the dependent variable?* (e.g., "what influence does religion have on voting behavior?" or "what impact does ideology have on how people vote?"). By convention, the independent variable is normally made the column variable in a crosstabulation. Thus it is column percentages that answer such questions, and crosstabulations most commonly displays column percentages.

Row percentages answer of this type question: *when cases are categorized with respect to the their values with respect to the row (dependent) variable, how do these categories differ with respect to column (independent) variable accounted for by the independent variable*? (e.g., "how do voting groups differ with respect to religion (or ideology)?").

Table percentages answer basically descriptive (rather than cause and effect) questions about *how the cases in the population as a whole are distributed among the categories defined by all possible combinations of values on the two variables* (e.g., "what percent of all voters are Catholic Democrats or conservative Republicans?").

SPSS Table Percentaging

As you would expect, SPSS crosstabulations can display any or all types of table percentages. In the Crosstabs dialog box, click on Cells and then check the desired percentages. If you wish, you

can suppress the display of (observed) case counts. You can also have SPSS calculate and display "expected case counts" or expected frequencies that would result in the absence of association between the variables (such as are displayed in Table 1B above). Some sample SPSS crosstabulations showing all types are percentages follow.

Suppose we are interested in the influence of IDEOLOGY on PRESIDENTIAL VOTE. Here is the basic SPSS crosstabulation (with some further editing) based on the SETUPS/NES 1992 data.

TABLE 2A. PRESIDENTIAL VOTE * IDEOLOGY CROSSTABULATION
(Case Counts and including Missing Data)

Presidential Vote	Ideology						Total
	1 Lib	2 SL	3 Mod	4 SC	5 Cons	6 NA	
1 Bush	11	26	126	177	206	24	570
2 Clinton	170	174	215	143	47	51	800
3 Perot	24	44	108	90	39	12	317
9 NA	38	83	132	165	50	98	566
Total	243	327	581	575	342	185	2253

If requested to calculate and display (column, row, or total) percentages, SPSS will delete the (shaded) missing data row and column shown above and produce Table 2B displayed on the next page. Note that the total number of cases has been reduced from 2253 to 1600 in the following manner:

2253 total number cases
– 185 missing on IDEOLOGY
– 566 missing on PRESIDENTIAL VOTE
+ 98 missing on both IDEOLOGY and VOTE so double subtracted above
1600 total number of valid cases (= sum of unshaded cells in Table 2A)

The resulting SPSS crosstabulation showing all types of percentages appears on the next page Note that SPSS labels row percentages as "% within Dependent Variable" and column percentages as "% with Independent Variable." Additional SPSS tables of PARTY IDENTIFICATION BY IDEOLOGY and of BUSH JOB APPROVAL BY PARTY IDENTIFICATION (parallel to the Student Survey crosstabulation you did in Problem Set #10) are attached at the end of the handout. The latter was produced by an earlier (DOS-based) version of SPSS, and it has also been reformatted into a "presentation grade" table that very clearly shows the "coloring" of PRESIDENTIAL APPROVAL by PARTY ID, especially in Presidential election year (this also is SETUPS/NES 1992 data).

Presidential Vote * Ideology Crosstabulation

Presidential Vote			Ideology					Total
			Liberal	Slightly Liberal	Moderate	Slightly Conservativ	Conservative	
	Bush	Count	11	26	126	177	206	546
		% within Presidential Vote	2.0%	4.8%	23.1%	32.4%	37.7%	100.0%
		% within Ideology	5.4%	10.7%	28.1%	43.2%	70.5%	34.1%
		% of Total	.7%	1.6%	7.9%	11.1%	12.9%	34.1%
	Clinton	Count	170	174	215	143	47	749
		% within Presidential Vote	22.7%	23.2%	28.7%	19.1%	6.3%	100.0%
		% within Ideology	82.9%	71.3%	47.9%	34.9%	16.1%	46.8%
		% of Total	10.6%	10.9%	13.4%	8.9%	2.9%	46.8%
	Perot	Count	24	44	108	90	39	305
		% within Presidential Vote	7.9%	14.4%	35.4%	29.5%	12.8%	100.0%
		% within Ideology	11.7%	18.0%	24.1%	22.0%	13.4%	19.1%
		% of Total	1.5%	2.8%	6.8%	5.6%	2.4%	19.1%
Total		Count	205	244	449	410	292	1600
		% within Presidential Vote	12.8%	15.3%	28.1%	25.6%	18.3%	100.0%
		% within Ideology	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	12.8%	15.3%	28.1%	25.6%	18.3%	100.0%

Recovering Original Case Counts

Suppose you are given a column percent (only) table such as Table 1E above, but you want (or are asked on a problem set or test) to answer a row percent question such as "Of all Democratic voters, what percent are Protestants?" First note that you cannot the answer this question immediately from the table and, in particular, the answer is *not* 46%. (46% is the answer to the question "Of all Protestants, what percent vote Democratic?")

But, *if and only if the* (column or row percent) *table shows the number of cases* corresponding to each 100% (i.e., each column, row, or table total), you use can use these case counts in conjunction with the percentages displayed in the table to *recover the original case counts* — for example to recover Table 1A from Table 1E. Once you are back to Table 1A, you can of course calculate row percentages and thereby answer any row percent question.

This is one reason why a table displaying percentages should always show the number of cases constituting each 100% of cases shown. In addition, sample size should always be specified. Table 1E would offer less persuasive evidence of the impact of religion on voting if the *n*'s were 65 and 35, rather than 650 and 350.

Confusing Row and Column Percentages

Debates and commentary concerning public affairs are sometimes off the mark because (in effect) row and column percentages have been confused. Here is a salient example.

After the (first) Gulf War, many news reports noted that about 40% of U.S. battle deaths resulted from "friendly fire," as opposed to about 5% in WWII, Korea, and Vietnam. Some commentators drew the inference from this statistic that U.S. military forces had become sloppy or careless. But our baseline expectation of what should be approximately constant from war to war if the competence and discipline of U.S. forces remains approximately constant is not (i) U.S. friendly-fire deaths *as a percent of all U.S. battle deaths suffered* but rather (ii) U.S. friendly-fire deaths *as a percent of all U.S. battle deaths inflicted* . (The 40% and 5% statistics are of the first type.) In a roughly balanced conflict in which each side suffers and inflicts about the same number of deaths, (i) and (ii) are about the same (though of course we typically have more precise information about the number of U.S. battle deaths suffered than about the number of U.S. battle death inflicted). But in a highly unbalanced conflict, such as the Gulf War (or the initial month of the more recent Iraq war), they are quite different. Percent (ii) is a very rough indicator of the competence and discipline of U.S. forces, while percent (i) is essentially an indicator of how unbalanced the conflict is. (After all, if an enemy is disarmed or surrenders before getting off a single shot, U.S deaths will be very low but 100% of them necessarily result from friendly fire — there being no unfriendly fire to inflict any U.S. deaths.)

Dangling Percentages: Row, Column, Total, or What?

I will distribute in class a graphical box that accompanied an article on "Who Lacks Health Insurance?" that appeared in the *Washington Post* weekly health section some years ago. The charts display a variety of percentages but they are all "dangling percentages" — that is, it is nowhere made clear what the base of each percentage is — in effect, whether it is a row, column, or total percentage. Therefore it is also unclear what questions these percentages may be answers to.

The *Washington Post* box focuses on, and appears to refine, the commonly quoted statistic that about 14% of the American population (about 16% of those under 65 -- virtually everyone 65 or older is covered by Medicare) lacks health insurance coverage at any given time. It appears to show how health insurance coverage (the dependent variable) is affected by various (independent) demographic variables -- namely, *age*, *region*, *size of employer*, and *income*. Normally, such relationships would be analyzed by means of a column percent table set up in the following manner.

HEALTH INSURANCE (DEP VAR)	**DEMOGRAPHIC CATEGORIES (INDEPENDENT VARIABLE)**			
	A	*B*	*C*	*etc.*
Covered	(100-A)%	(100-B)%	(100-C)%	. . .
Not Covered	A%	B%	C%	. . .
Total	100%	100%	100%	

Age. For example, with respect to age, we might expect to be shown how coverage rates vary with age. Thus we might initially suppose that the first panel of the box is, in effect, presenting us with the following column percent table:

HEALTH INSURANCE STATUS	**AGE CATEGORY**						
	0-4	5-17	18-24	25-34	35-54	55-59	60-64
Covered	94%	83%	81%	76%	74%	96%	96%
Not Covered	6%	17%	19%	24%	26%	4%	4%
Total	100%	100%	100%	100%	100%	100%	100%

But closer consideration shows that this cannot be correct, for the following reasons.

(i) The table above does not make substantive sense. Why would preschool (age 0-4) children be so much more completely covered than older (5-17) children? Why would people in their prime working years (35-54) have the lowest rate of coverage, when health insurance comes primarily through employment (for those under age 65)? And why would older people (55-64), some of whom have stopped working and lost employer-provided insurance (but are not old enough to be covered by Medicare), be covered at the highest rate?

(ii) We note that 6% +17% +19% +24% +26% +4% +4% = 100%, which *strongly* suggests that the *Post*'s percentages are (in the sense of the table as set up above) *row* percentages, not *column* percentages. (The pie-chart format of the display reinforces this interpretation.) So the chart is actually presenting a row table of the following sort.

HEALTH INSURANCE STATUS	**AGE CATEGORY**							
	0-4	5-17	18-24	25-34	35-54	55-59	60-64	Total
Covered	?%	?%	?%	?%	?%	?%	?%	100%
Not Covered.	6%	17%	19%	24%	26%	4%	4%	100%

Thus the 26% in the 35-54 column is *not* saying that 26% of all people in this age category lack health insurance (column %) but rather that 26% of all people (under age 65) who lack health insurance are in this age category (row %). Now the 26% makes sense because this is by far the "widest" age category (20 years wide) and (by including most of the "baby-boom" generation) may be the most "densely" populated category as well. That is, even though this age group may

have the highest rate of health insurance coverage (for reasons noted above), it includes such a large fraction of the population under age 65 (perhaps upwards of 40%) that people in this age category still constitute about a quarter of those without health insurance. If this interpretation is correct, we can also understand why the "narrowest" age categories consistently are associated with the smallest percentages.

Note that, if this interpretation is correct (and I am essentially sure that it is), the *Post*'s graphic is *not* telling us what we (probably) want to know — i.e., how health insurance coverage varies by age. Moreover, we cannot recover original case counts because we do not know (from the information the Post presents) the size of the *n*'s constituting 100% in each row.

Region. Again with respect to region, we might expect to be shown how coverage rates vary by region, so we might initially suppose that the second panel (map of U.S.) is, in effect, presenting us with the following column percent table:

HEALTH INSURANCE STATUS	**REGION**			
	West	*Midwest*	*South*	*Northeast*
Covered	75%	82%	58%	86%
Not Covered	25%	18%	42%	14%
Total	100%	100%	100%	100%

So interpreted, these percentages make a certain amount of sense. We would expect the South to have the lowest rate of coverage, as it remains the poorest area of the country. Perhaps the mobility of the population in the West leads to a somewhat lower rate of coverage than in the Midwest and Northeast. But there is a basic problem — the average rate of non-coverage across all four (roughly equally populated) regions appears to be about 25%, while at the outset we were told the national non-coverage rate was 16% (14% if people over 65 are included). Moreover, we see that 25% +18% + 42% +14% = 99% (≈ 100% with rounding error). So it appears very likely that these percentages also are actually row percentages.

HEALTH INSURANCE STATUS	**REGION**				
	West	*Midwest*	*South*	*Northeast*	*Total*
Covered	?%	?%	?%	?%	100%
Not Covered	25%	18%	42%	14%	100%

Size of Employer and Income. These do appear to be the expected column percentages such as would appear in the table as set up at the outset (so the *Post*'s graphic is mixing types of percentages without warning). We can conclude this for three reasons: (i) the percentages are all plausible when so interpreted; (ii) it is plausible that they average out to about 16% (when we recognize that probably about half of all workers work for firms with 100 or more workers and that at least half of all families have incomes of more than 200% of the poverty level); and (iii)

the percentages in these panels do not add up to 100% (even approximately).

Sometimes percentages are discussed in even more profoundly confused ways. I will also distribute in class a more recent news story in *Washington Post*, which provides another illustration of the problem of "dangling percentages" in (newspaper reports of) public policy debates.

The opening paragraph, echoing the lead title of the story and citing a Northeastern University report, says that "recent immigrants . . . account[ed] for half of the new wage earners who joined the labor force in [the past decade]." The second paragraph make the even more striking claim that "eight of 10 new male workers in the decade were immigrants who arrived during that time." Put otherwise, the first claim appears to be that 50% of all the people who entered the labor force from 1990 to 2000 were immigrants had who entered the country during the same decade, while the second claim appears to be that 80% of all the males who entered the labor force during the decade were immigrants.

But back-of-the-envelope calculations show that such claims are manifestly absurd. Three to four million babies have been born in the U.S. each year since the end of WWII. Almost all enter the workforce some 25 years later – at least three million a year and at least 30 million over the past decade. If an equal number of new immigrants entered the labor force over the same period (thereby constituting 50% of all new entrants), it must be that considerably more than 30 million immigrants (allowing for children and other non-workers) entered the U.S. during the decade. This greatly exceeds any reasonable estimate of (legal and illegal) immigration during the decade (which the news story itself puts at about 13 million, of whom 8 million joined the labor force). Things get even more bizarre when we make similar calculations based on the 80% figure for males only. (They lead to the conclusion that more than 60 million male immigrants entered the country [and its labor force] over the decade.)

A hint that the data has been misinterpreted in the opening paragraphs appears in the third paragraph, for otherwise it is puzzling why the story, having given percentages for Maryland and Virginia corresponding with those for the nation as a whole, declines to give such a percentage for the District of Columbia, on the grounds that its "workforce declined [but] immigrants prevented further shrinkage." If these percentages really refer to the percent of new workers who are new immigrants, they can be calculated in exactly the same manner regardless of whether the size of the overall workforce has increased, decreased, or remained constant.

Actually, the first sentence in the third paragraph (which the news story seems to treat as parallel to the sentences quoted above), as well as the headline for the continuation of the story and heading for the table that appears below the pie chart, make it reasonably clear what these percentages actually refer to — namely, that they result from dividing the number of new immigrants who entered the labor force during the decade, not by the total number of labor market entrants over the same period, but by the *net absolute growth* (if any) *in the labor force over the period*. Given this interpretation, the percentages are plausible (and unsurprising), but they really don't convey much information of interest, since the numerator and denominator are

proverbial "apples and oranges." Note that, while the magnitude of the resulting percentage depends in part on (and is a positive function of) the magnitude of the numerator, it is much more sensitive to (and is a negative function of) the magnitude of the denominator. In fact, the numerator can easily exceed the denominator, so the resulting percent can easily exceed 100% (though this threshold has no distinctive importance) — indeed, the percent is infinite if net growth is zero (and incoherent if net growth is negative, as in the case of D.C). And whatever percent of net growth new immigrants may constitute, new native-born workers constitute a much larger percentage of net growth (i.e., about 200%, based on the back-of-envelope calculations above).

Having seemingly corrected itself in the third paragraph, in the last sentence of the fifth paragraph, the story reverts to the original manifestly incorrect interpretation of the percentages. (It should say that "the report said 8 million immigrants joined the labor force . . . over a period when the total number of new workers *exceeded the total number of departing workers* by 16 million" [or the report itself should have said this].)

The "theme" of the article, and apparently of the research report on which it is based, is evidently that immigrants (and recent immigrants in particular) make up an increasingly large proportion of the U.S. labor force. This information is of interest (though not of surprise). The pie chart displays the relevant profile of the U.S. labor force in 2001: native-born Americans constituted 86% of the labor force and all immigrants the remaining 14%; and the subset of immigrants who entered the country since 1990 constituted 5.7% of the labor force. One wonders why such percentages could not have been calculated for successive decades and the trend from decade to decade straightforwardly presented.

Later the news story refers to "a puzzling decline in the share of U.S.-born men in the workforce." We can wonder whether this unspecified "share" refers to (a) U.S.-born men in the workforce as a percent of all U.S.-born men or (b) U.S.-born men in the workforce as percent of all members (or all U.S.-born members) of the workforce. The suggested explanations (e.g., earlier retirement) pertain to (a), but, given the earlier confusions, I wonder whether the story may actually be referring to (b), which surely has declined quite dramatically over the past generation (as more women have entered the labor force). If so, this would be the same type of "row percent" vs. "column percent" confusion as in the friendly fire statistics discussed earlier.

Answering Questions from Crosstabulations

Questions pertaining to crosstabulations are all of this general form: "Of all cases for which A is true, for what fraction (or percent) of cases is B also true?", where "A" and "B" refer to values of the variables in the table (though A may refer to all values and thus be true of all cases).

The remainder of this handout suggests a step-by-step procedure to answer any such question. It assumes that you are starting with a crosstabulation displaying *absolute frequencies* (or total percentages). Thus, if you are given a table displaying row or column percentages, you

must first *recover* the absolute frequencies in each cell of the table, in the manner discussed above. Do this by multiplying the number of cases corresponding to each 100% (row or column total) by the relative frequency (percent) in that cell. (A properly constructed table displays such cases counts; otherwise you cannot recover the absolute frequencies and you can answer only row percent or column percent questions, according to the type of percentages displayed.) You now have the "core" of the crosstabulation, i.e., the absolute frequencies of the interior cells of the table that represent particular combinations of values on the two variables, and from which any type of percentage question can be answered.

(1) *First*, put a *double line* (or other distinctive marking) around all the cells of the table for which A is true.

   (a) If A refers to all cases in the table, the double line goes around the entire table.

   (b) If A refers to all cases with a specified value (or set of values) on the column variable, the double line goes around the appropriate column (or set of columns).

   (c) If A refers to all cases with a specified value (or set of values) on the row variable, the double line goes around the appropriate row (or set of rows).

   (d) If A refers to all cases with specified combinations of values on the row and column variables, the double line goes around the appropriate cells.

(2) *Second*, *shade in* (or otherwise indicate) all the cells of the table (1) which are within the double lines *and* (2) for which B is true (where B, like A, refers to one or more rows, columns, or cells in the table).

(3) *Finally*, the answer to the question is simply the *fraction* formed by dividing the number of cases (or the sum of total percentages) in the shaded cells by the number of cases (or the sum of total percentages) in the portion of the table enclosed by double lines. This fraction can be straightforwardly converted into a percentage by using a calculator (or even paper and pencil). Many tables, including many SPSS tables, include (a) row and column totals and/or (b) row and/or column and/or total percentages for each cell, which may save you from making calculations. However *the "core" of the table* from which everything else can be calculated is the set of case counts (absolute frequencies) in each cell of the table.

Here are some examples that have been worked out in this manner. Consider the following SPSS crosstabulation of PARTY IDENTIFICATION by IDEOLOGY in the 1992 NES data. PARTY IDENTIFICATION has been recoded so that it reflects answers given to the basic Party ID question only (i.e., Question 1 on the Student Survey). Likewise, IDEOLOGY has been recoded to combine the "liberal" and "slightly liberal" categories and likewise the "conservative" and "slightly conservative" categories. Thus we have the 3 × 3 table that appears on the following page.

Crosstabulation: V8 Party Identification
By V53 Ideology

V8	V53→	Liberal 1	Moderate 2	Conservative 3	Row Total
Democrat	1	304 42.7 53.5 14.8	199 27.9 34.5 9.7	209 29.4 22.9 10.2	712 34.6
Independent	2	211 26.8 37.1 10.2	241 30.7 41.9 11.7	334 42.5 36.7 16.2	786 38.2
Republican	3	54 9.6 9.4 2.6	136 24.4 23.7 6.6	369 66.0 40.4 17.9	559 27.2
	Column Total	569 27.6	577 28.0	912 44.3	2057 100.0

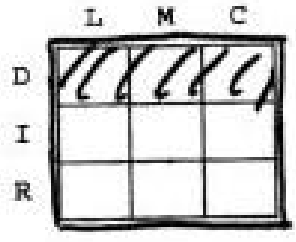
Here is the "core" of the crosstabulation. All totals and percentages can be calculated from this "core" information. =>

	L	M	C
D	304	199	209
I	211	241	334
R	54	136	369

Now let us answer some typical questions pertaining to this crosstabulation.

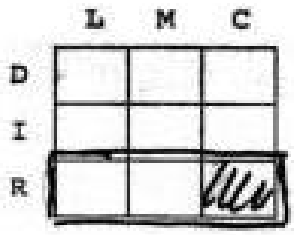
1. Of all respondents, what fraction (or percent) are Democrats?

$\frac{712}{2057} = 34.6\%$ (T%)



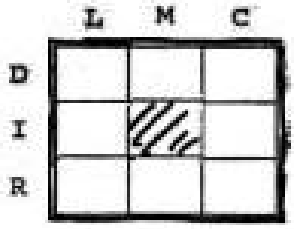
2. Of all Republicans, what fraction (or percent) are conservatives?

$\frac{369}{559} = 66.0\%$ (R%)



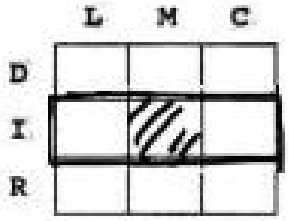
3. Of all respondents, what fraction (or percent) are Independent moderates?

$\frac{241}{2057} = 11.7\%$ (T%)



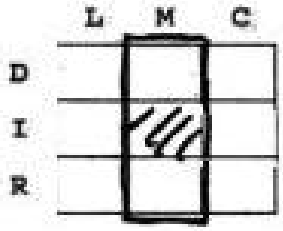
4. Of all Independents, what fraction (or percent) are moderates?

$\frac{241}{786} = 30.7\%$ (R%)



5. Of all moderates, what fraction (or percent) are Independents?

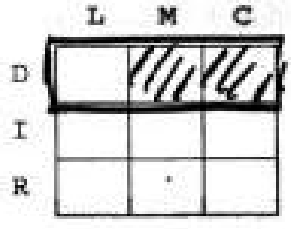
$\frac{241}{517} = 41.9\%$ (C%)



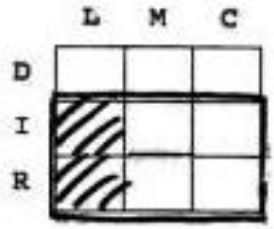
6. Of all Democrats, what fraction (or percent) are non-liberals (i.e., either moderates or conservatives)?

$\frac{199+209}{712} = \frac{408}{712}$

$= 27.9\% + 29.9\% = 57.8\%$ (R%+R%)



7. Of all non-Democrats (i.e., Independents and Republicans), what fraction (or percent) are liberals?

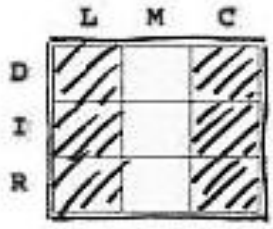


$\frac{211+54}{786+559} = \frac{265}{1345} = \frac{10.2\%+2.6\%}{38.2\%+27.7\%} = \frac{12.8\%}{65.9\%}$

$= 19.6\%$ (R% in combined rows)

Let us call respondents *ideologues* if they are *non-moderate*, i.e., either liberals or conservatives. Let us call respondents *partisans* if they are *non-Independents*, i.e., either Democrats or Republicans. And let us call respondents *consistent partisan ideologues* if they are either liberal Democrats or conservative Republicans.

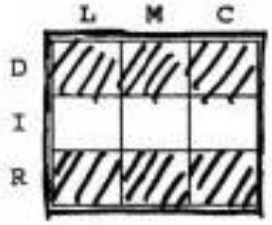
8. Of all respondents, what fraction (or percent) are ideologues?



$\frac{569+912}{2057} = \frac{1481}{2057} = 27.6\%+44.3\%$

$= 79.9\%$ (T%)

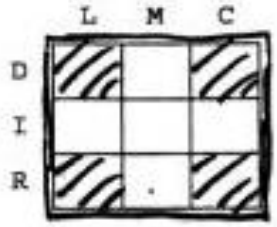
9. Of all respondents, what fraction (or percent) are partisans?



$\frac{712+559}{2057} = \frac{1271}{2057} = 34.6\%+27.2\%$

$= 61.8\%$ (T%)

10. Of all respondents, what fraction (or percent) are partisan ideologues?



$\frac{304+209+54+369}{2057} = \frac{936}{2057}$

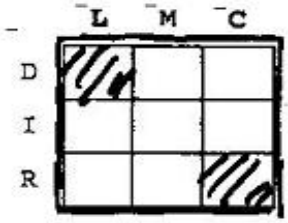
$= 14.8\%+10.2\%+2.6\%+17.9\%$

$= 45.5\%$ (T%)

11. Of all respondents, what fraction (or percent) are consistent partisan ideologues?

$\frac{304+369}{2057} = \frac{673}{2057} = 14.8\% + 17.9\%$

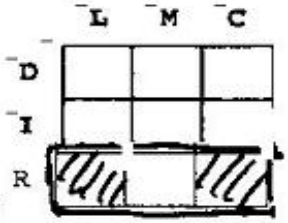
$= 32.7\%$ (T%)



12. Of all Republicans, what fraction (or percent) are ideologues?

$\frac{54+369}{559} = \frac{423}{559} = 9.6\% + 66.0\%$

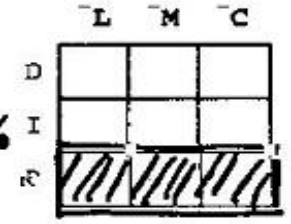
$= 75.6\%$ (R% + R%)



13. Of all Republicans, what fraction (or percent) are partisans?

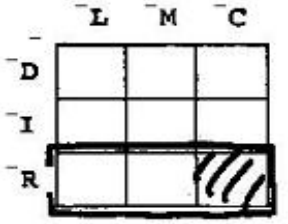
$\frac{54+136+369}{559} = \frac{559}{559} = 9.6\% + 24.6\% + 66.0\%$

$= 100\%$ (R% + R% + R%)



Of all Republicans, what fraction (or percent) are consistent partisan ideologues?

$\frac{369}{559} = 66.0\%$ (~Q2)

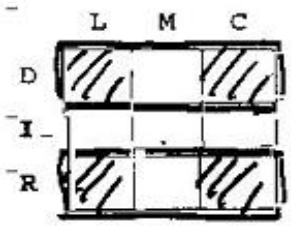


15. Of all partisans, what fraction (or percent) are ideologues?

$\frac{304+209+54+369}{712+559} = \frac{936}{1271}$

$= \frac{14.8\% + 10.2\% + 2.6\% + 17.9\%}{34.6\% + 27.2\%}$

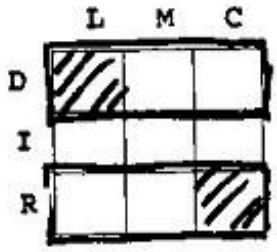
$= \frac{45.5\%}{61.8\%} = 73.6\%$ (R% + R% in combined rows)



16. Of all partisans, what fraction (or percent) are consistent partisan ideologues.

$$\frac{304+369}{712+559} = \frac{673}{1271} = \frac{14.8\%+17.9\%}{34.6\%+27.2\%}$$

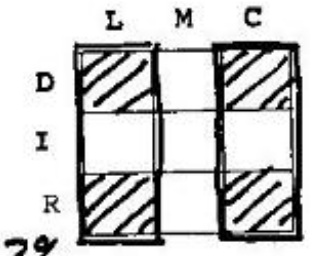
$$= \frac{32.7\%}{61.8\%} = 52.9\%$$



17. Of all ideologues, what fraction (or percent) are partisans?

$$\frac{304+209+54+369}{569+912} = \frac{936}{1481}$$

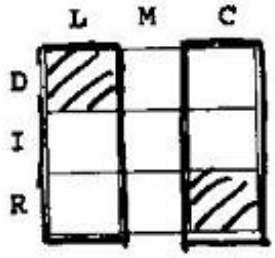
$$= \frac{14.8\%+10.2\%+2.6\%+17.9\%}{27.6\%+44.3\%} = \frac{45.5\%}{71.9\%} = 63.3\%$$



18. Of all ideologues, what fraction (or percent) are consistent partisan ideologues?

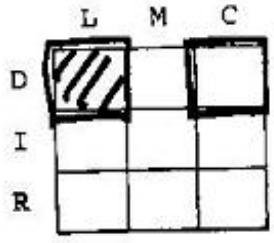
$$\frac{304+369}{569+912} = \frac{673}{1481} = \frac{14.8\%+17.9\%}{27.6\%+44.3\%}$$

$$= \frac{32.7\%}{71.9\%} = 45.5\%$$



19. Of all Democratic ideologues, what fraction (or percent) are consistent?

$$\frac{304}{304+209} = \frac{304}{513} = 59.3\%$$



20. Of all consistent partisan ideologues, what fraction (or percent) are Democrats?

$$\frac{304}{304+369} = \frac{304}{673} = 45.2\%$$