22. Cars. A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

Origin	Driver: Student	Driver: Staff
American	107	105
European	33	12
Asian	55	47

a) What percent of all the cars surveyed were foreign?
b) What percent of the American cars were owned by students?
c) What percent of the students owned American cars?
d) What is the marginal distribution of origin?
e) What are the conditional distributions of origin by driver classification?
f) Do you think that origin of the car is independent of the type of driver? Explain.

This is a 2 way table, where we are asked to find relationship between the 2 categorical variables:

ORIGIN - "ethnicity" of auto in a college parking lot

DRIVER - "status" (classification) of the auto's driver

possible motivating question (purpose) of the study: "Can you tell which cars are driven by staff (or students) just by the car's country of origin?

In this context, DRIVER is the response variable & ORIGIN is the explanatory variable.

traditionally, the response variable is the column variable, with explanatory variable being the row variable -- this "tradition" is frequently broken, however

(a) % of cars surveyed which were foreign

$$= \frac{\#\text{ of foreign cars}}{\text{Total \# of cars in lot}} = \frac{33+12+55+47}{107+105+33+12+55+47}$$

$$= \frac{147}{359} = 0.4095$$

So .4095 ≈ 41% of cars in lot are foreign.

(b) % of American cars owned by students

$$= \frac{\#\text{ of student American cars}}{\#\text{ of American cars}} = \frac{107}{107+105}$$

= .5047 ≈ 50.5% of the American cars are driven by students

(c) % of students who own American cars

$$= \frac{\#\text{ student American cars}}{\#\text{ student cars}} = \frac{107}{107+33+55}$$

= 0.549 ≈ 55% of student cars are American

(d) In order to answer questions about marginal & conditional distributions, we must first add total column, total % column, total row, total % row, & total-total column (with total-total % = 100%). So table rewritten is

DRIVER / ORIGIN	Student	Staff	Total	Total %
American	107	105	212	59.05%
European	33	12	45	12.53%
Asian	55	47	102	28.42%
Total	195	164	359	
Total %	54.32%	45.68%		100%

NOTE: On percentages, hundredths place has been slightly altered to total up to 100% for rows & columns

The row total number = column total number = 359 ← this is how you can "sanity check" your addition

marginal distribution of origin (column marginal):

= 59.05% → American
12.53% → European
28.42% → Asian

note that marginal distribution ALWAYS has the total-total as denominator of fraction used in percentage

(e) conditional distribution of origin by driver classification (column conditional)

Student conditional — Staff Conditional

$\frac{107}{195} = .5487$ ← American → $\frac{105}{164} = .6402$

$\frac{33}{195} = .1692$ ← European → $\frac{12}{164} = .0732$

$\frac{55}{195} = .2821$ ← Asian → $\frac{47}{164} = .2866$

So, conditional of origin by driver

Student		Staff
≈ 54.9%	American	64.0%
17.0%	European	7.3%
28.2%	Asian	28.7%

Note: The denominator of the fraction for conditional distributions is NEVER the total-total, but rather is the column (or row) total -- in this case the column total

(F) We can determine independence relationships by comparing conditional distributions, & how close the conditionals match the marginal distributions, & then verify our conclusions by producing segmented bar graphs of the

conditional distributions. If the conditional bars look "similar", then the marginal bar will also be similar to the similar conditional bars, & we will conclude that the 2 categorical variables are independent.
This corresponds (visually) to saying that the conditional distributions are quite similar to each other, & to the marginal distribution, therefore, the 2 variables are independent.

If I have 2 variables which are "independent", then if you ask me the % of 1 level of a variable, I can answer your question (rather "independently) by just looking at the marginal distribution of that variable.

If I have 2 variables which are "not independent", then if you ask me the % of 1 level of a variable, I cannot answer your question (thereby being not independent) without you first telling me which level of that 2nd variable you are interested in.