Statistical Science 1024

Chapter 6

Two-Way Tables

	L	R	
M	34	36	70
F	56	39	95
	90	75	165

Example:
U.S. Survey of Youth in Custody (1987)

Two variables: **Race** and **Type of crime**

Race

- 1 = white
- 2 = black
- 3 = other (aboriginal, Asian, etc)

Type of crime

- 1 = violent (murder, rape, robbery, assault)
- 2 = property (burglary, larceny, arson, fraud)
- 3 = drug (possession or trafficking)
- 4 = public order (weapons violation, perjury, failure to appear in court)
- 5 = juvenile offence (truancy, running away, incorrigible behavior)

Cross Tabulation

- 1939 total cases (a subsample of a bigger survey file) classified two ways
- the cross tabulation shows the number that fall into each category defined by race ***and*** each category defined by type of crime

distribution of types of crimes

incarcerated

	Violent	Property	Drugs	Public Order	Juvenile Offence	Total
White	373	492	52	64	20	1001
Black	427	294	71	24	9	825
Other	56	47	5	4	1	113
Total	856	833	128	92	30	1939

MARGINAL AND CONDITIONAL DISTRIBUTIONS

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

A **conditional distribution** of a variable is the distribution of values of that variable among only individuals who have a given value of the other variable. There is a separate conditional distribution for each value of the other variable.

Marginal distributions

We can look at each categorical variable in a two-way table separately by studying the row totals and the column totals. They represent the **marginal distributions,** expressed in counts or percentages (they are written as if in a margin).

The marginal distributions can then be displayed on separate bar graphs, typically expressed as percents instead of raw counts. Each graph represents only one of the two variables, completely ignoring the second one.

Marginal Distributions

Race (row variable)

- calculate the counts for each type of race
- in the given table these are the row totals
- calculate proportions or percentages

Crime type (column variable)

- calculate the counts for each crime type
- in the given table these are the column totals
- calculate proportions or percentages

Marginal Distributions from the Survey

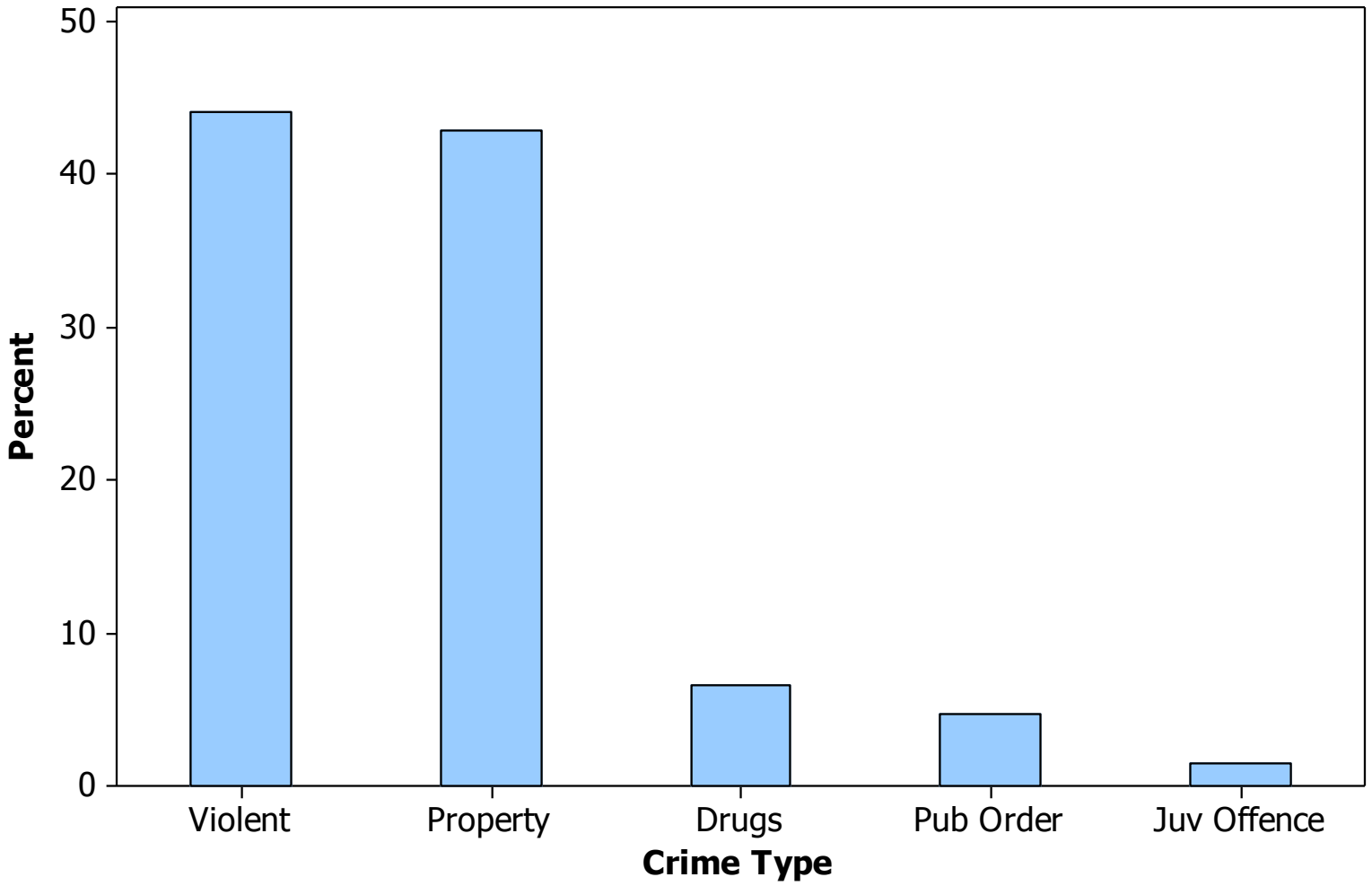
- race

	Number	Percent
White	1001	51.6
Black	825	42.5
Other	113	5.8
Total	1939	99.9

- crime type

	Violent	Property	Drugs	Public Order	Juvenile Offence	Total
Number	856	833	128	92	30	1939
Percent	44.1	43.0	6.6	4.7	1.5	99.9

Visual Display of the Marginal Distribution for Crime Type



Relationships between categorical variables

The **marginal distributions** summarize each categorical variable independently. But the two-way table actually describes the relationship between both categorical variables.

The cells of a two-way table represent the intersection of a given level of one categorical factor with a given level of the other categorical factor.

Because counts can be misleading (for instance, one level of one factor might be much less represented than the other levels), we prefer to calculate percents or proportions for the corresponding cells. These make up the **conditional distributions.**

The counts or percents within the table represent the **conditional distributions.** Comparing the conditional distributions allows us to describe the “relationship" between both categorical variables.

Conditional Distributions on Race Categories

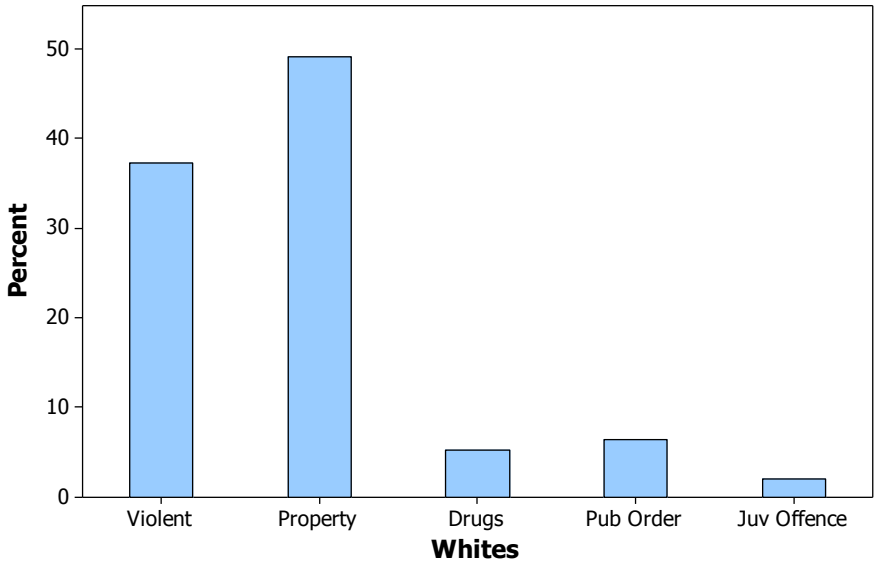
- whites

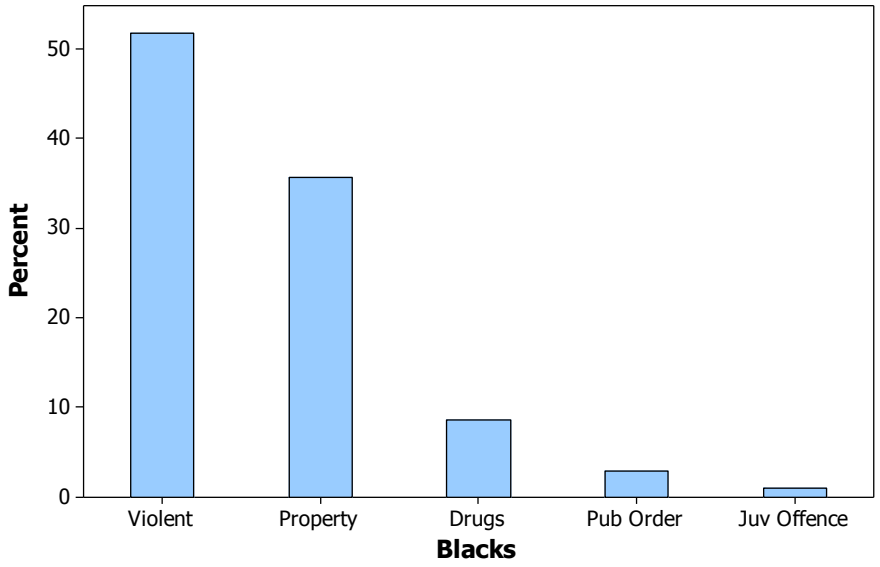
	Violent	Property	Drugs	Public Order	Juvenile Offence	Total
Number	373	492	52	64	20	1001
Percent	37.3	49.2	5.2	6.4	2.0	100.1

- blacks

	Violent	Property	Drugs	Public Order	Juvenile Offence	Total
Number	427	294	71	24	9	825
Percent	51.8	35.6	8.6	2.9	1.1	100.0

Visually Comparing Two Conditional Distributions for Crime Types

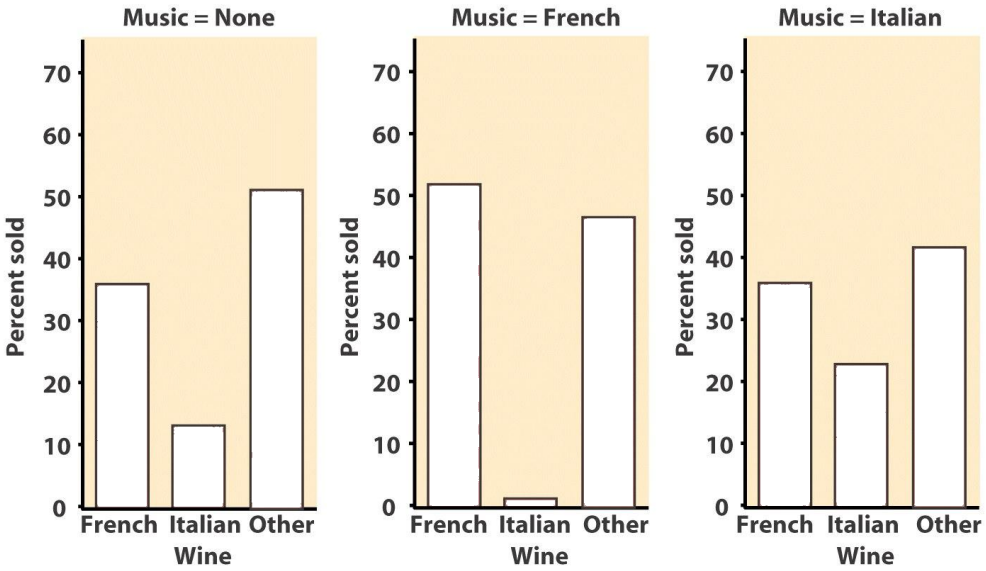




For every two-way table, there are two sets of possible conditional distributions.

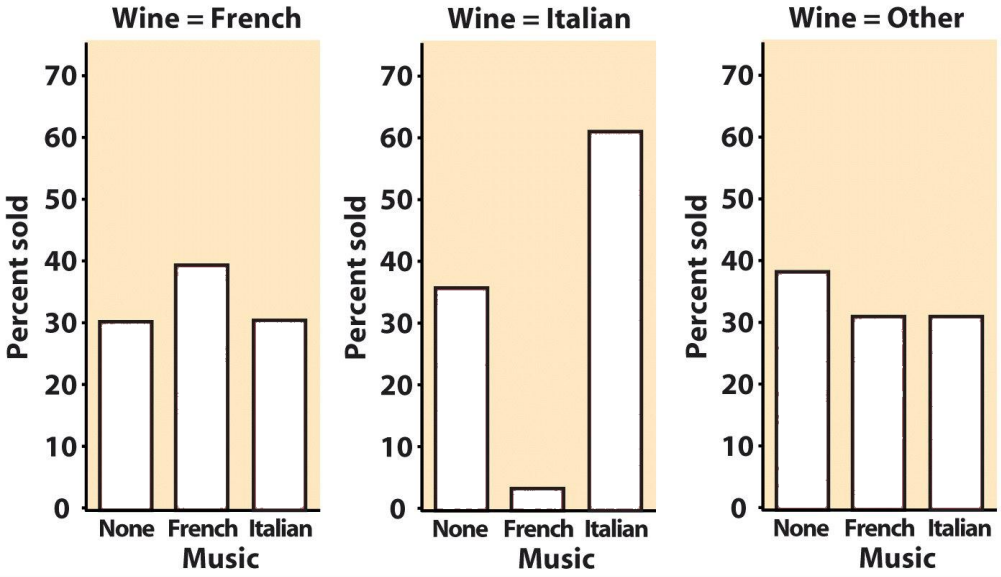
	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Does background music in supermarkets influence customer purchasing decisions?



Wine purchased for each kind of music played (column percents)

Music played for each kind of wine purchased (raw percents)



RMS Titanic: 1912

	First Class		Second Class		Third Class	
	death D	survivor S	D	S	D	S
Men	118	57	154	14	387	75
Women	4	140	13	80	89	76
Children	1	5	0	24	52	27

1316 passengers on the Titanic classified by class of accommodation, gender and whether or not they survived

Source: Board of Trade report on the disaster

Survival of the Richest

Marginal Distribution – sum over gender

	Deaths	Survivors
First Class	123	202
Second Class	167	118
Third Class	528	178

Women and Children First?

Marginal Distribution – sum over accommodation type

	Deaths	Survivors
Men	659	146
Women	106	296
Children	53	56

Children not first?

Focus on the Children

	Deaths	Survivors
First Class	1	5
Second Class	0	24
Third Class	52	27

Poorer children – prelude to Simpson’s Paradox

Survival Percentages by Gender and Accommodation

	First Class	Second Class	Third Class
Men	32.6	8.3	19.3
Women	87.2	81.7	46.1
Children	83.3	100.0	34.2

Note: to do any manipulations of the table we need to deal with counts

SIMPSON'S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox.**

Simpson's paradox

Beware of lurking variables

Example: Hospital death rates

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800
% Surv.	97.0%	98.0%

On the surface, Hospital B would seem to have a better record.

But once patient condition is taken into account, we see that, in fact, Hospital A has a better record for both patient conditions (good and poor).

Patients in good condition

	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600
% surv.	99.0%	98.7%

Patients in poor condition

	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200
% surv.	96.2%	96.0%

Here patient condition was the lurking variable.

Another Example: Smoking and Mortality

- survey of 839 women between the ages 35 and 74 in Whickham, United Kingdom
- survey carried out over the years 1972 to 1974
- each woman was followed up after 20 years

Cross Tabulation for the Sample

- Counts

	Dead	Alive	Total
Smoker	121	269	390
Nonsmoker	160	289	449

- Distributions conditional on smoking status

	Dead	Alive	Total
Smoker	31.0	69.0	100.0
Nonsmoker	35.6	64.4	100.0

- Does it really pay to smoke?

Age: the Lurking Variable

	Age Group							
	35 – 44		45 – 54		55 – 64		65 – 74	
	Smoker?		Smoker?		Smoker?		Smoker?	
	Yes	No	Yes	No	Yes	No	Yes	No
Dead	14	7	27	12	51	40	29	101
Alive	95	114	103	66	64	81	7	28
% deaths	12.8	5.8	20.8	15.4	44.3	33.0	80.6	78.3

- The death rate for smokers is greater than that of nonsmokers in every age group