Section 5.6
Relations in Categorical Data

A **two-way table** organizes the data for two categorical variables.

The totals of each row and column are considered **marginal distributions** because they appear in the margins of the table.

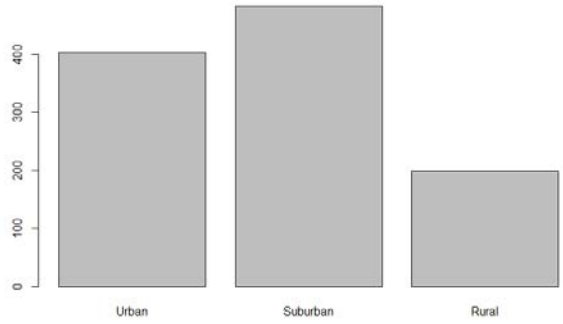
Example: The following two-way table describes the age groups and localities of residents for people in and around a certain small town

	Urban	Suburban	Rural	Totals
Under 25	110	150	65	325
25-50	240	220	75	535
Over 50	53	112	58	223
Totals	403	482	198	1083

a. Construct the marginal distributions of this table in counts.

b. Draw a bar chart to display the marginal distribution of locality. *From Section 1.5.*
Commands:

```
x = c(403, 482, 198)
barplot(x, names.arg = c("Urban", "Suburban", "Rural"))
```



c. What percent of urban dwellers are over 50?

$\frac{53}{403} = 0.1315 \Rightarrow 13.15\%$

d. What percent of the residents are under 25 and live in urban areas?

Intersection!

$\frac{110}{1083} = 0.1016 \Rightarrow 10.16\%$

A **conditional distribution** is made up of the percentages that satisfy a given condition.

e. Compare the conditional distributions of locality for the Under 25 and Over 50 age groups. Use percentages.

	Urban	Suburban	Rural
Under 25 (total) 325	$\frac{110}{325} = 33.85\%$	$\frac{150}{325} = 46.15\%$	$\frac{65}{325} = 20\%$

	Urban	Suburban	Rural
Over 50 (total) 223	$\frac{53}{223} = 23.77\%$	$\frac{112}{223} = 50.22\%$	$\frac{58}{223} = 26.01\%$

Here's an example for observation:

A drug company tests two new treatments for an illness. Here is a table for each trial. Which treatment would you conclude is better based on these tables?

Trial 1	Cured	Total	Percentage
Drug A	45	200	22.5%
Drug B	32	200	16%

A

Trial 2	Cured	Total	Percentage
Drug A	85	100	85%
Drug B	400	500	80%

A

Next, the data is put together into one table and the percentage of cured for the aggregated data is shown. Now which treatment would you conclude is better based on this table?

Combined Trials 1 and 2	Cured	Total	Percentage
Drug A	130	300	43.3%
Drug B	432	700	61.7%

B

As seen in the example above, one should always use caution when combining data to form a single group. **Simpson's Paradox** is the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.