ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

SUMMARIZING CATEGORICAL DATA

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

We will:

- Visualize categorical data using bar plots and their variations
- Examine categorical data using tables
- Compare numerical data across groups

Contingency Tables for Two Categorical Variables

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Table from Open-intro Statistics textbook, Chapter 2.

- The table above is called a **contingency table.**
- It summarizes data for two categorical variables, in this case: homeownership type and loan application type.

Row and Column Proportions

- We can modify contingency tables to show the fractional breakdown of one variable in another.

	homeownership			
	rent	mortgage	own	Total
individual	3496	3839	1170	8505
joint	362	950	183	1495
Total	3858	4789	1353	10000

- New table shows the **row proportions** for the homeownership/loan data, computed as the counts divided by their row totals.
  - Example: 3496 at the intersection of individual and rent is replaced by 3496/8505 = 0.411 corresponding to the proportion of individual applicants who rent.

	rent	mortgage	own	Total
individual	0.411	0.451	0.138	1.000
joint	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

- A contingency table of the **column proportions** is similar: each column proportion is the count divided by the corresponding column total.

	rent	mortgage	own	Total
individual	0.906	0.802	0.865	0.851
joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

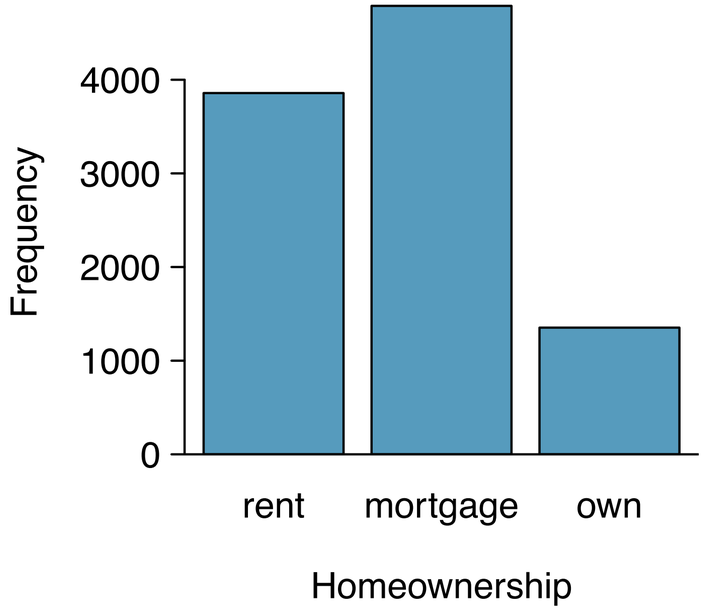
Tables from Open-intro Statistics textbook, Chapter 2.

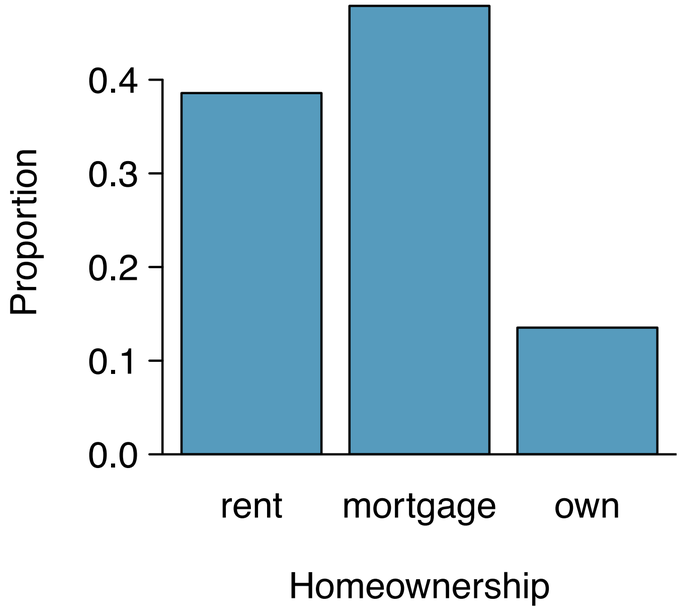
Tables and Bar Plots

- The table on this slide shows the overall numbers for a single category.
- A bar plot is another way to display a single categorical variable.

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

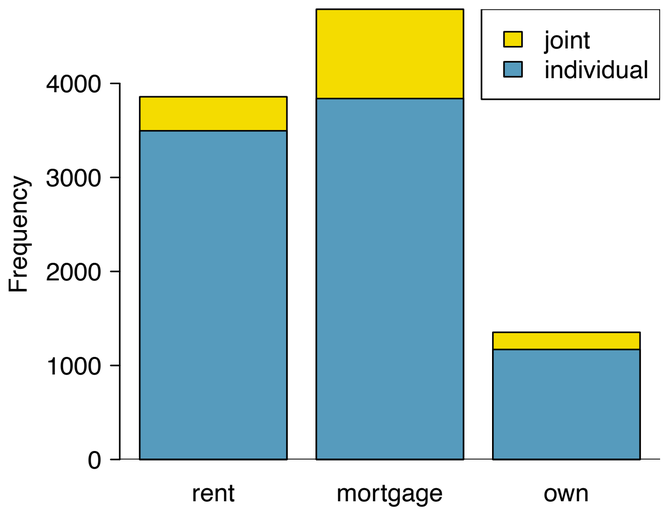
Table and figures from Open-intro Statistics textbook, Chapter 2.



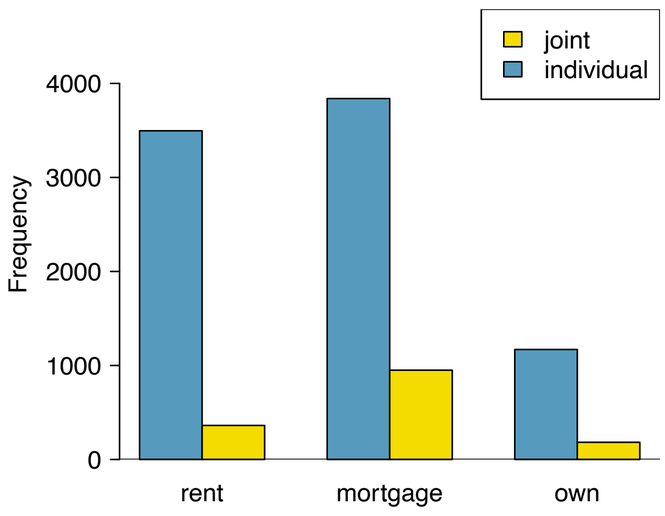


Bar Plots with Two Variables

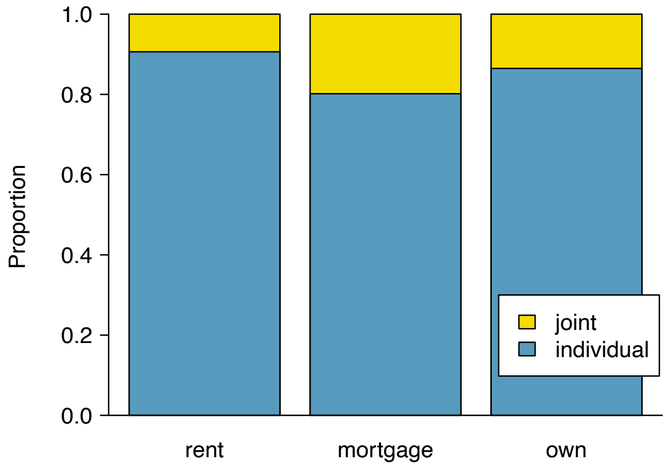
- A **stacked bar plot** (Figure a) is a way to visualize contingency table information.
- A **side-by-side bar plot** (Figure b) is similar.
- A **standardized stacked bar plot** can help visualize, e.g., column proportions (Figure c).



(a)



(b)



(c)

Comparing Numerical Data across Groups

- To examine numerical data across groups, one can simply make a numerical plot for each group and plot them on the same graph.

- For example, we can do
  - Side by side box plots
  - Hollow histograms

Median Income for 150 Counties, in $1000s

Population Gain						No Population Gain		
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
44.6	51.8	40.7	48.1	56.4	41.9	39.3	40.4	40.3
40.6	63.3	52.1	60.3	49.8	51.7	57	47.2	45.9
51.1	34.1	45.5	52.8	49.1	51	42.3	41.5	46.1
80.8	46.3	82.2	43.6	39.7	49.4	44.9	51.7	46.4
75.2	40.6	46.3	62.4	44.1	51.3	29.1	51.8	50.5
51.9	34.7	54	42.9	52.2	45.1	27	30.9	34.9
61	51.4	56.5	62	46	46.4	40.7	51.8	61.1
53.8	57.6	69.2	48.4	40.5	48.6	43.4	34.7	45.7
53.1	54.6	55	46.4	39.9	56.7	33.1	21	37
63	49.1	57.2	44.1	50	38.9	52	31.9	45.7
46.6	46.5	38.9	50.9	56	34.6	56.3	38.7	45.7
74.2	63	49.6	53.7	77.5	60	56.2	43	21.7
63.2	47.6	55.9	39.1	57.8	42.6	44.5	34.5	48.9
50.4	49	45.6	39	38.8	37.1	50.9	42.1	43.2
57.2	44.7	71.7	35.3	100.2		35.4	41.3	33.6
42.6	55.5	38.6	52.7	63		43.4	56.5	

Table from Open-intro Statistics textbook, Chapter 2.

Comparing Numerical Data across Groups

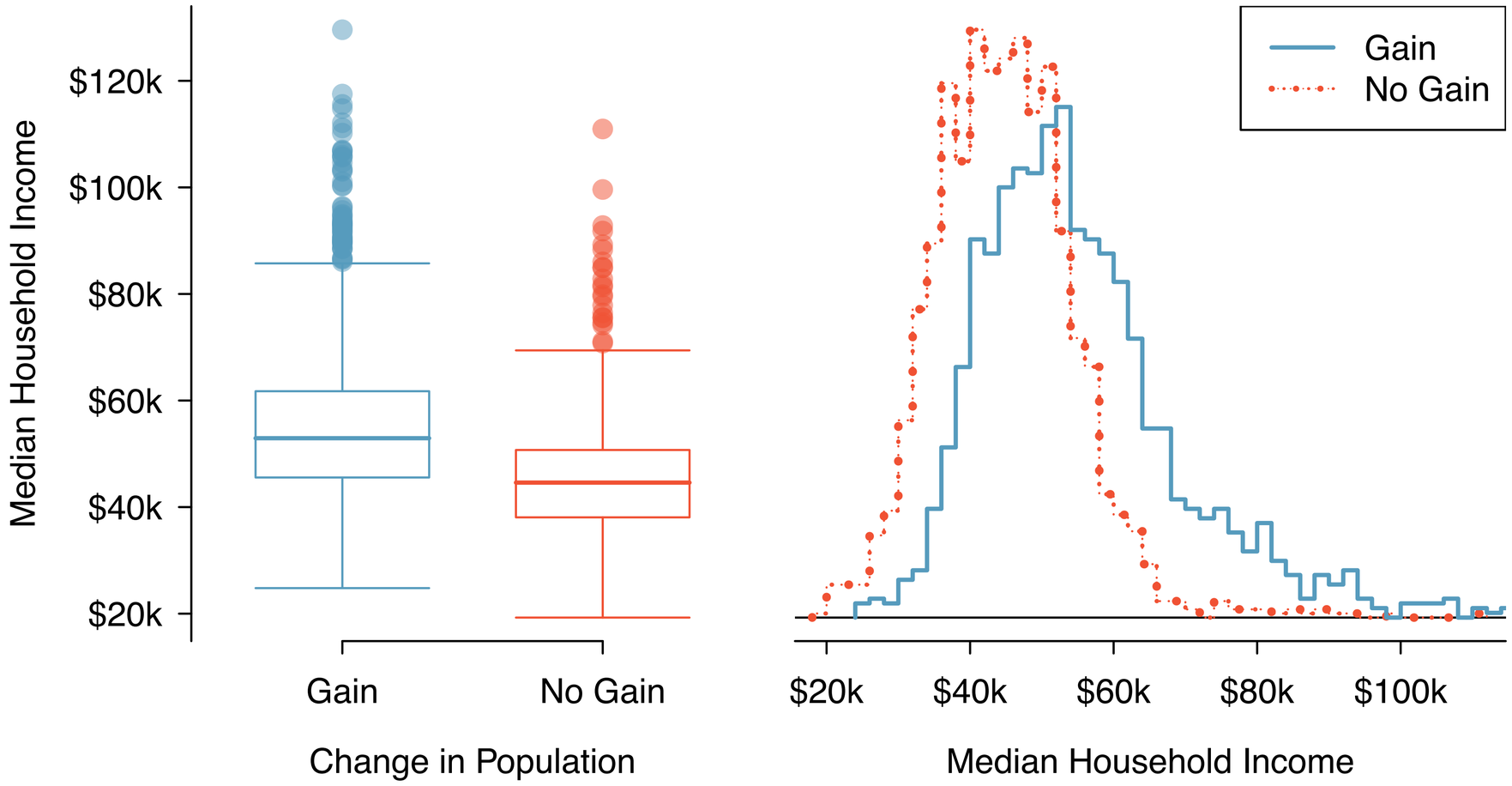


Figure from Open-intro Statistics textbook, Chapter 2.