Topic 5: Examining Categorical Data

Contingency Tables

We can use a **contingency table** to break down data with respect to 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

Contingency table: contains aggregate data for 2 categorical variables, combined

Question of interest: Does there appear to be a relationship between loan application type and homeownership status? We can start to answer this question by examining row proportions:

- % of all applicants who rent:

total proportion who rent

$\frac{3858}{10000} = 0.3858 = 38.58\%$

- % of individual applicants who rent:

$\frac{3496}{8505} \approx 0.41$ or 41%

Numerator: rent AND individual
Denominator: all individual

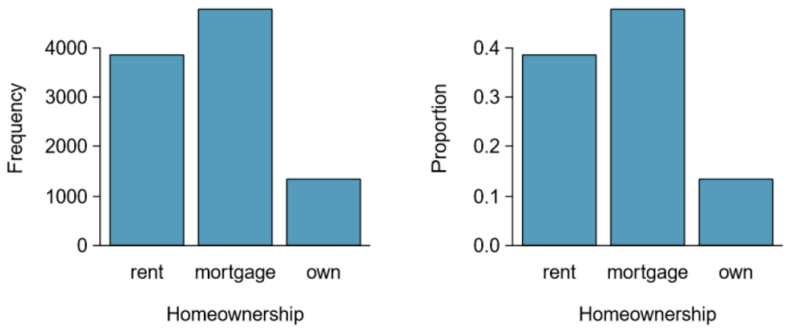
- % of joint applicants who rent:

$\frac{362}{1495} \approx 0.24$ or 24%

Numerator: rent AND joint
Denominator: all joint

(You can do this same thing with column proportions, proportions using multiple cells... any combination that meaningfully contributes to your research question!

Bar Plots



Bar plot: displays frequencies (or proportions, i.e. relative frequencies) for all categories of 1 categorical variable

Differences between bar plots and histograms?

Histogram: numberline on x-axis, you get to choose suitable bins

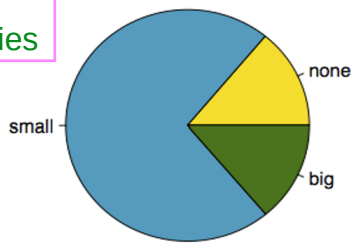
Bar plot: categories on x-axis, fixed by categories for the variable

Standard convention: histogram bins are directly adjacent to each other according to numberline arrangement

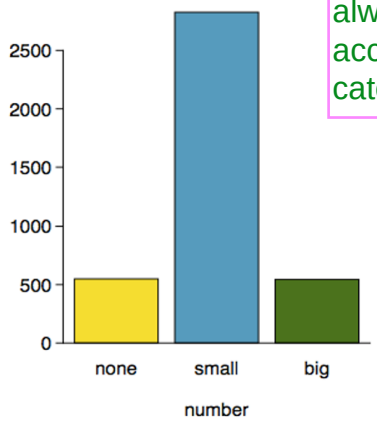
Standard convention: include gaps between bars on a bar plot, to separate categories

Pie Charts

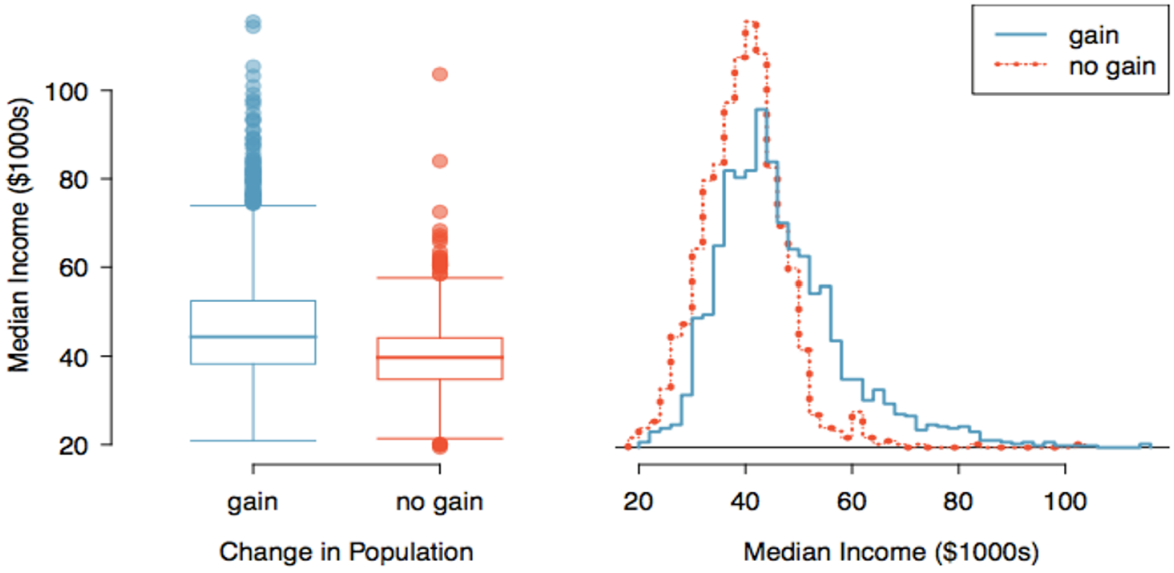
Visually distinct, but make it hard to see RELATIVE sizes of categories



Bar plot is basically always better for accurately displaying categorical data!



Side-by-Side Plots



Can separate out data from a numerical variable according to categories from a categorical variable!

Then make numerical plots (e.g. box plot, histogram) for each category, but put them on the same figure

Allows side-by-side viewing of the influence of the categorical variable! This is using a categorical variable as EXPLANATORY for some numerical RESPONSE variable