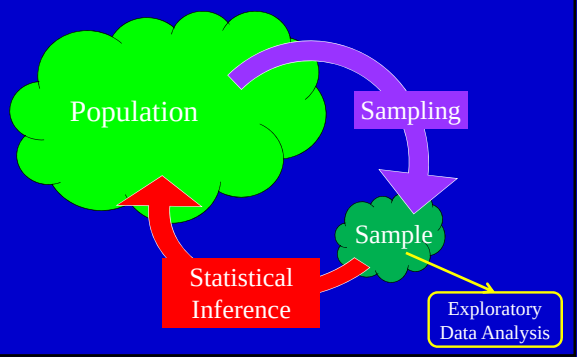
Describing Data:
Categorical and Quantitative Variables

The Big Picture



Descriptive Statistics

In order to make sense of data, we need ways to summarize and visualize it.

Summarizing and visualizing variables and relationships between two variables is often known as exploratory data analysis (also known as descriptive statistics).

The type of summary statistics and visualization methods to use depends on the type of variables being analyzed (i.e., categorical or quantitative).

Community Coalitions

(n = 175)

One Categorical Variable

"What is your race/ethnicity?"

- White
- Black
- Hispanic
- Asian
- Other

Display the number or proportion of cases that fall into each category.

Frequency Table

A frequency table shows the number of cases that fall into each category:

"What is your race/ethnicity?"

White	Black	Hispanic	Asian	Other	*Total*
111	29	29	2	4	*175*

Proportion

The sample proportion ($\hat{p}$) of directors in each category is

$$\hat{p} = \frac{\text{number of cases in category}}{\text{total number of cases}}$$

Proportion

White	Black	Hispanic	Asian	Other	*Total*
111	29	29	2	4	*175*

The sample proportion of directors who are white is:

$$\hat{p} = \frac{111}{175} \approx .63\ (63\%)$$

Proportion and percent can be used interchangeably.

Relative Frequency Table

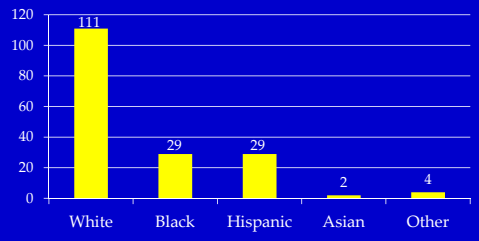
A relative frequency table shows the proportion of cases that fall in each category.

White	Black	Hispanic	Asian	Other
.63	.17	.17	.01	.02

All the numbers in a relative frequency table sum to 1.

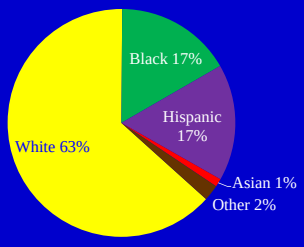
Bar Chart

In a bar chart, the height of the bar corresponds to the number of cases that fall into each category.



Pie Chart

In a pie chart, the relative area of each slice of the pie corresponds to the proportion/percentage in each category.



Two Categorical Variables

Look at the relationship between two categorical variables

1. Race/Ethnicity
2. Gender

Two-Way Table

	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

It doesn't matter which variable is displayed in the rows and which in the columns.

Two-Way Table

	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

What proportion of female directors are Hispanic?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

Two-Way Table

	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

What proportion of Hispanic directors are female?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

Two-Way Table

	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

The proportion of female directors that are Hispanic $\neq$ The proportion of Hispanic directors that are female

Two-Way Table

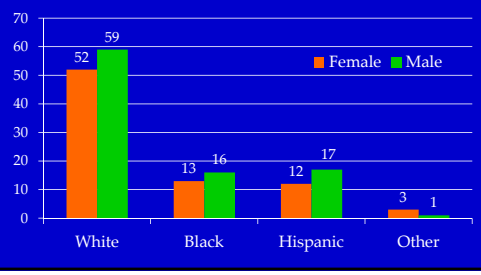
	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

What proportion of directors are female and Hispanic?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

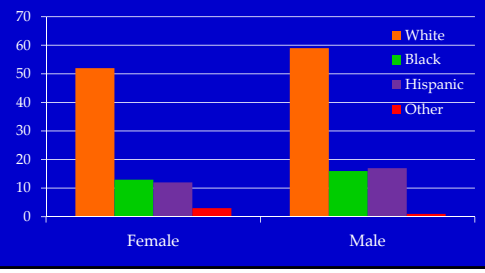
Side-by-Side Bar Chart

In a side-by-side bar chart, the height of each bar corresponds to the number of cases that fall into each category of the table



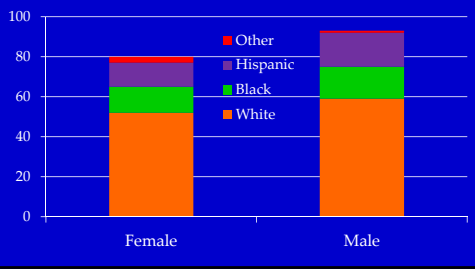
Side-by-Side Bar Chart

In a side-by-side bar chart, the height of each bar corresponds to the number of cases that fall into each category of the table



Segmented Bar Chart

A segmented bar chart is like a side-by-side bar chart, but the bars are stacked instead of side-by-side



Difference in Proportions

A difference in proportions is…

the difference in proportions for one categorical variable (e.g., the proportion who are Hispanic)

calculated for the different levels of another categorical variable (e.g., gender)

Difference in Proportions

What is the difference in proportion of male directors who are Hispanic and female directors who are Hispanic?

$\hat{p}_{M_H}$ = sample proportion of male directors who are Hispanic

$\hat{p}_{F_H}$ = sample proportion of female directors who are Hispanic

Difference in Proportions = $\hat{p}_{M_H} - \hat{p}_{F_H}$

Two-Way Table

	Female	Male	*Total*
White	52	59	*111*
Black	13	16	*29*
Hispanic	12	17	*29*
Other	4	2	*6*
Total	*81*	*94*	*175*

What is the difference in gender proportions among Hispanic directors? $\hat{p}_{M_H} - \hat{p}_{F_H}$

The proportion of male directors who are Hispanic – 17/94

–The proportion of female directors who are Hispanic – 12/81

.033

One Quantitative Variable

When describing quantitative variables we are interested in the distribution of the values – it's shape, center, and spread.

Shape: Form of the distribution of values

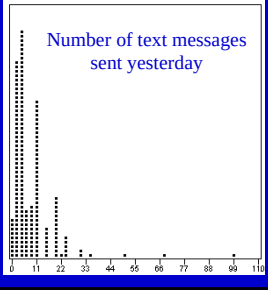
Center: Main peak

Spread: Relative deviation of the values

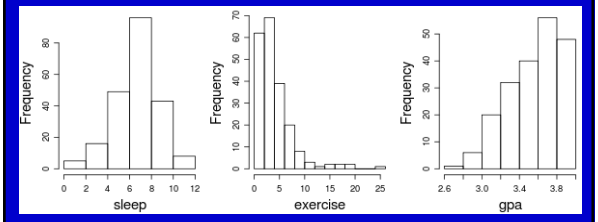
To understand these concepts we'll look at quantitative variables from the student survey.

Dotplot

In a dotplot, each case is represented by a dot and dots are stacked.



Histogram



Create "bins" (i.e., value intervals) and place each case in the appropriate bin based on its value for the variable of interest.

The height of the each bar corresponds to the number of cases that have values falling within that particular interval.

Bar Charts vs. Histograms

Although they look similar, a histogram is not the same as a bar chart.

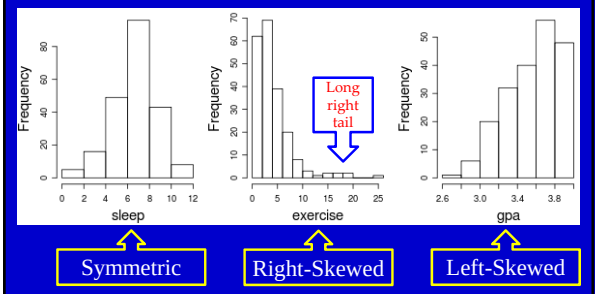
A bar chart is for categorical data, and the x-axis has no numeric scale.

A histogram is for quantitative data, and the x-axis is numeric.

For a categorical variable, the number of bars equals the number of categories, and the number in each category is fixed.

For a quantitative variable, the number of bars (or bins) in a histogram is up to you, and the appearance can differ with different number of bars.

Shape



Measures of Center

Notation

The sample size, the number of cases in the sample, is denoted by n.

A variable is often denoted by x, and $x_1, x_2, \ldots, x_n$ represent the n values of the variable x.

Example: x = The number of body piercings

gender	intro_extro	piercings	sleep
female	extravert	6	6
male	introvert	0	9
female	introvert	1	7
female	introvert	5	7
male	extravert	NA	8
female	extravert	2	9
female	introvert	2	3
female	extravert	2	7
male	introvert	0	7

$x_1 = 6$
$x_2 = 0$
$x_3 = 1$
$x_4 = 5$
…

Measures of Center
Mean

The sample mean ($\bar{x}$) is the average, and is computed by adding up all the numbers and dividing by the number of cases.

$$\text{Sample Mean: } \bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

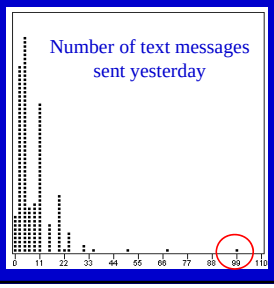
Measures of Center
Median

The sample median (m) is the middle value when the data is ordered.

If there are an even number of values, the median is the average of the two middle values.

Outliers

An outlier is a value that is notably different from the other values (e.g., much larger or smaller than the other values)



Resistance

A statistic is resistant if it is not heavily affected by outliers.

The median is resistant, the mean is not resistant.

Number of text messages sent per day:

	Mean	Median
Outlier Included	32.6	8
Outlier Removed	9.2	8

Outliers

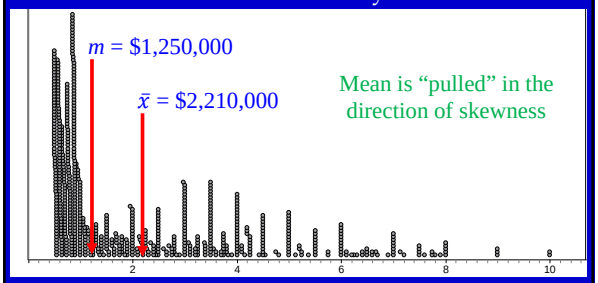
When calculating statistics that are not resistant to outliers, look for outliers and decide whether the outlier is a mistake.

If not, you have to decide whether the outlier is part of your population of interest or not.

Usually, for outliers that are not a mistake, it's best to run the analysis twice, once with the outlier(s) and once without, to see how much the outlier(s) affect the results.

Measures of Center

Distribution of NHL Player Salaries



Assignment

Part I

Graded Problems

2.18 and 2.60

Additional Practice Problems *(not to be turned in)*:

2.11 and 2.57

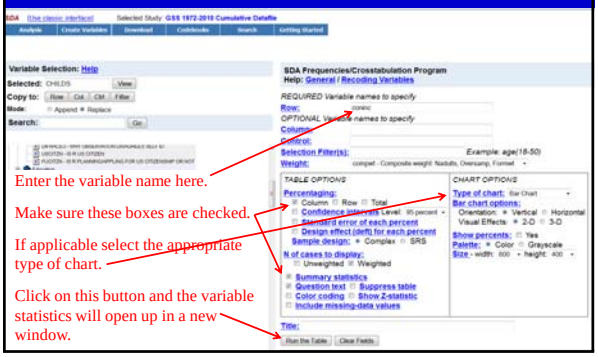
Part II

Goto http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10

Find 3 categorical variables and provide the proportion for each category for each variable.

Find 3 quantitative variables and provide the mean & median and whether the distribution of values are symmetric, right skewed, or left skewed.

Getting Variable Statistics from the GSS



Summary:
One Categorical Variable

Summary Statistics

- Proportion
- Frequency table
- Relative frequency table

Visualizations

- Bar chart
- Pie chart

Summary:
Two Categorical Variables

Summary Statistics

- Two-way table
- Difference in proportions

Visualizations

- Side-by-side bar chart
- Segmented bar chart