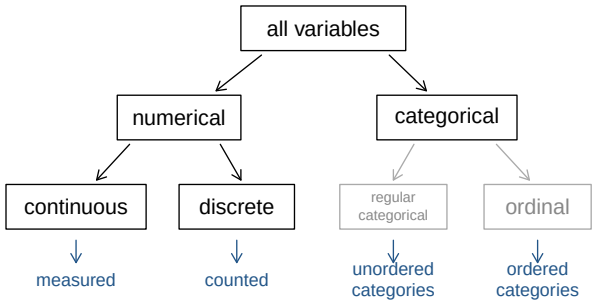
Lecture 3: Data Description

STA3100: Programming with Data

Types of variables



Variability

Variability is a very important concept in statistics. The higher the variability, the harder it is to draw conclusions.

Common sources of variability:

- Variability across measurements - measurement error can cause the recorded length to vary from measurement to measurement.
- Variability across subjects - average pulse rate varies from moment to moment and from person to person
- Variability across groups - in a comparison between men and women, need to take into account variability within men and women.

Single Variable: Numerical descriptive measures

The most common numerical descriptive measures are:

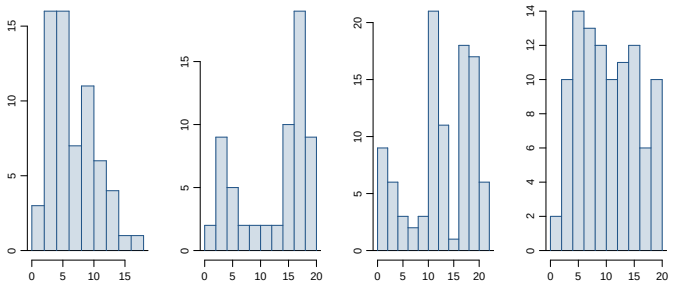
- Central tendency measures: mean, median, mode
- Variability measures: variance, standard deviation, range, IQR

population $\longrightarrow$ *parameters*
sample $\longrightarrow$ *statistics*

	mean	variance	SD
sample	$\bar{x}$	s^2	s
population	μ	σ^2	σ

Shape of a distribution: modality

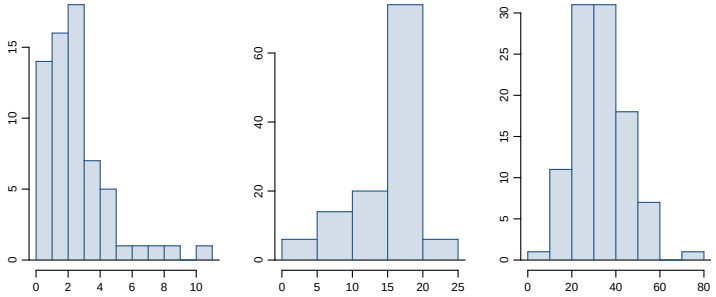
The *mode* is defined as the most frequent observation in the data set. Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



In order to determine modality, it's best to step back and imagine a smooth curve over the histogram.

Shape of a distribution: skewness

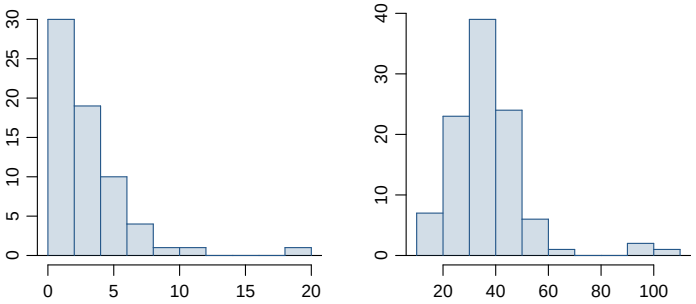
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

Are there any unusual observations or potential *outliers*?

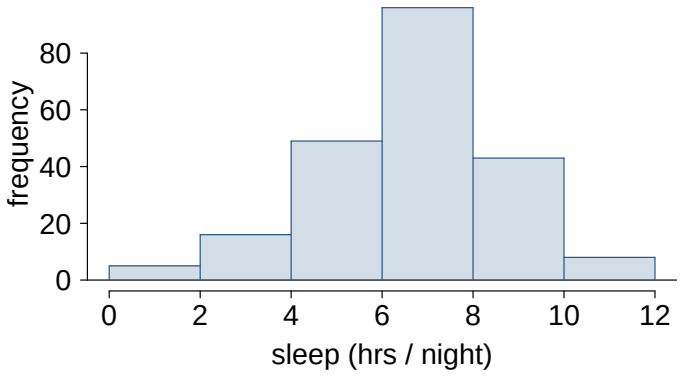


How would you expect each of the following to be distributed?

a. Weights of adult males
b. Salaries of people from Florida
c. Exam scores on an easy test
d. Birthdays of classmates (day of the month)

Variability in data

How would you describe the amount of variability in the number of hours of sleep students get per night?



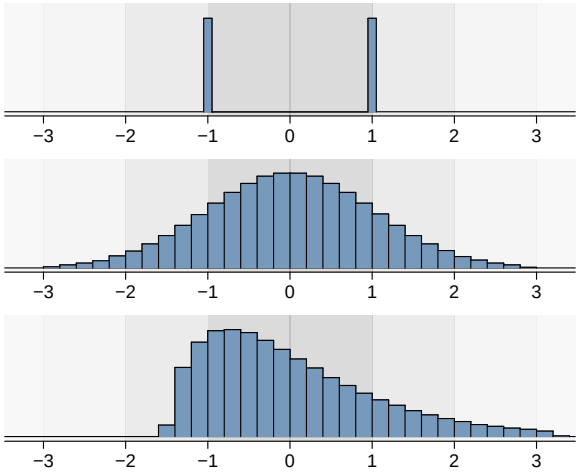
Sample Variance and Standard Deviation

```
sort(d$sleep)
  [1]  1  1  1  2  2  3  3  3  3  3  3  3  3  3  3  3  3  3  4  4  5  5  5  5  5  5  5  5  5  5  5  5
 [35]  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5
 [69]  6  6  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
[103]  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
[137]  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  8  8  8  9  9  9  9  9
[171]  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9
[205]  9  9  9  9 10 11 11 11 11 11 11 11 12
```

```
mean(d$sleep, na.rm=TRUE)
[1] 6.709677
median(d$sleep, na.rm=TRUE)
[1] 7
var(d$sleep, na.rm=TRUE)
[1] 4.11
sd(d$sleep, na.rm=TRUE)
[1] 2.02
```


Describing distributions

When describing distributions make sure to talk about the shape, center, spread, and if any, unusual observations.



Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, *Q3*.

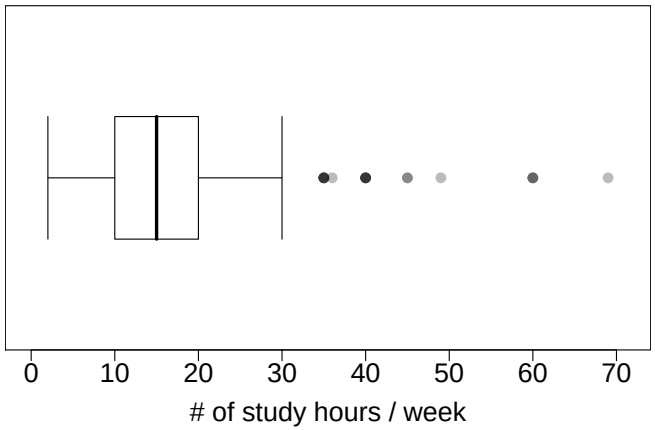
```
summary(d$study_hours)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2.00   10.00   15.00   17.45   20.00   69.00   20.00
quantile(d$study_hours, prob=0.25, na.rm=TRUE)
   25\%
   10
```

- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

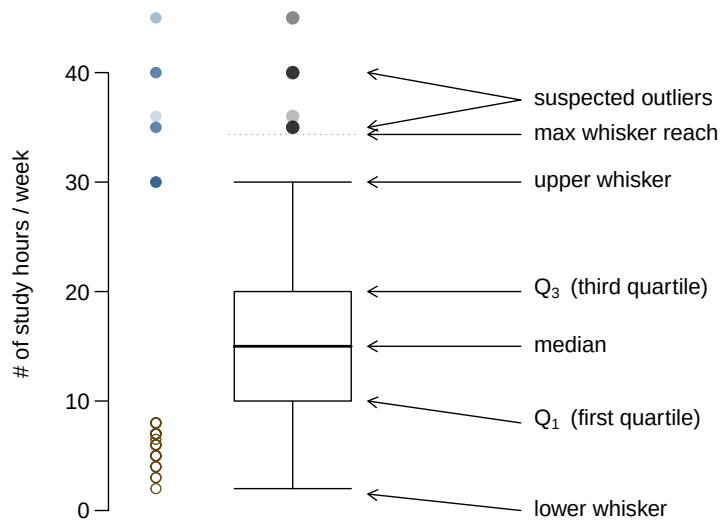
$$IQR = 20 - 10 = 10$$

Box plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a box plot



Whiskers and outliers

- *Whiskers* of a box plot can extend up to 1.5 * IQR away from the quartiles.

  max upper whisker reach : $Q3 + 1.5 * IQR = 20 + 1.5 * 10 = 35$

  max lower whisker reach : $Q1 - 1.5 * IQR = 10 - 1.5 * 10 = -5$

- The *whiskers* are calculated as the min/max value between the data points within the max whisker reach and the max whisker reach limit:

  upper whisker : $min(max(x), Q3 + 1.5 * IQR) = 30$

  lower whisker : $max(min(x), Q1 - 1.5 * IQR) = 2$

- An *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

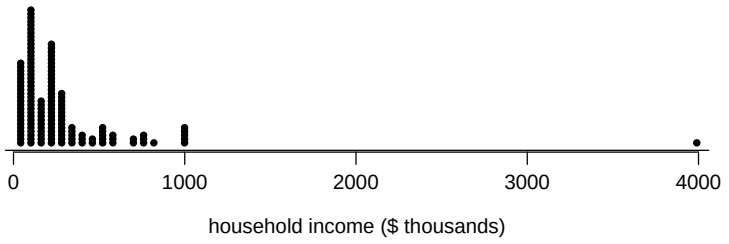
Outliers (cont.)

Why is it important to look for outliers?

Extreme observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million? What if the smallest value was replaced with $10 million?

Robust statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	165K	150K	211K	180K
move largest to $10 million	165K	150K	398K	1,422K
move smallest to $10 million	190K	163K	4,186K	1,424K

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is more appropriate to use median and IQR to describe the center and spread
- for symmetric distributions it is more appropriate to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a UF student, would you be more interested in the mean or median income?

More than one variable: Calculating the correlation

- Pearson correlation coefficient:

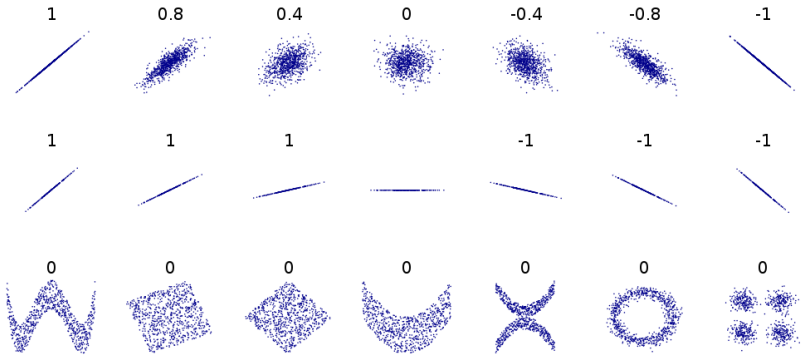
$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- *Correlation* describes the strength and direction of the *linear* relationship between two variables.
- It takes values between -1 (perfect negative relationship) and +1 (perfect positive relationship).
- A value of 0 indicates no linear relationship.
- Using R:

```
cor(x, y)
```

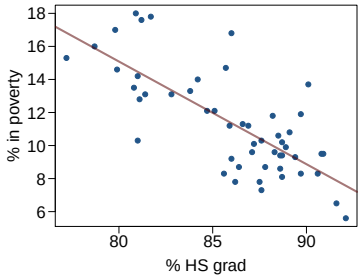
Quantifying the relationship

Correlation means linear association!



Describing the relationship: Shape, Direction, and Strength

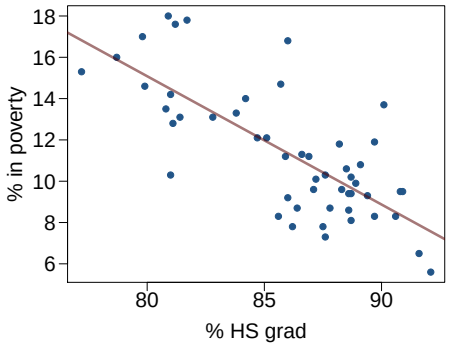
The *scatterplot* below shows the relationship between HS graduate rate in all 50 states and the District of Columbia vs the % of residents who live below the poverty line ($<$ \$22,350 for a family of 4).



How would you describe the relationship between % in poverty and % HS grad?

Guessing the correlation

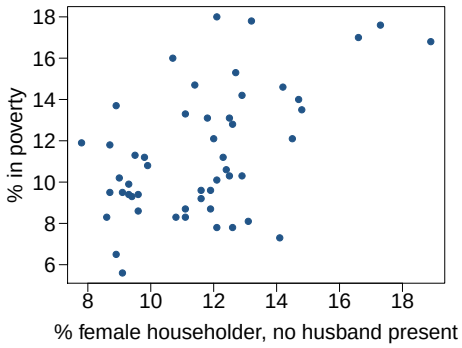
Which of the following is the best guess for the correlation between % in poverty and % HS grad?



(a) 0.6 (b) -0.75 (c) -0.1 (d) 0.02 (e) -1.5

Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?



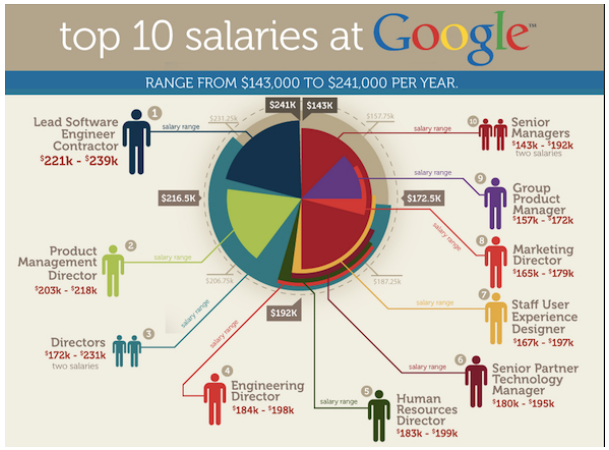
(a) 0.1 (b) -0.6 (c) -0.4 (d) 0.9 (e) 0.5

Graphical description

Depending on the type of measurement, common plots are pie charts, bar charts, histograms, box plots, and density plots.

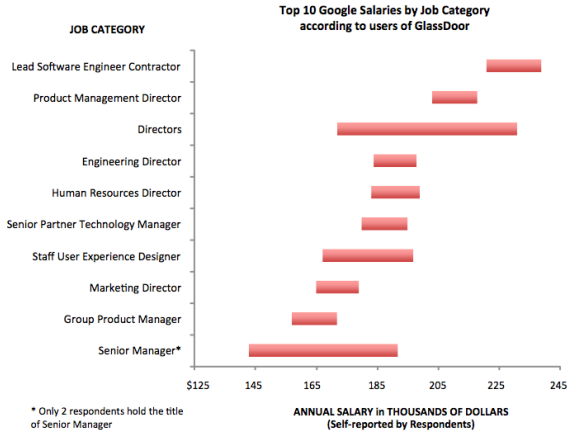
- *Pie charts:* Used to describe any variable type.
  - Continuous numeric variables must be collapsed into bins or buckets.
  - The size of the sectors of the pie represent the relative frequency of each category.
- *Bar charts:* Nominal or ordinal data.
  - The variable levels are arrayed on the bottom (or left side) of the plot.
  - Bars above (or beside) the levels represent the range of values taken by the response variable or the frequency or relative frequency of the number of observations belonging to the various categories.

Problem with Pie Charts



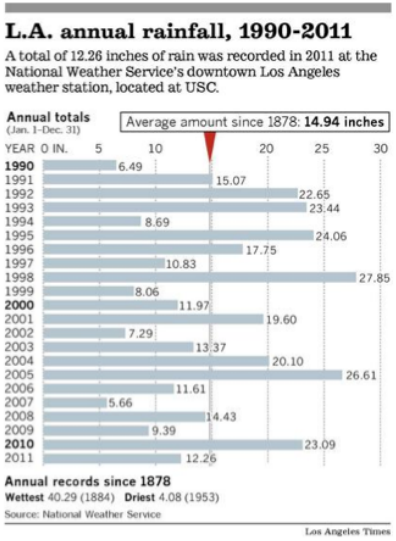
http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

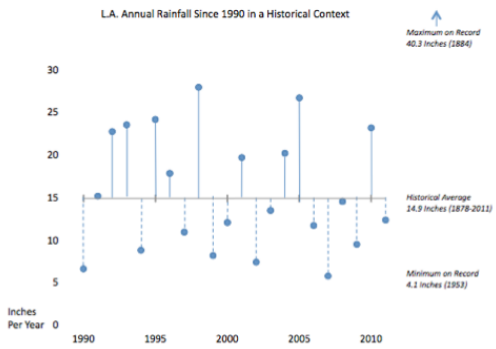
Problem with Pie Charts



http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

Axis Placement - What story to tell





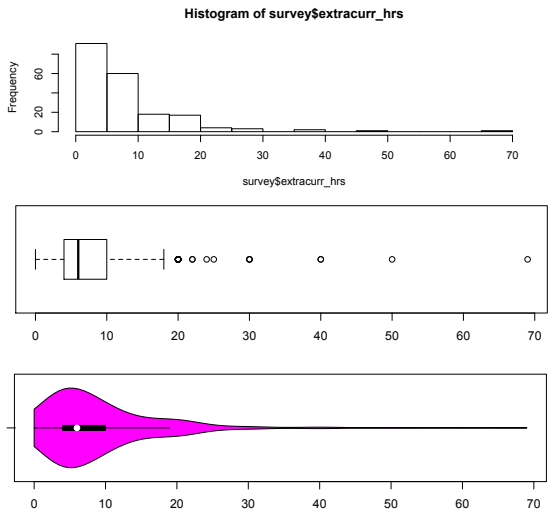
Graphical description (cont.): Numeric Variables

- *Histograms:* The heights of the bars above the bins represent the frequency or relative frequency of the various bins.
- *Box plots:* They identify particular percentiles of a distribution and are useful in detecting outlying observations and spread in the distribution.
- *Density plots:* They represent smoothed histograms describing the proportion of measurements within some distance of each point on the continuum.
- *Violin plots:* Similar to box-plot but allows to see better the shape of the distribution.

Univariate Plots

- Categorical
  - "Good" - Bar plot
  - "Bad" - Pie chart
- Numerical
  - "Good" - boxplot (outliers), histogram/density (shape), violin plot (shape and outliers)
  - "Bad" - Stem and leaf plot

Univariate Plots - Numerical

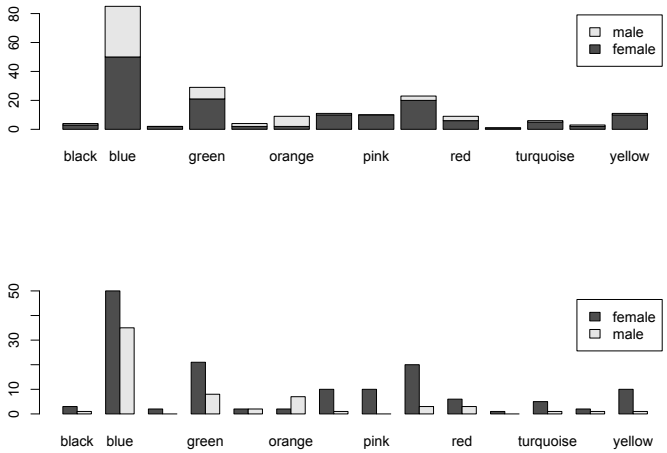


Bivariate Plots

- Categorical vs Categorical
  - Stacked or grouped bar plot, mosaic plots
- Numerical vs Categorical
  - Side by side boxplot or violin plot
- Numerical vs Numerical
  - Scatter plot, line plot

Bivariate Plots - Bar plots

Favorite color by gender:



Bivariate Plots - Mosaic plot

Titanic survivors by gender: