Lecture 1 - Data and Data Summaries

Statistics 102

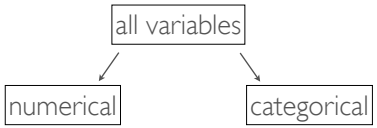
Colin Rundel

January 13, 2013

Announcements

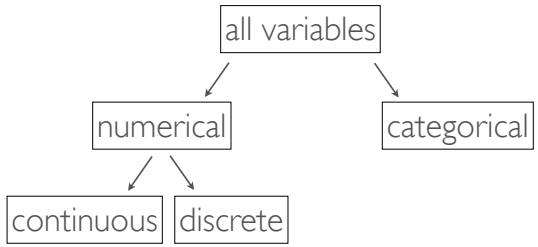
- Homework 1 - Out 1/15, due 1/22
- Lab 1 - Tomorrow
  - RStudio accounts created this evening
  - Try logging in at *http://beta.rstudio.org*
- Practice Quiz - In class 1/15
  - Not graded, chance to try clicker submission process.

Data



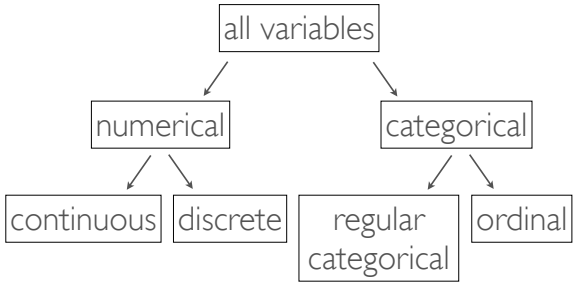
- *Numerical (quantitative)* - takes on a numerical values
  - Ask yourself - is it sensible to add, subtract, or calculate an average of these values?
- *Categorical (qualitative)* - takes on one of a set of distinct categories
  - Ask yourself - are there only certain values (or categories) possible? Even if the categories can be identified with numbers, check if it would be sensible to do arithmetic operations with these values.

Numerical Data



- *Continuous* - data that is measured, any numerical (decimal) value
- *Discrete* - data that is counted, only whole non-negative numbers

Categorical Data



- *Ordinal* - data where the categories have a natural order
- *Regular categorical* - categories do *not* have a natural order

Example - Class Survey

Students in an introductory statistics course were asked the following questions as part of a class survey:

1. What is your gender?
2. Are you introverted or extraverted?
3. On average, how much sleep do you get per night?
4. When do you go to bed: 8pm-10pm, 10pm-12am, 12am-2am, later than 2am?
5. How many countries have you visited?
6. On a scale of 1 (very little) - 5 (a lot), how much do you dread this semester?

What type of data is each variable?

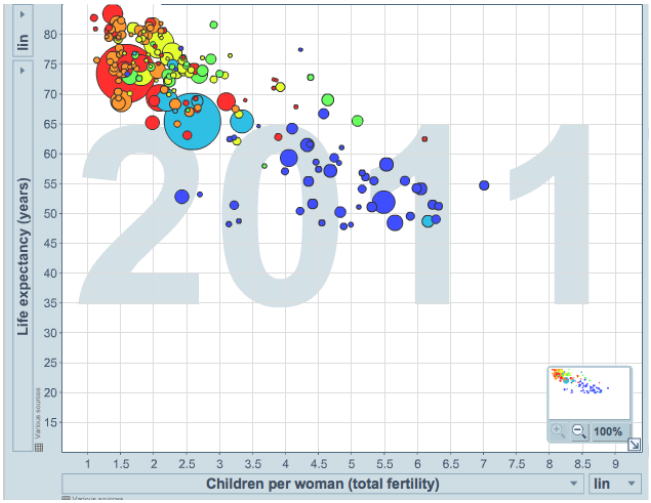
Representing Data - Class Survey

We use a *data matrix (data frame)* to represent responses from this survey.

- Columns represent *variables*
- Rows represent *observations (cases)*

student	gender	intro_extra	sleep	bedtime	countries	dread
1	male	extravert	8	10-12	13	3
2	female	extravert	8	8-10	7	2
3	female	introvert	5	12-2	1	4
4	female	extravert	6.5	12-2	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
86	male	extravert	7	12-2	5	3

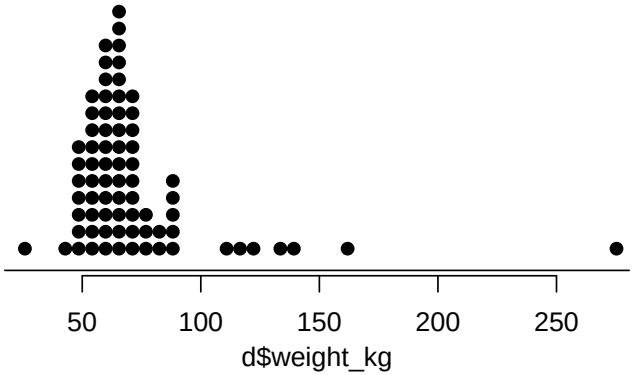
Scatterplots



http://www.gapminder.org/world

Dot plots

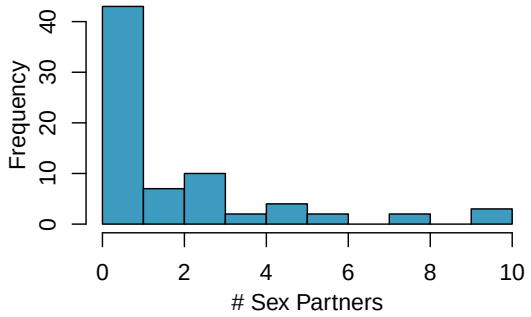
Useful for visualizing a single numerical variable, especially useful when individual values are of interest.



Do you see anything out of the ordinary?

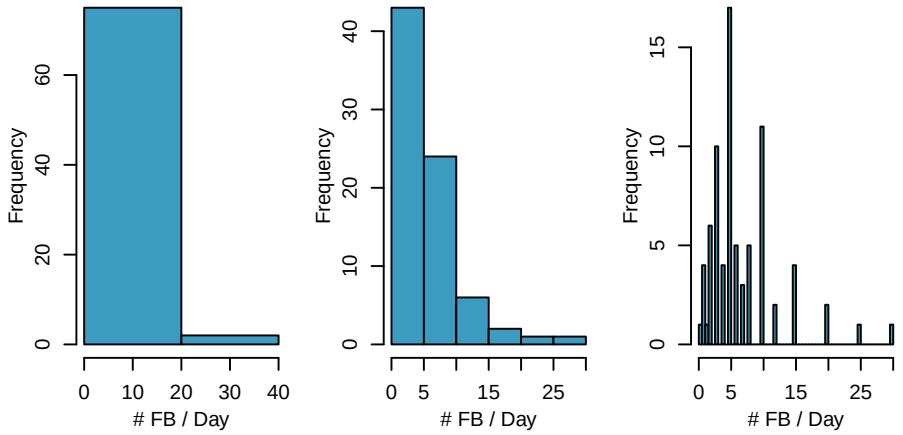
Histograms

- Preferable when sample size is large but hides finer details like individual observations.
- Histograms provide a view of the data's *density*, higher bars represent where the data are more common.
- Histograms are especially useful for describing the *shape* of the distribution.



Bin width

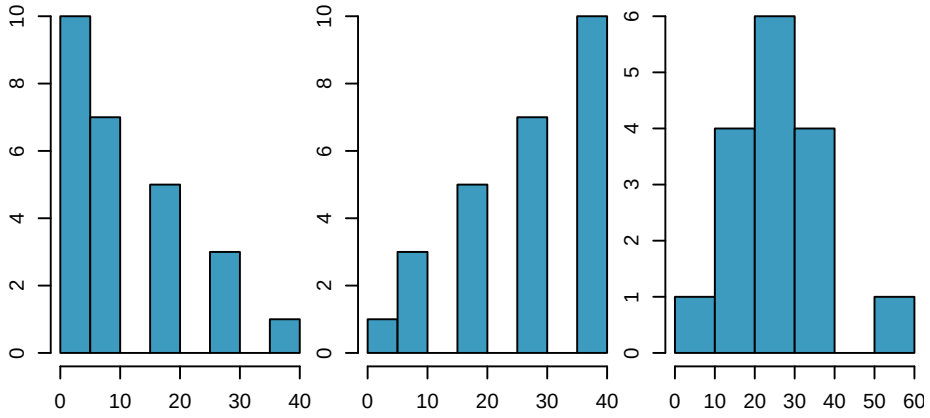
The chosen *bin width* can alter the story the histogram is telling.



Which histogram is the most useful? Why.

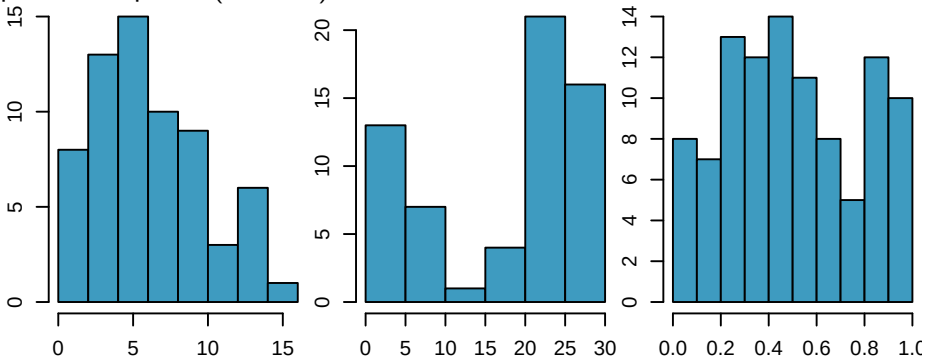
Skewness

Histograms are said to be skewed towards the direction with the longer tail. A histogram can be *right skewed*, *left skewed*, or *symmetric*.



Modality

This describes the pattern of the peaks in peaks in the histogram - a single prominent peak (*unimodal*), several (*bimodal/multimodal*), or no prominent peaks (*uniform*)?



Note: *In order to determine modality, it's best to step back and imagine a smooth curve (limp spaghetti) over the histogram.*

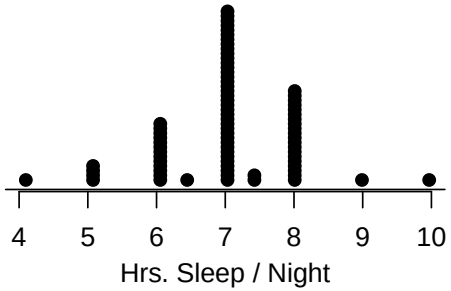
Clicker Question

Which of the following variables is most likely to be uniformly distributed?

1. weights of adult females
2. salaries of a random sample of people from North Carolina
3. exam scores
4. birthdays of classmates (day of the month)

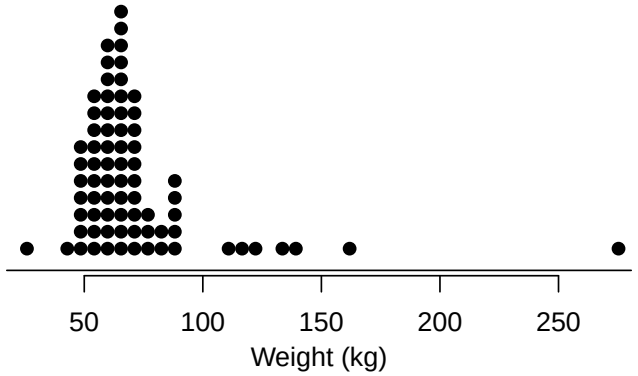
Guess the center

What would you guess is the average numer of hours students sleep per night?



Guess the center, cont.

What would you guess is the average weight of students?



Mean

- *Sample mean* ($\bar{x}$) - Arithmetic average of values in sample.

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Population mean* (μ) - Computed the same way but it is often not possible to calculate μ since population data is rarely available.

$$\mu = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The sample mean is a *sample statistics*, or a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population) it is usually a good guess.

Are you typical?

http://www.youtube.com/watch?v=4B2xOvKFFz4

Variance

Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Population Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Roughly the average squared deviation from the mean.

Why do we use the squared deviation in the calculation of variance?

Standard deviation

Sample SD

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

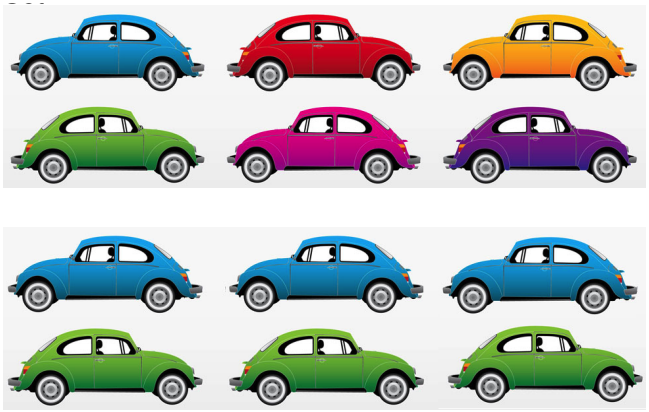
Population SD

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Note that variance has square units while the SD has the same units as the data - this leads to a more natural interpretation.

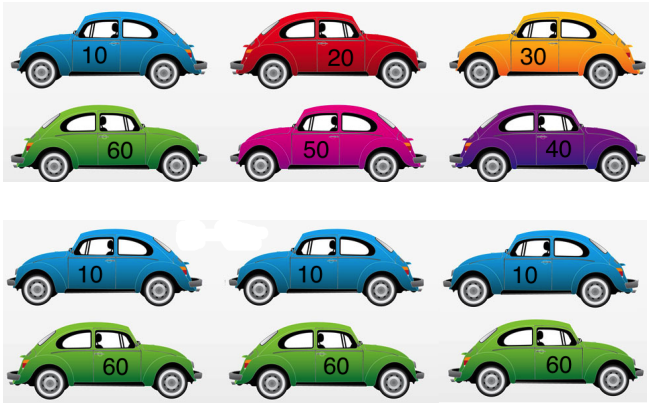
Diversity vs Variability

Which group of cars has a more diverse set of colors?



Diversity vs Variability (cont.)

Which group of cars has a more variable mileage?



Median, Quartiles, and IQR

- The *median* is the value that splits the data in half when ordered in ascending order, i.e. 50^{th} *percentile*.

$$0, 1, 2, 3, 4$$

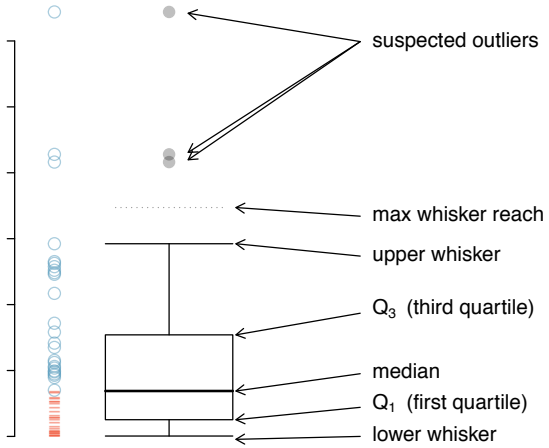
  - If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 75^{th} percentile is also called the third quartile, *Q3*.
- The range the middle 50% of the data span is called the *interquartile range*, or the *IQR*.

Box plot

A *box plot* visualizes the median, the quartiles, and suspected outliers.



Box plot - Example

Resting Pulse

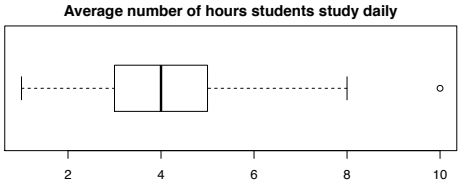
62, 64, 68, 70, 70, 74, 74, 76, 76, 78, 78, 80

Steps:

1. Calculate median, Q1, Q3, IQR, min, and max
2. Calculate upper and lower fences (Q1 - 1.5 IQR, Q3 + 1.5 IQR)
3. Find the location of the upper and lower wiskers
4. Consider data points outside whiskers as potential outliers

Clicker Question

Which of the following is false about the distribution of average number of hours students study daily




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.821   5.000  10.000
```

a) There are no students who don't study at all.
b) 75% of the students study more than 5 hours daily, on average.
c) 25% of the students study less than 3 hours, on average.
d) IQR is 2 hours.

¨¡¼

Robust statistics

The median and IQR are examples of what are known as robust statistics - because they are less affected by skewness and outliers than statistics like mean and SD.

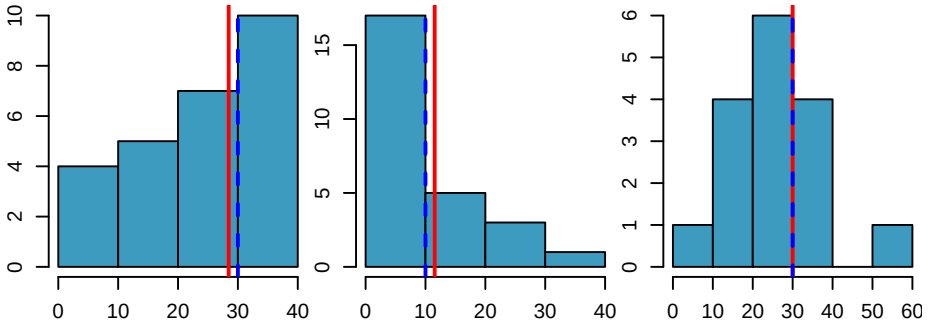
As such:

- 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 were searching for a car and are price conscious, should you be more interested in the mean or median vehicle price when considering a car?

Mean vs. median

- If the distribution is symmetric, center is the mean
  - Symmetric: mean = median
- If the distribution is skewed or has outliers center is the median
  - Right-skewed: mean > median
  - Left-skewed: mean < median

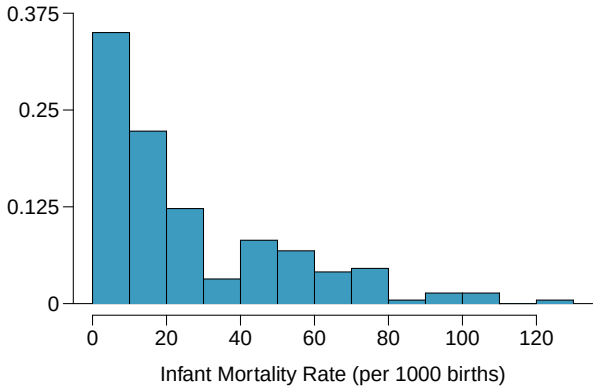


red solid - mean, black dashed - median

Relative Frequency Histograms

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. The relative frequency histogram below shows the distribution of estimated infant death rates in 2012 for 222 countries.

Where would you estimate the third quartile to be located?



Tables and Contingency tables

We might be interested in looking at if there is a relationship between religion belief in God and gender, in which case we need to summarize both variables:

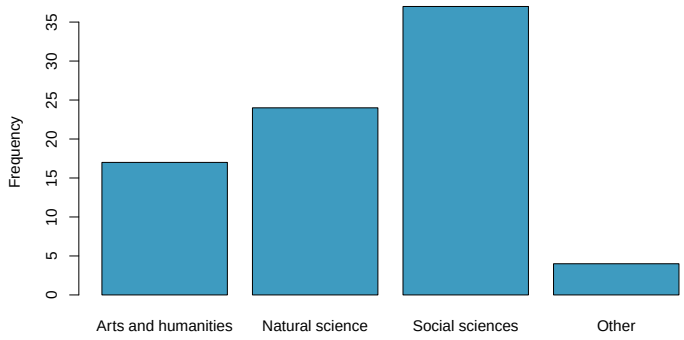
No	Somewhat	Yes
22	23	36

Female	Male
57	25

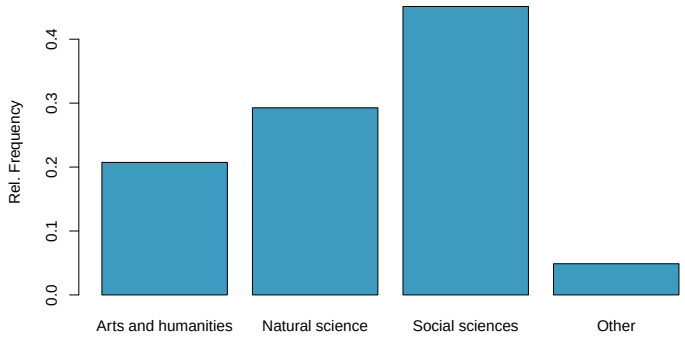
but this is not enough alone.

Female	Male
14	8
16	7
26	10

Barplots - Absolute vs Relative



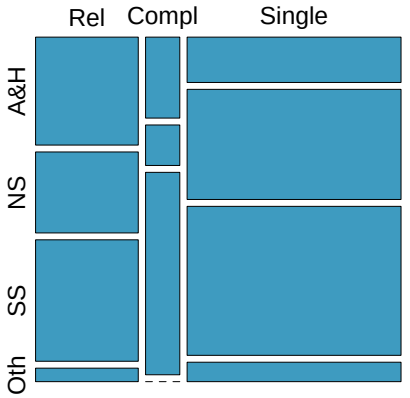
Barplots - Absolute vs Relative



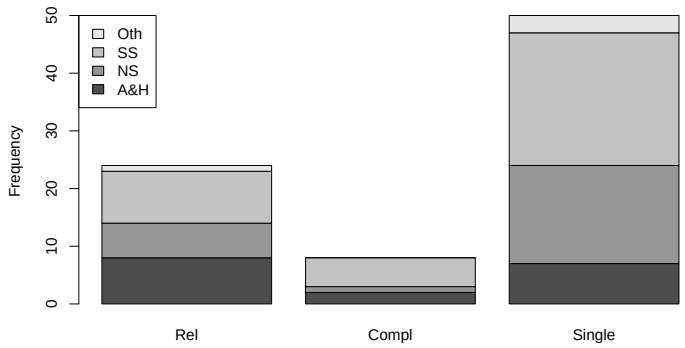
Mosaic plots

Is there a relationship between major and relationship status?

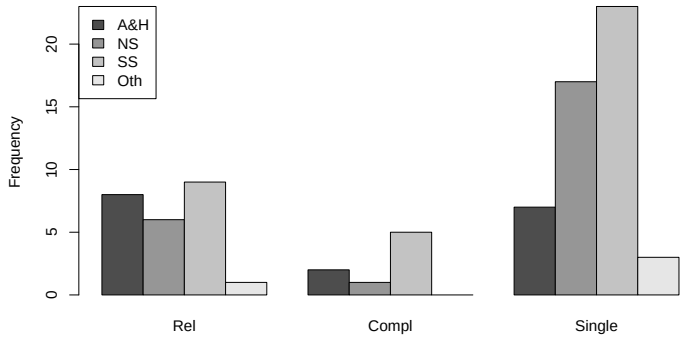
Rel	Compl	Single
8	2	7
6	1	17
9	5	23
1	0	3



Bivariate Barplots - Stacked vs Juxtaposed

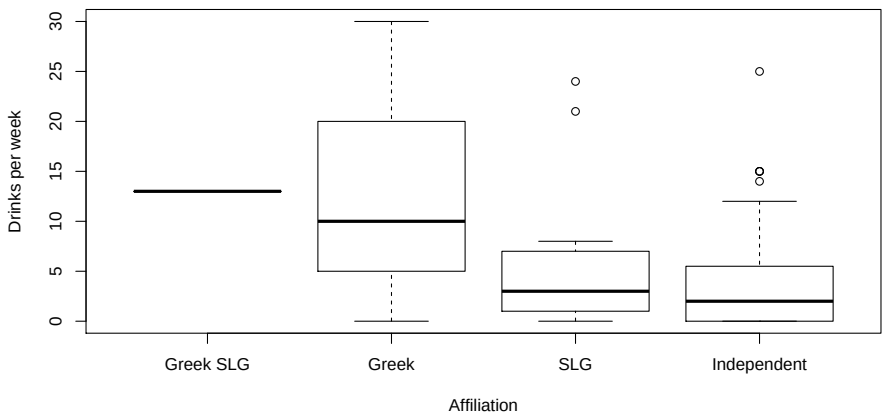


Bivariate Barplots - Stacked vs Juxtaposed



Side-by-side box plot

How does number of drinks consumed per week vary by affiliation?



Visualization Summary

- Single numeric - dot plot, box plot, histogram
- Single categorical - bar plot (or a table)
- Two numeric - scatter plot
- Two categorical - mosaic plot, stacked or side-by-side bar plot
- Numeric and categorical - side-by-side box plot

Tufte's Principles:

1. Above all else show data.
2. Maximize the data-ink ratio.
3. Erase non-data-ink.
4. Erase redundant data-ink.
5. Revise and edit