Descriptive Statistics (Numerical Data)

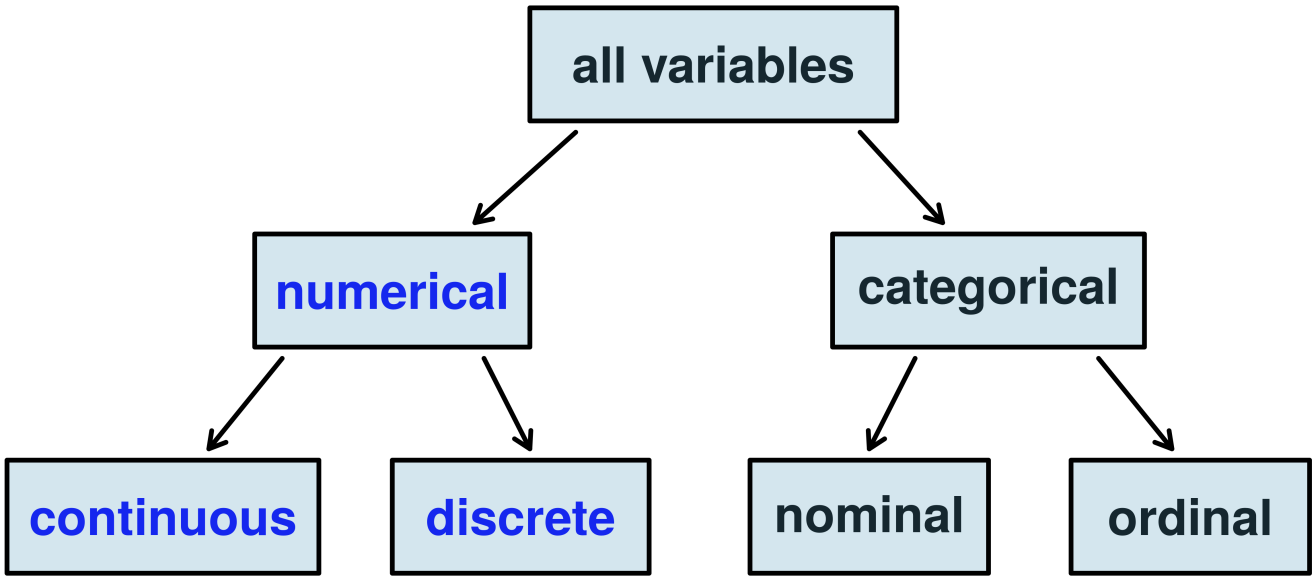
Sumanta Basu

Reading

Textbook section 1.6

Numerical Variables

Recall: Types of Variables



Numerical Variables

Numerical variables

- vary in magnitude; has units of measure
- "on number line"; add/subtract makes sense
- **continuous** (e.g., height)
- **discrete** (e.g., number of…)

As opposed to **Categorical**

- vary in "level", no units of measure
- **nominal** (e.g., favorite food)
- **ordinal** (e.g., highest degree attained)

Recall: summarizing categorical data

Summarize categorical data

- frequencies (counts), proportions

Visualize categorical data

- barplots (of counts or proportions)

> Question: Can we use these ideas to summarize and visualize numerical data as well?

The Titanic

Goal: describe survival patterns for the passengers

The data: - ~1300 passengers

- variables:
- age (in years)
- sex (male/female)
- class (first, second, third)
- survived (yes/no)

The raw data

Open `titanic.txt`.

Loading into R

```
dat <- read.table("titanic.txt", header=TRUE)
head(dat, n=3)
```

```
##                                     Name PClass Age    Sex Survived
## 1        Allen, Miss Elisabeth Walton    1st  29 female        1
## 2         Allison, Miss Helen Loraine    1st   2 female        0
## 3 Allison, Mr Hudson Joshua Creighton    1st  30   male        0
```

Frequency Distribution

Recall: main idea of summarizing variables

- List number of distinct levels (e.g. PClass 1st, 2nd, 3rd), calculate their **frequencies** in data

```
table(dat$PClass)
```

```
##
## 1st 2nd 3rd
## 322 280 711
```

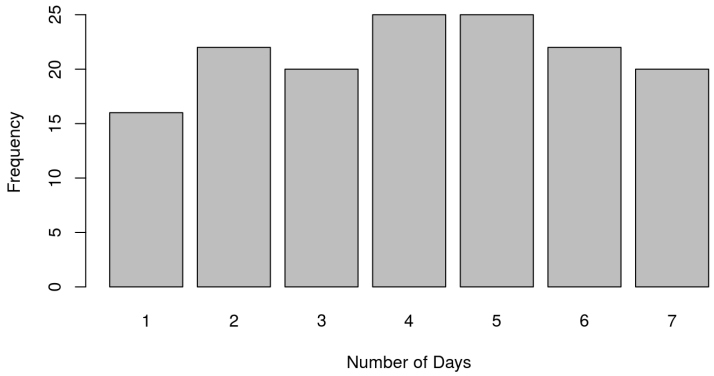
Same concept can be used to summarize **discrete** numerical variables (e.g. number of days in a week a student checks his facebook account)

```
set.seed(1) # set seed to ensure reproducibility
days = sample(1:7, 150, replace = TRUE) # simulate data on 150 students
table(days)
```

```
## days
##  1  2  3  4  5  6  7
## 16 22 20 25 25 22 20
```

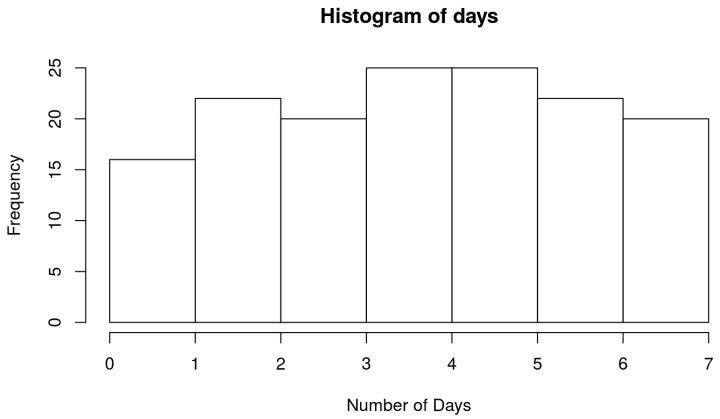
Visualizing Frequency Distributions

```
barplot(table(days),
        xlab = 'Number of Days',
        ylab = 'Frequency')
```



> barplot is for categorical variables, levels may not have numerical meaning

```
hist(days, breaks = 0:7,
     xlab = 'Number of Days',
     ylab = 'Frequency')
```



histogram is for numerical/quantitative variables, plot values on **real line** in the x-axis

How do we look at the ages?

There are too many to list…

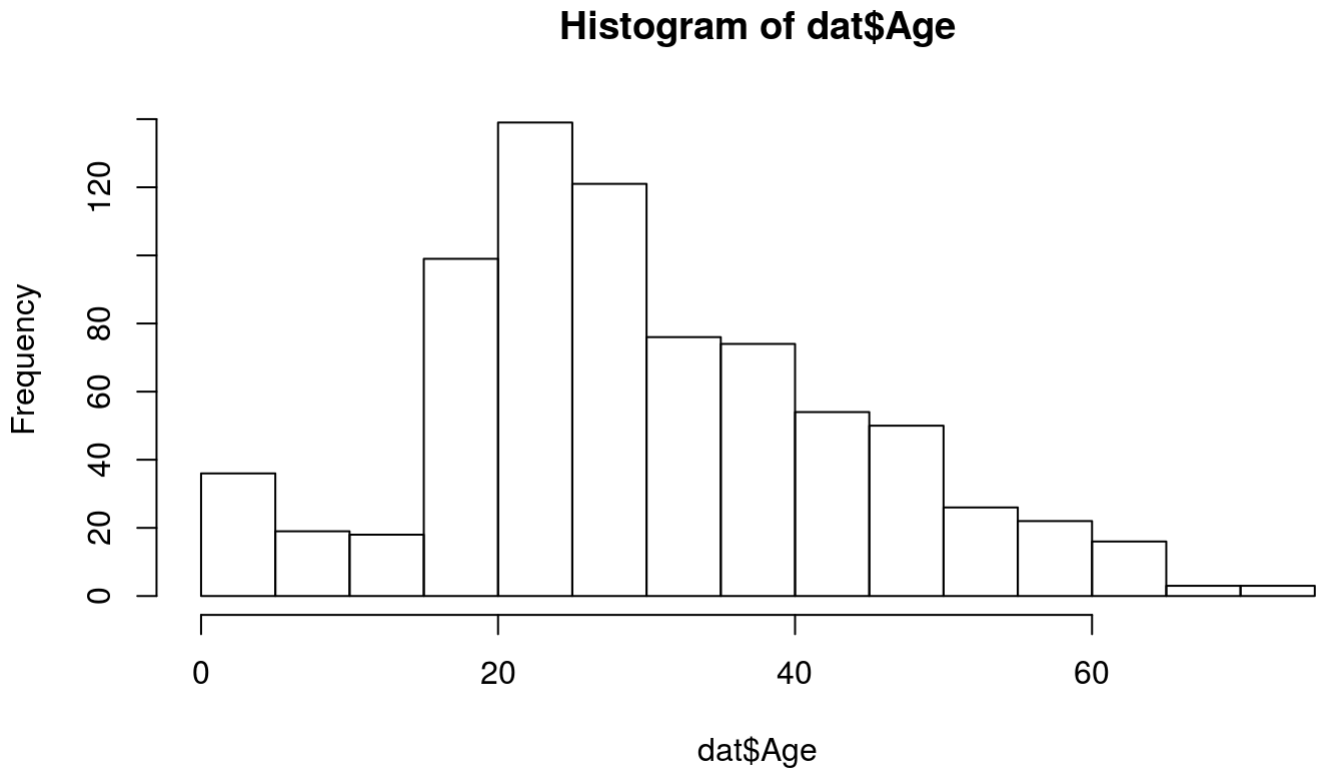
```
library(doBy)
library(openintro); data(COL); twocolors=COL[1:2,1]

dat$Age[1:50] # a look at first 50
```

```
##  [1] 29.00  2.00 30.00 25.00  0.92 47.00 63.00 39.00 58.00 71.00 47.00
## [12] 19.00    NA    NA    NA 50.00 24.00 36.00 37.00 47.00 26.00 25.00
## [23] 25.00 19.00 28.00 45.00 39.00 30.00 58.00    NA 45.00 22.00    NA
## [34] 41.00 48.00    NA 44.00 59.00 60.00 45.00    NA 53.00 58.00 36.00
## [45] 33.00    NA    NA 36.00 36.00 14.00
```

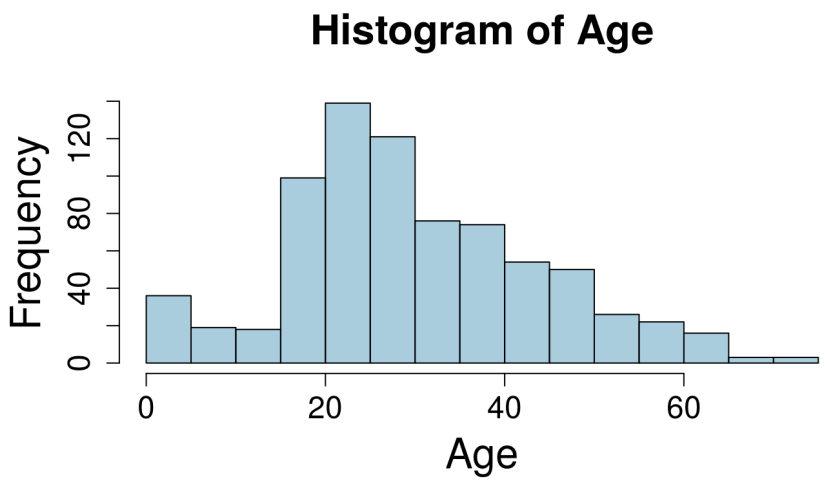
> NA: missing values

Histograms

```
hist(dat$Age)
```



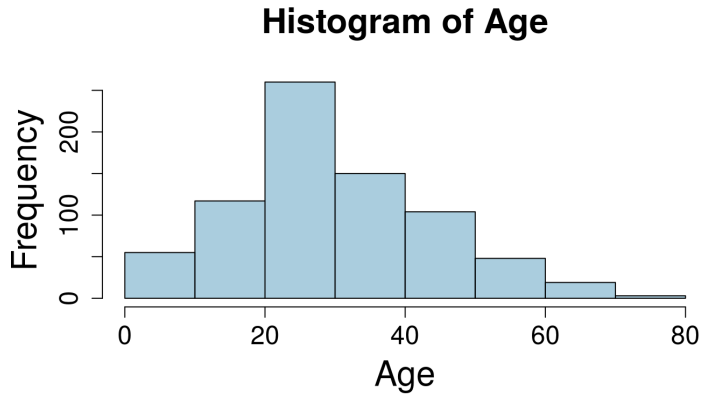
Histograms



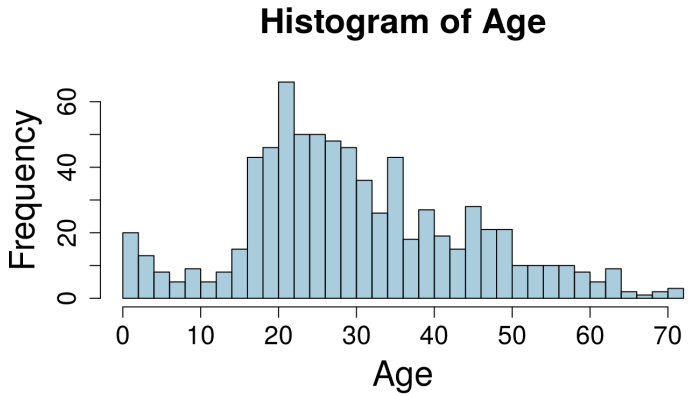
- divide up variable into "bins"
- height shows how many people fall in each bin
- summarizes "shape" of distribution

Effect of bin size

```
hist(dat$Age, breaks=10)
```



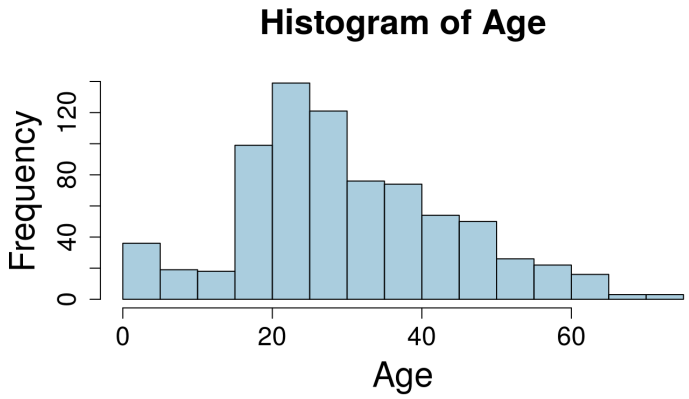

```
hist(dat$Age, breaks=50)
```



> Question: can you separate sub-populations of children and adults?

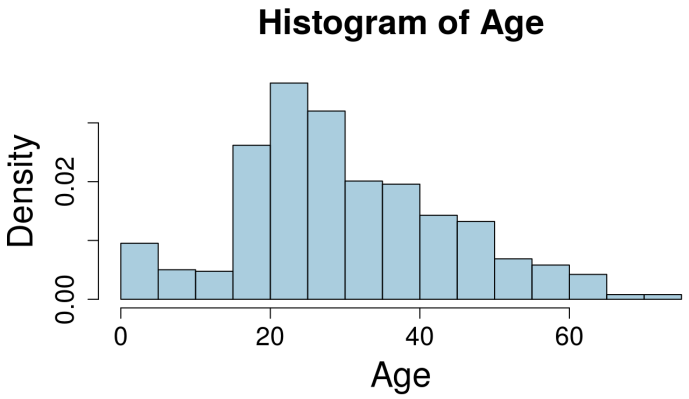
Frequency versus density

```
hist(dat$Age, freq = TRUE)
```



`freq = TRUE`: heights of bars correspond to counts

```
hist(dat$Age, freq = FALSE)
```



`freq = FALSE`: area of a bar correspond to proportion

- total area of all bars will be 1
- height of a single bar (density) can be > 1

Histograms for Exploratory Data Analysis

Old Faithful Geyser (Yellowstone, 1948)

Old Faithful Waiting Times

```
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

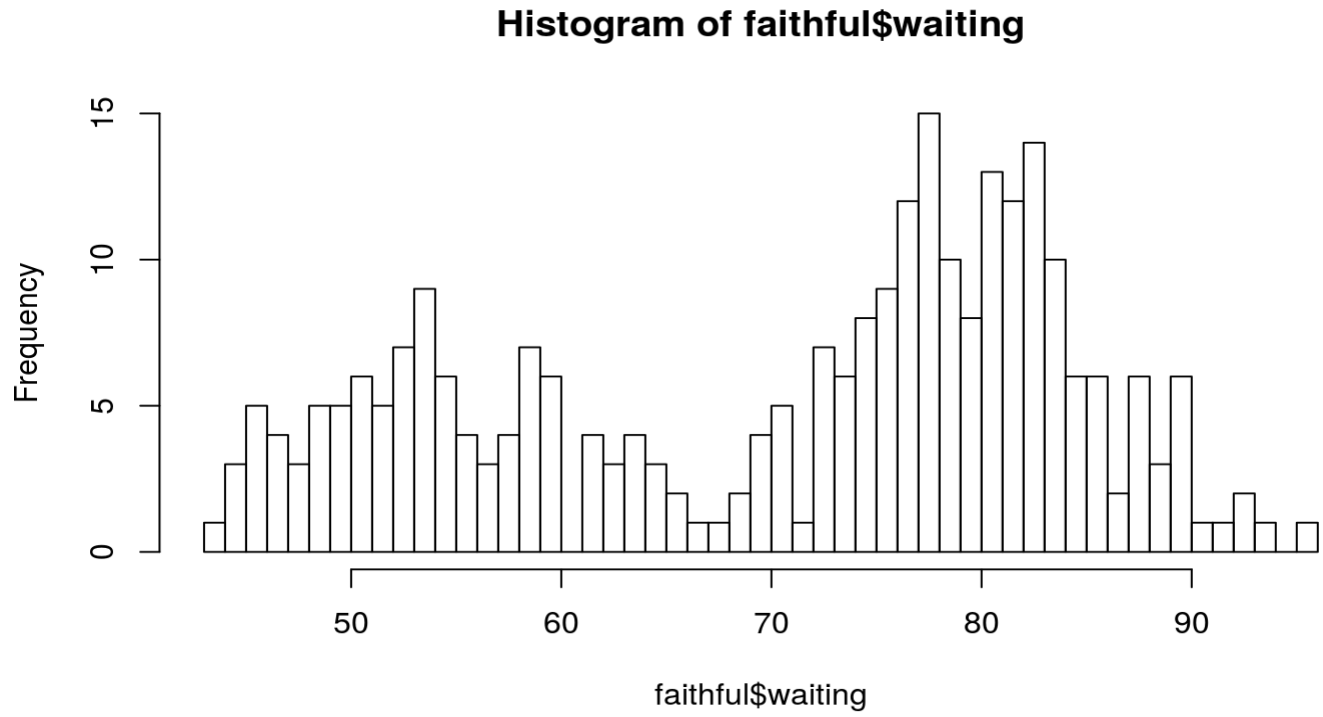
```
mean(faithful$waiting)
```

```
## [1] 70.89706
```

Old Faithful Waiting Times

Histogram shows a bimodal distribution – two subgroups of events?

```
hist(faithful$waiting, breaks=50)
```



Old Faithful Waiting Times

Bimodal Distribution related to length of eruption?

```
# Mean eruption length for shorter waiting times
mean(faithful$eruptions[faithful$waiting < 70])
```

```
## [1] 2.134951
```

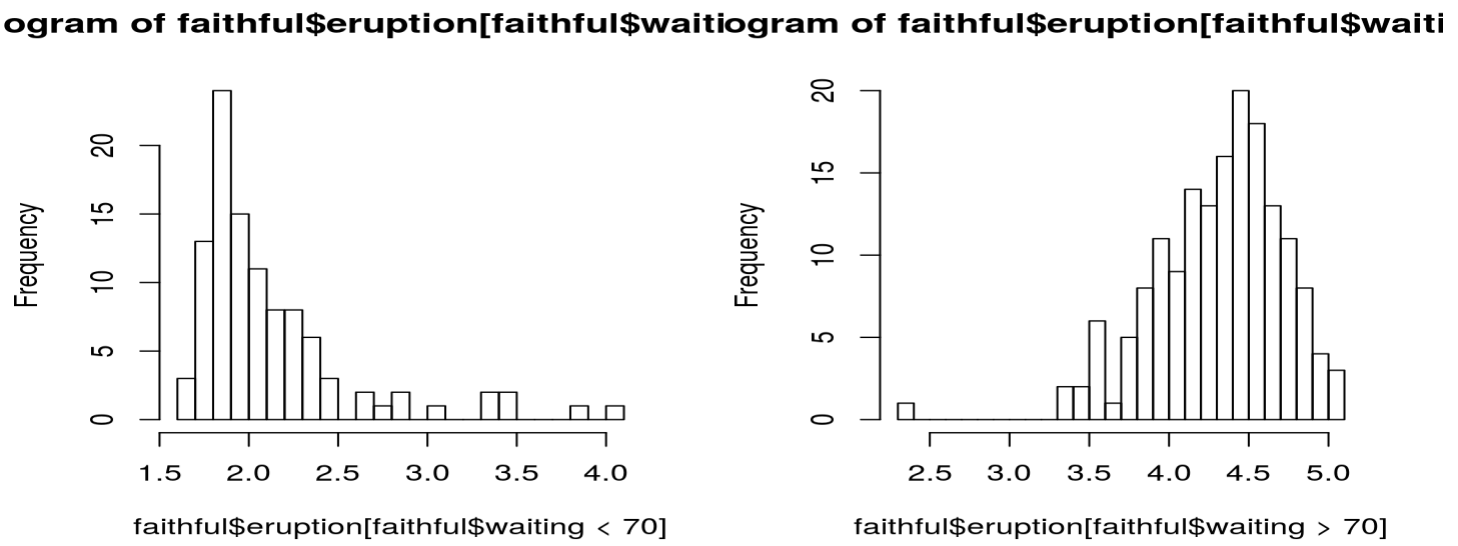
```
# Mean eruption length for longer waiting time
mean(faithful$eruptions[faithful$waiting > 70])
```

```
## [1] 4.319255
```

Old Faithful Eruption Lengths

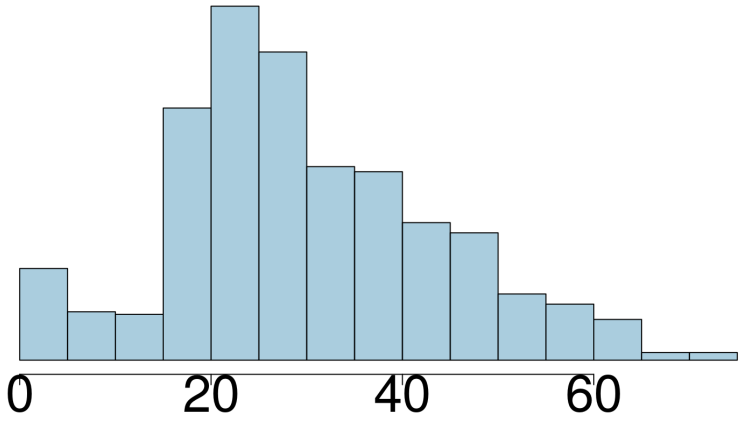
Distribution of Eruption Lengths – bimodal? multimodal?

```
par(mfrow=c(1,2))
hist(faithful$eruption[faithful$waiting < 70], breaks=25)
hist(faithful$eruption[faithful$waiting > 70], breaks=25)
```



No

What do we look for in a histogram?



Clustering: (number of **modes**)

- No bumps: “uniform”
- 1 bump: “unimodal”
- 2 bumps: “bimodal”

Center: where is the “middle”?

Spread: how much variability is there?

Skewness: lack of symmetry (e.g., one “tail” longer than the other)

Outliers: extreme values that stand out?

> Qn: How do we objectively measure the above five?

Summarizing continuous variables

Overview

- **Goal:** reduce data to a few “talking points”
- For **categorical variables** - counts/frequencies
- For **numerical variables** - more options
- Can be at population level or **sample level**

> A **statistic** is a numerical summary computed from a sample of data.

Measuring Center

Measures of central tendency

Multiple notions of *center*:

Mean - average value

Median - “midpoint” of the data when ordered from smallest to largest

Mode - most common value(s)

> Modes can be used to detect number of sub-populations as well

Other measures:

- quartiles
- quantiles / percentiles

Sample mean (in R)

```
x=c(2,1,2,8,2)
sum(x) / length(x)
```

```
## [1] 3
```

```
mean(x)
```

```
## [1] 3
```

```
mean(dat$Age, na.rm=TRUE) # calculate mean after removing missing values
```

```
## [1] 30.39799
```

```
mean(dat$Age)
```

```
## [1] NA
```

Sample Mean ("in math")

Given n values, $x_1, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

(pronounced "x bar")

or

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

x_i is the i th value in the data set.

Calculate mean from first principle

```
x = c(1, 2, 3, 5, 6, 7)
mean(x)
```

```
## [1] 4
```

How about using a `for` loop?

```
temp = 0
for (i in 1:6){
  temp = temp+x[i]
  # print(i) # see how the value of i changes within a loop
  # print(temp) # see how the value of temp changes within a loop
  # readline()  # this command pauses the output until you press ENTER
}
print(temp/6) # this is the value of mean(x)
```

```
## [1] 4
```

Sample Median (in R)

```
x=c(2,1,2,8,2)
sort(x)
```

```
## [1] 1 2 2 2 8
```

```
[1] 1 2 2 2 8
```

```
median(x)
```

```
## [1] 2
```

```
median(dat$Age, na.rm=TRUE) # calculate median after removing missing values
```

```
## [1] 28
```

Calculating median from first principle

```
age = dat$Age[!is.na(dat$Age)] # remove missing values from the column of Age
length(age) # see how many observations are left
```

```
## [1] 756
```

```
age_sort = sort(age) # sort the ages from smallest to largest
age_sort[756/2] # pick the middle value
```

```
## [1] 28
```

```
age_sort[756/2 + 1] # or the next one
```

```
## [1] 28
```

> 50% of observations in your data are smaller than the median, that is why it is also called the 50^{th} **percentile**, or 0.5 **quantile**

Sample Median, Quartiles and Percentiles

- Given n values, $x_1, \ldots, x_n$, the **sample median** is the **50th percentile**, i.e. the value such that half of the data lies below and half lies above.
- **Equal area of histogram** on either side of median.
- **first quartile (Q_1)**: 0.25 quantile or 25^{th} percentile
- **third quartile (Q_3)**: 0.75 quantile or 75^{th} percentile
- **decile**, **quintile** are defined similarly

> Median is more **robust** measure of center than mean

What if we change 8 to 80?

```
x=c(2,1,2,8,2)
mean(x)
```

```
## [1] 3
```

```
median(x)
```

```
## [1] 2
```

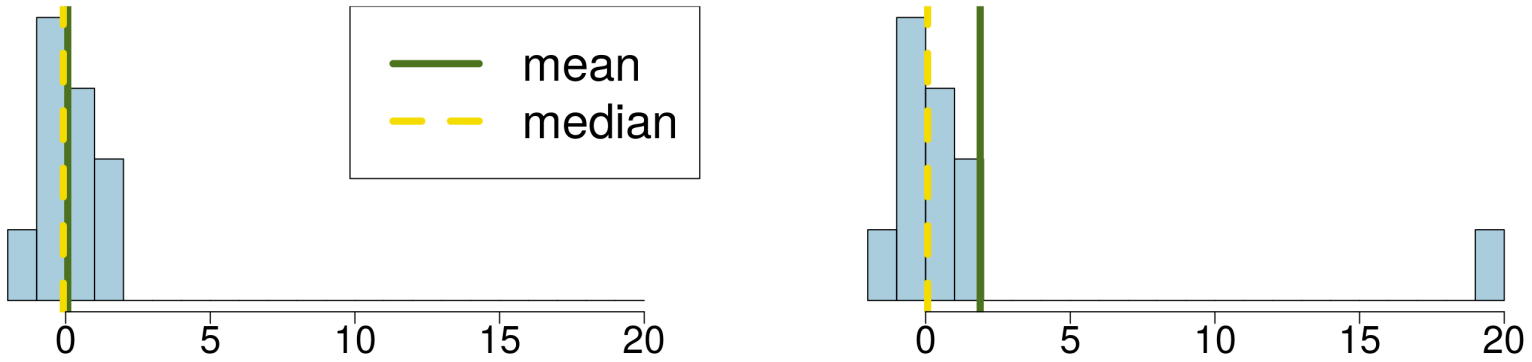
```
xx=c(2,1,2,80,2)
mean(xx)
```

```
## [1] 17.4
```

```
median(xx)
```

```
## [1] 2
```

In pictures

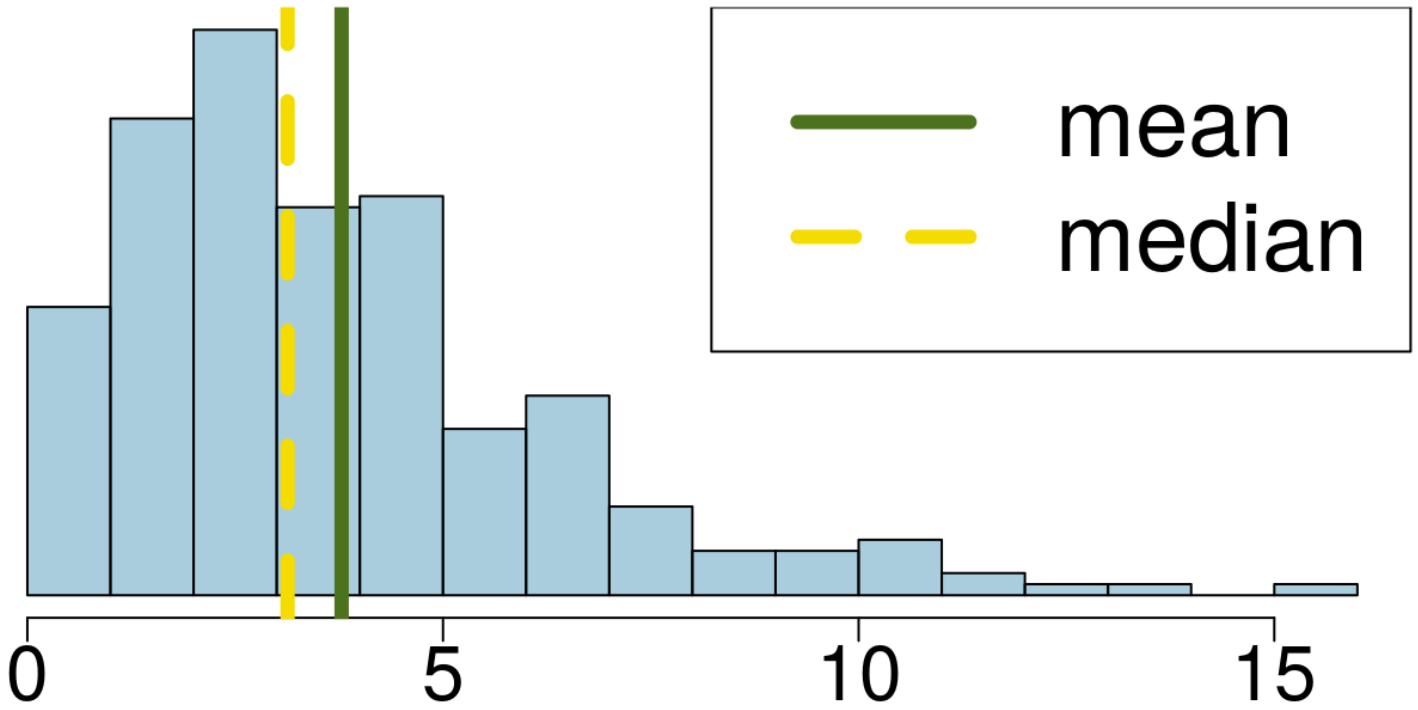


Median: Equal area

Mean: Balancing point

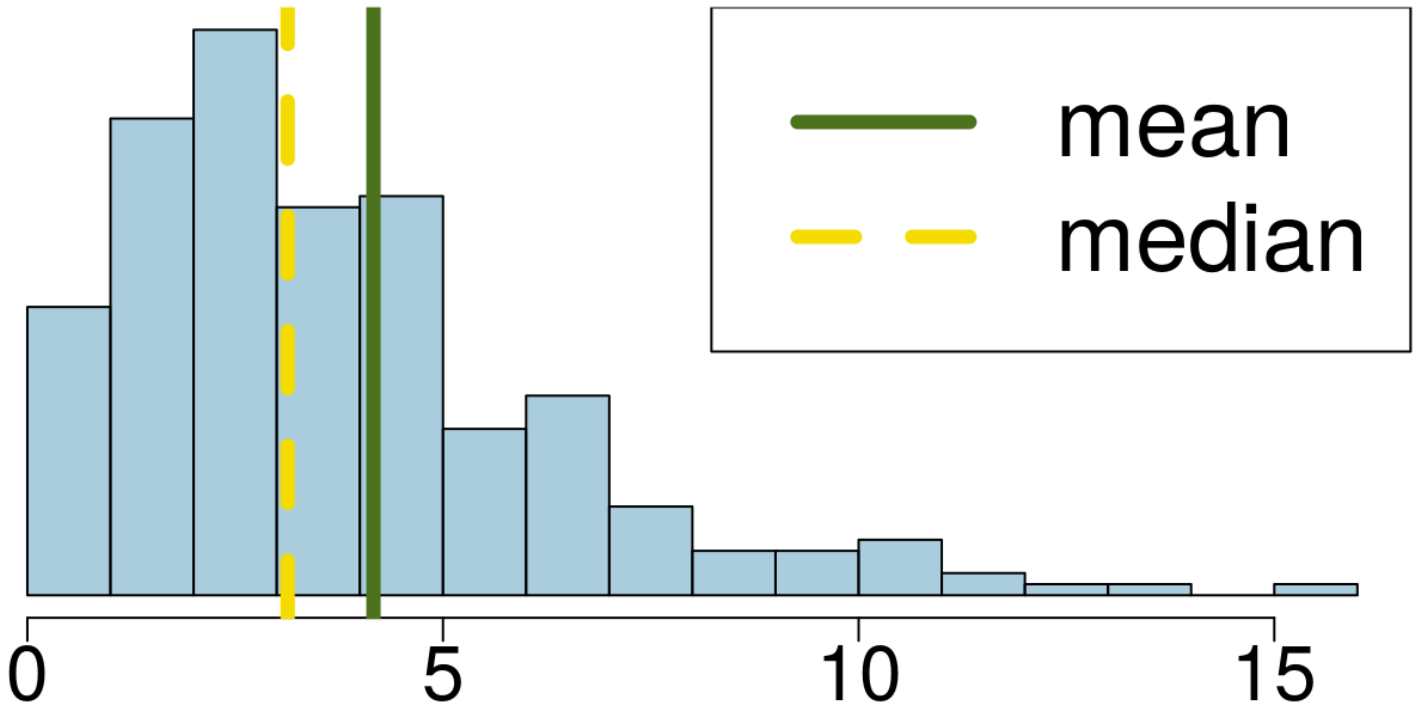
Right skewed

long right tail (250 observations total)



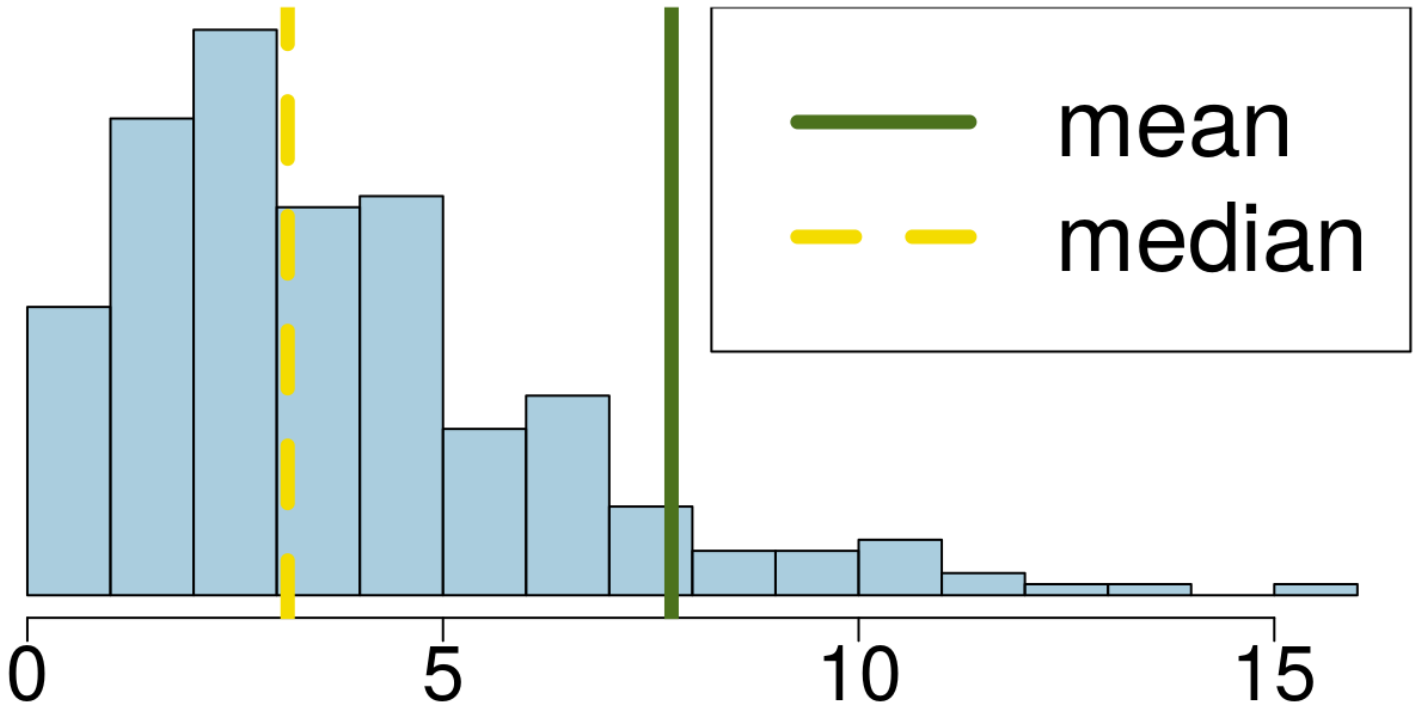
Right skewed

adding an observation of size 100…



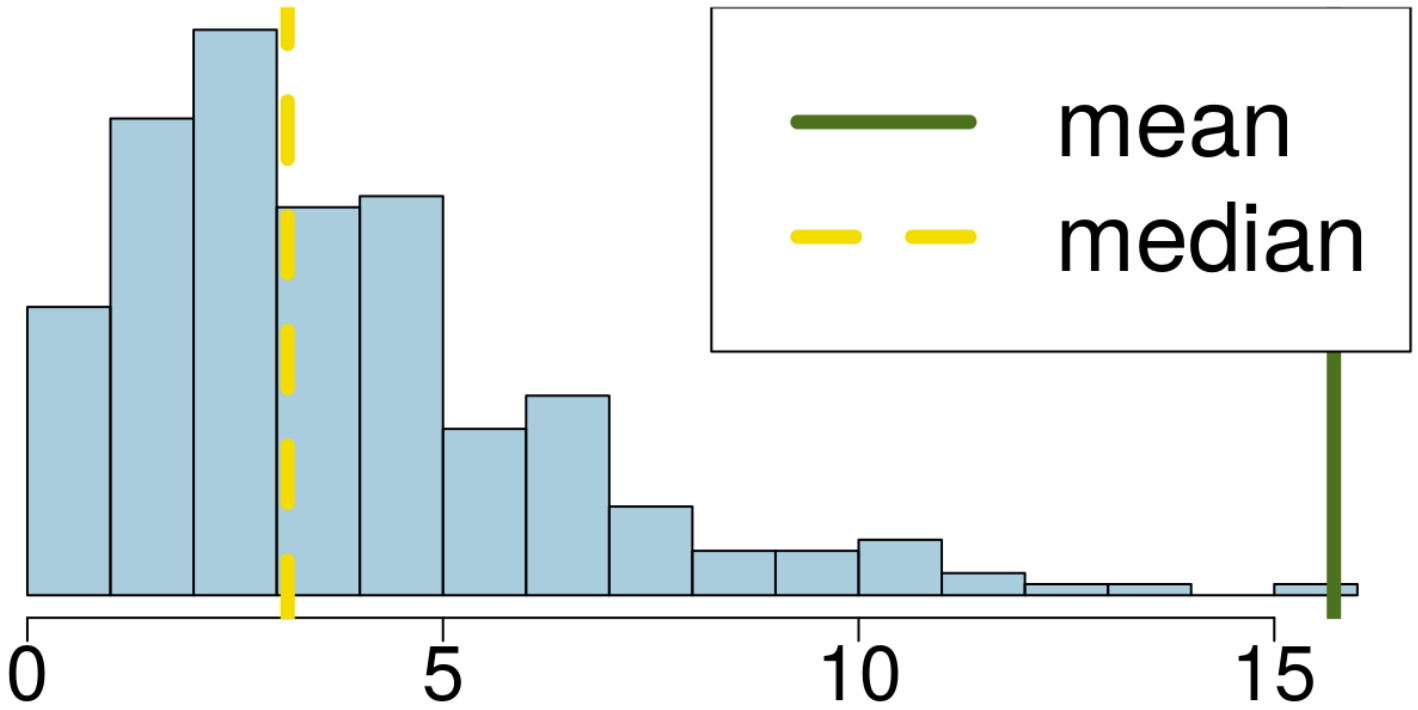
Right skewed

adding an observation of size 1000…



Right skewed

adding an observation of size 3000…

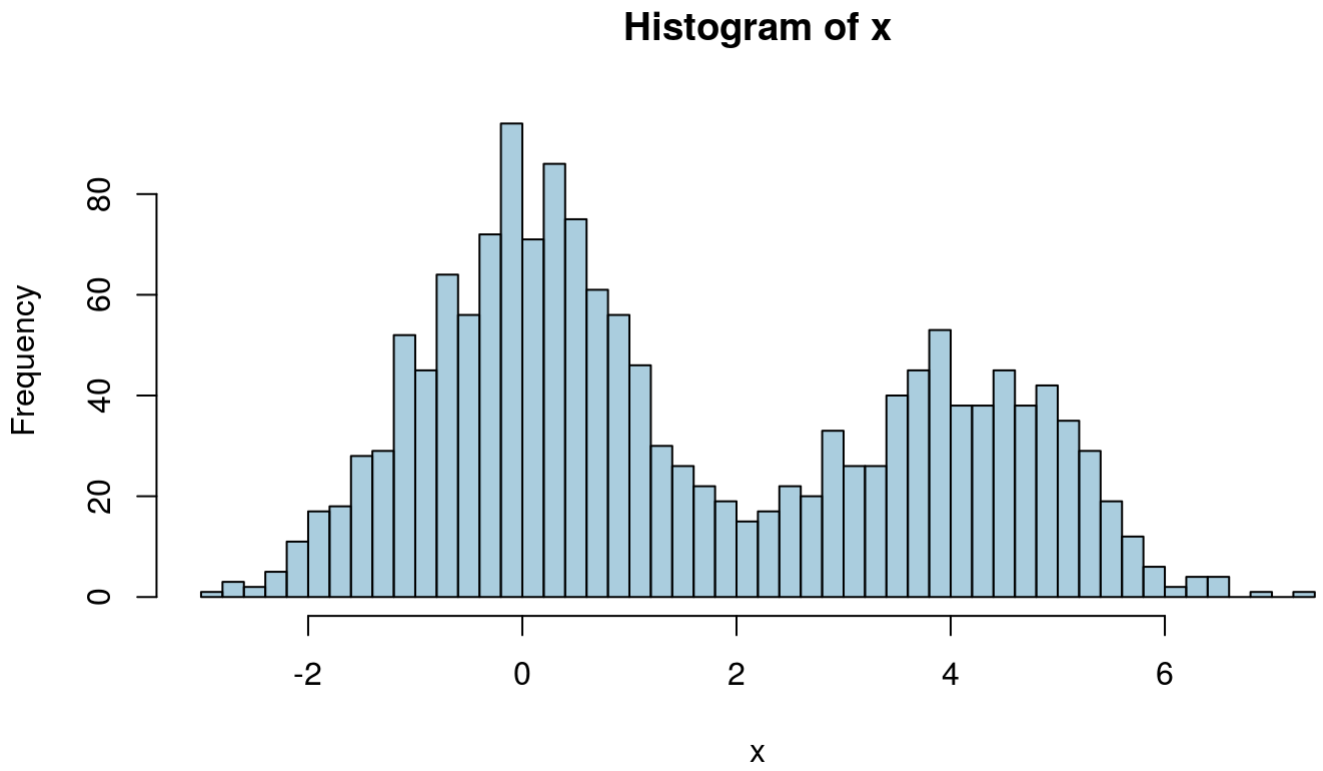


Mode

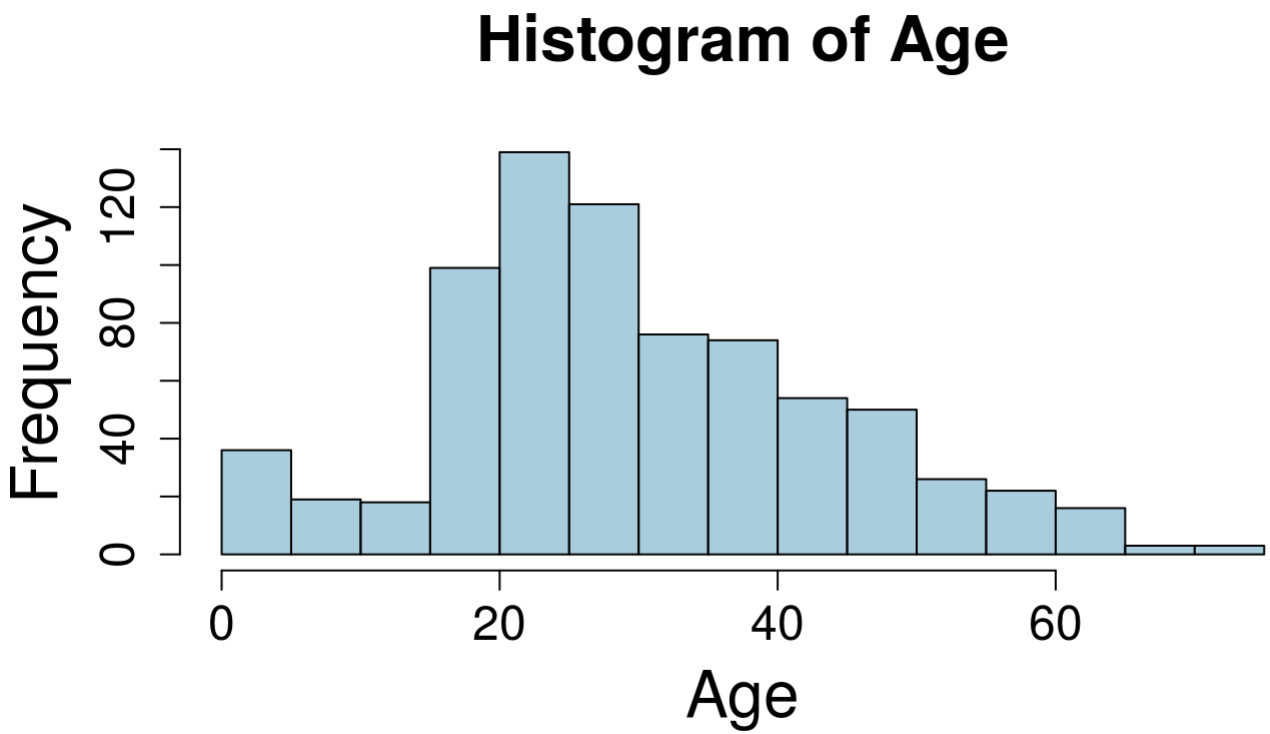
```
x=c(2,1,2,8,2)
table(x)
```

```
## x
## 1 2 8
## 1 3 1
```

In pictures



Titanic ages



Weighted Average

If a discrete random variable takes values $x_1, \ldots, x_k$ with frequencies $f_1, \ldots, f_k$, its average can be calculated as

$$\frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{f_1 + f_2 + \ldots + f_k}$$

```
head(days)
```

```
## [1] 2 3 5 7 2 7
```

```
table(days)
```

```
## days
##  1  2  3  4  5  6  7
## 16 22 20 25 25 22 20
```

```
mean(days)
```

```
## [1] 4.113333
```

```
(1*16+2*22+3*20+4*25+5*25+6*22+7*20)/150
```

measures of central tendency and loss minimization

Mean, median and quantiles are less ad hoc than you would think

Imagine you are trying to approximate a data set of numerical variables by a single number c (short for “center”)

It makes sense to find a c which minimizes a reasonable notion of “overall distance” from all the data points

measures of central tendency and loss minimization

```
x = c(1, 2, 3, 4, 5, 6, 6, 7)
```

Find a number c such that $\sum_{i=1}^{n} |x_i - c|$ is as small as possible

```
c = 1
sum(abs(x-c)) # same as |x1-c| + |x2-c| + ... + |xn - c|
```

```
## [1] 26
```

Try different values of c and see if `c=median(x)` gives you the minimum value

Try on your own: can you write a `for` loop to search over different values of c?

measures of central tendency and loss minimization

```
x = c(1, 2, 3, 4, 5, 6, 6, 7)
```

Find a number c such that $\sum_{i=1}^{n}(x_i - c)^2$ is as small as possible

```
c = 1
sum((x-c)^2) # same as (x1-c)^2 + (x2-c)^2 + ... + (xn - c)^2
```

```
## [1] 116
```

Try different values of c and see if `c=mean(x)` gives you the minimum value

Measuring Spread or Dispersion

Quantifying "spread"

Range: maximum minus minimum

rth quantile, aka *(100r)th percentile*: value such that (100r)% of values are below it and 100(1-r)% are above.

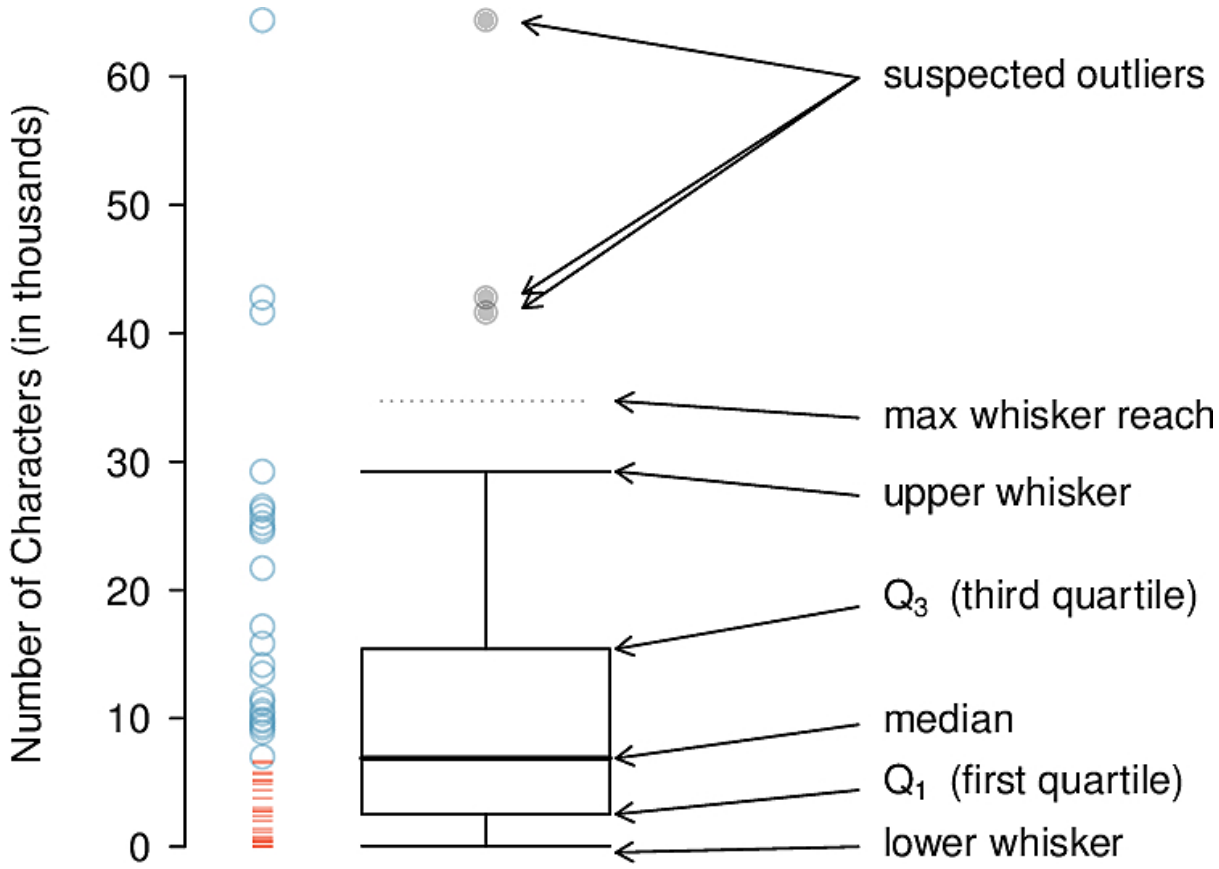
Quartiles

- 1st quartile: 25th percentile
- 2nd quartile: 50th percentile (aka…?)
- 3rd quartile: 75th percentile

Interquartile range (IQR): 3rd quartile minus 1st quartile

Variance and **standard deviation**

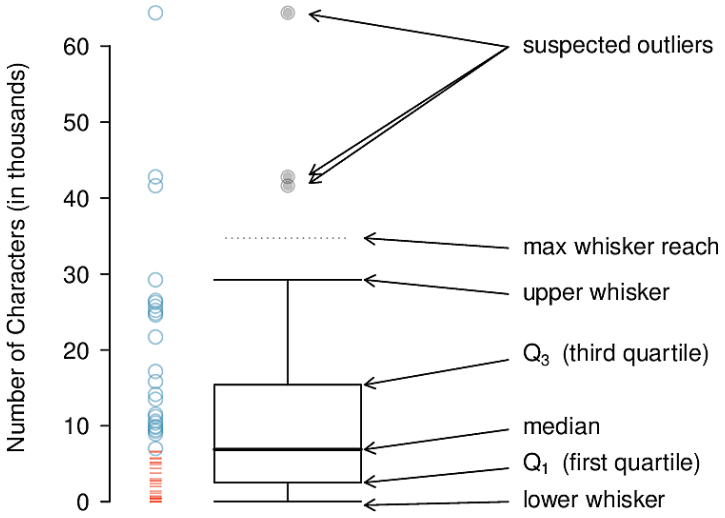
The boxplot


Number of Characters (in thousands)
60
50
40
30
20
10
0
suspected outliers
max whisker reach
upper whisker
Q_3 (third quartile)
median
Q_1 (first quartile)
lower whisker

The boxplot

Labeled dot plot and boxplot of **number of characters in 50 emails**

$n = ?$



- Upper whisker is at most $Q_3 + 1.5 \times IQR$ (but falls back to largest point below that), likewise for lower whisker.

- **Q1**: 25% of emails have at least __?__ characters
- **Q2**: 25% of emails have at most __?__ characters
- **Q3**: __?__ emails have "unusually" high number of characters
- **Q4**: Distribution of characters is **left/right(?)** skewed

John Tukey (1915-2000)

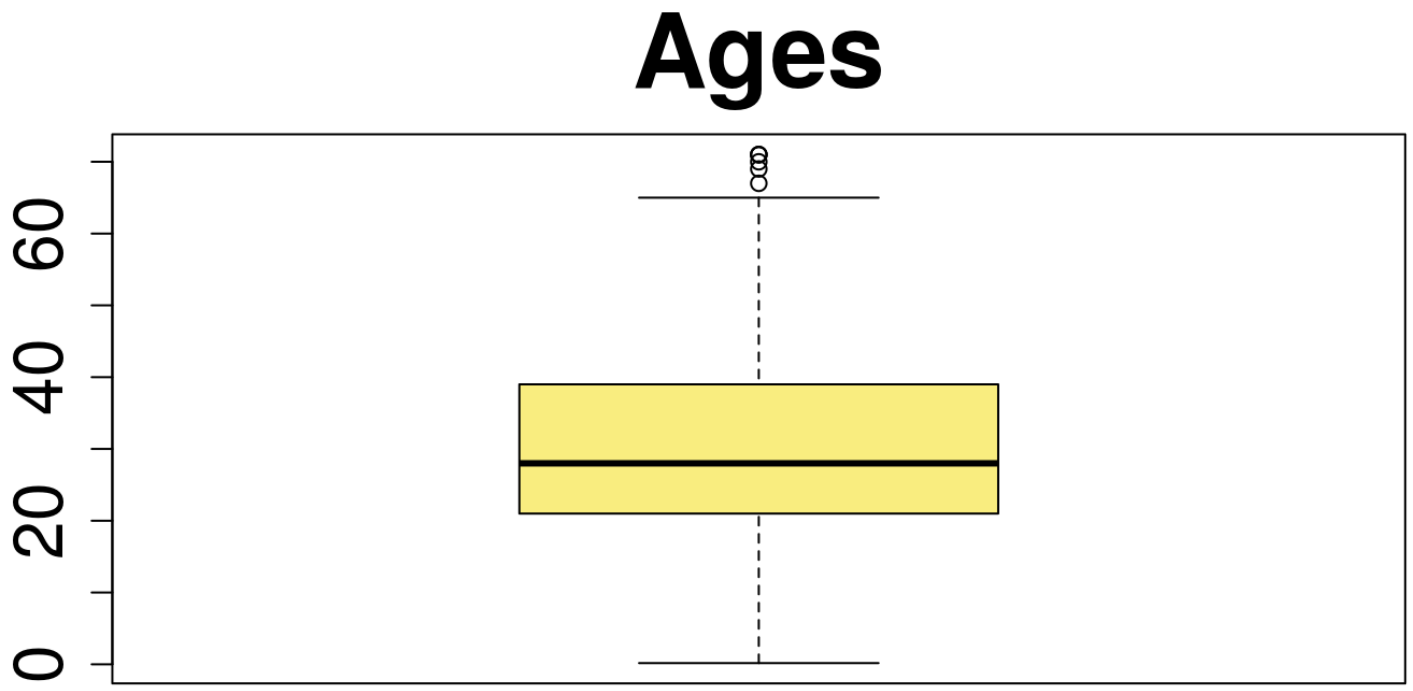
> "Numerical quantities focus on expected values, graphical summaries on unexpected values." –John Tukey

Selected contributions (many more!)

- box plot
- Fast Fourier Transform
- coined terms: "bit" and "software" (and "scagnostics")

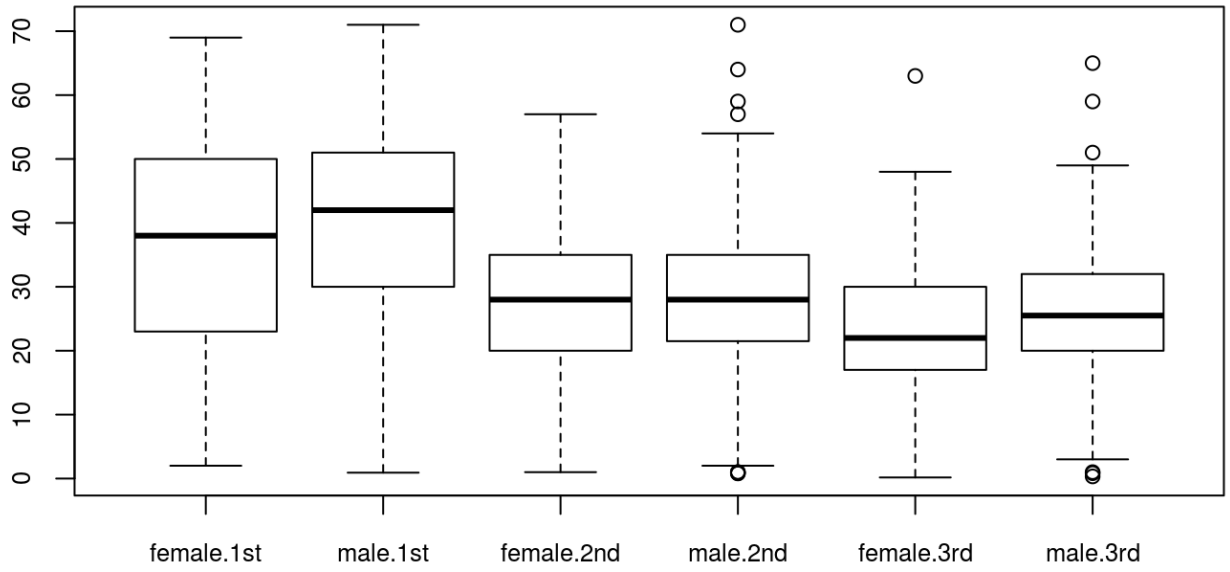
Titanic ages

```
boxplot(dat$Age, main="Ages")
```

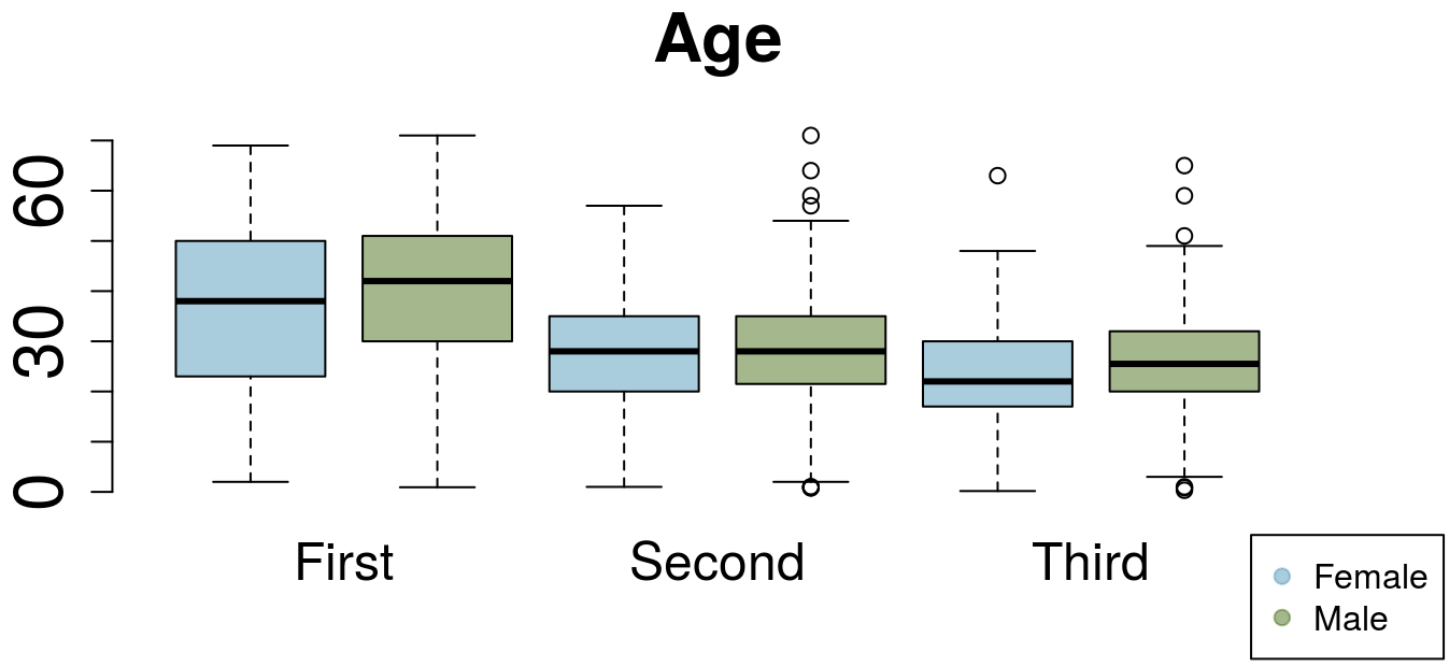


Titanic ages

```
boxplot(Age ~ Sex + PClass, data=dat)
```



Titanic ages



Variance and standard deviation

Sample variance - Measures *dispersion* of individual data points about $\bar{x}$, expressed in *squared* units of x

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

```
x=c(2,1,2,8,2)
xbar = mean(x)
n = length(x)
sum((x-xbar)^2) / (n-1)
```

```
## [1] 8
```

```
var(x)
```

```
## [1] 8
```

Variance and standard deviation

Sample standard deviation - Measures *dispersion* of individual data points about $\bar{x}$, expressed in *the same units* as x

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

```
x=c(2,1,2,8,2)
sqrt(var(x))
```

```
## [1] 2.828427
```

```
sd(x)
```

```
## [1] 2.828427
```

Calculate variance using “for” loop

```
x=c(2,1,2,8,2)
xbar = mean(x)
n = length(x)
```

```
xvar = 0
for (i in 1:n){
  xvar = xvar+(x[i]-xbar)^2
}
xvar = xvar/(n-1)
xvar
```

```
## [1] 8
```

Variance and standard deviation

- sensitive to every observation
- always non-negative
- they're zero *if and only if* all values are equal
- don't specify shape!

> Report std dev in words:

- "roughly, the average distance of observations from the mean …"

More on quantifying spread

Why $n-1$ and not n in sample standard deviation?

> Population Parameters: N units in a population, data: $x_1, x_2, \ldots, x_N$

- Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Population std dev: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$

One "degree of freedom" lost to calculate sample mean ($\bar{x}$) (on a more technical term, using $1/n$ leads to biased estimate of σ)

More on quantifying spread

Why not just use $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$?

- This is also used in practice, known as **mean absolute deviation (MAD)**
- Not differentiable, so hard to do calculus or analytical work

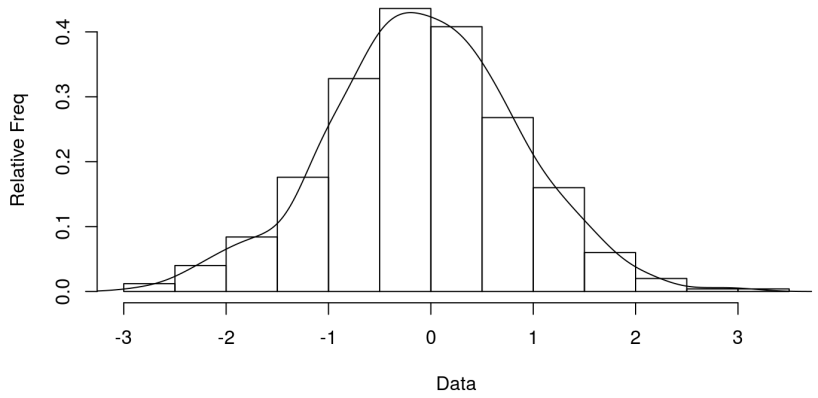
Percentiles and z-scores

Bell-shaped Distribution

Unimodal, reasonably symmetric data distributions are fairly common

Normal (Gaussian / bell-shaped) distributions often used as an approximation

Bell-shaped distributions are completely specified by mean and std dev



Empirical Rule

For bell-shaped distributions, roughly

- 68% of data are within ± 1 std dev of mean
- 95% of data are within ± 2 std dev of mean
- 99.7% of data are within ± 3 std dev of mean

Example

Prelim 1 results are out. The histogram of scores looks roughly bell-shaped, with mean 60 and standard deviation 5.

You scored 65. What percentage of students scored more than you?

What is your percentile score?

About 95% of students scored between ___?___ and ___?___

z-score

Alice scored 70 in prelim 1. How many std dev above/below mean is her score?

Standardized score (or z-score): $z = \frac{observed\ value - mean}{stddev}$

Measures how far an individual value falls from mean, in the unit of std deviation

Standardization helps (e.g., compare student scores across sections)

Common thumbrule to detect extreme observations (z-score beyond -3 or 3)

Empirical Rule (in terms of z-scores)

For bell-shaped distributions, roughly

- 68% of data have z-scores between -1 and 1
- 95% of data have z-scores between -2 and 2
- 99.7% of data have z-scores between -3 and 3

Why not use the actual percentile calculated from data, rather than bell-curve approximation?

- easier to calculate, commonly used in practice

Skewness

Measure of Skewness

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{(x_i-\mu)}{\sigma}\right)^3$$

Skewness in R

```
library(e1071)
x = c(1,2,2,3,3,3,4,4,5)
# hist(x, breaks = 0:5)
skewness(x)
```

```
## [1] 0
```

```
x = c(1,2,2,3,3,3,4,4,10)
skewness(x)
```

```
## [1] 1.53069
```

```
x = c(-10,2,2,3,3,3,4,4,5)
skewness(x)
```

```
## [1] -1.890539
```

Relationships between variables

Graphical displays

Categorical vs. Categorical - stacked/unstacked bar charts

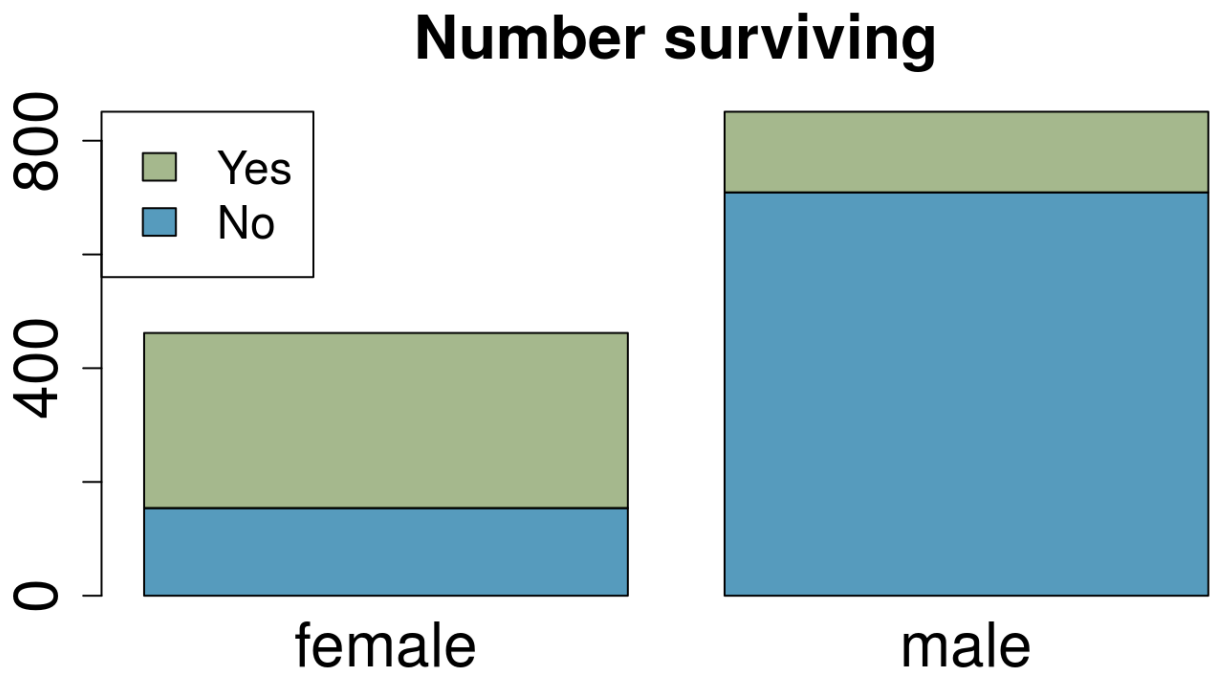
Numerical vs. Categorical - side-by-side boxplots

Numerical vs. Numerical - scatterplots

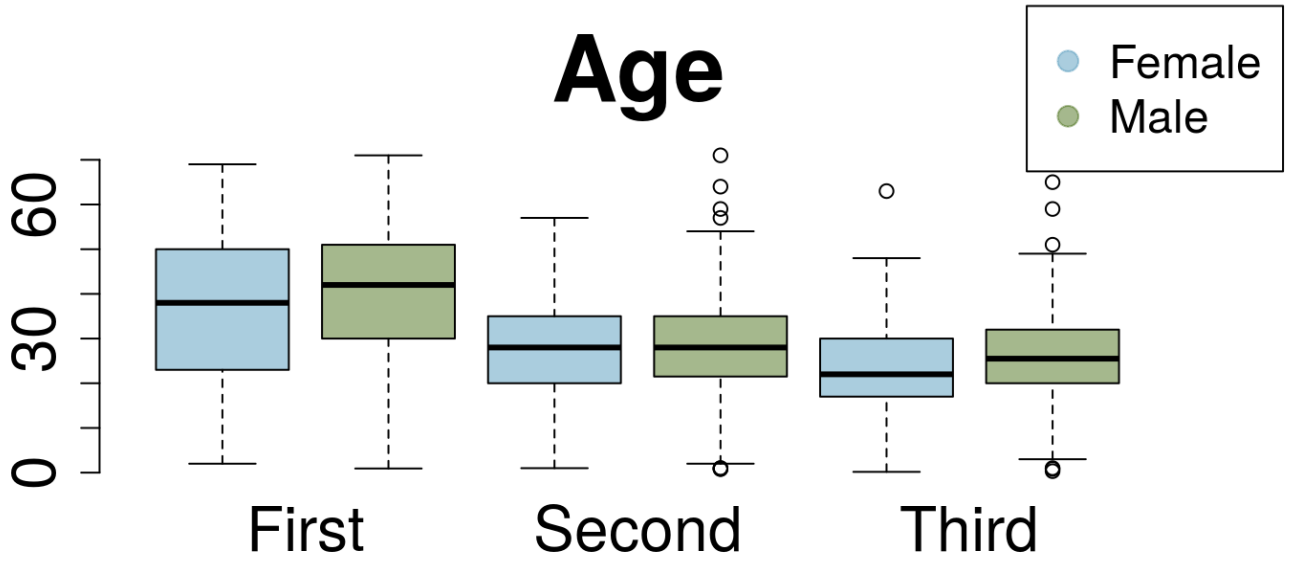
Numerical vs. Numerical vs. Categorical

- coded scatterplots

Categorical vs. Categorical



Numerical vs. Categorical



Canadian Prestige Data (1971)

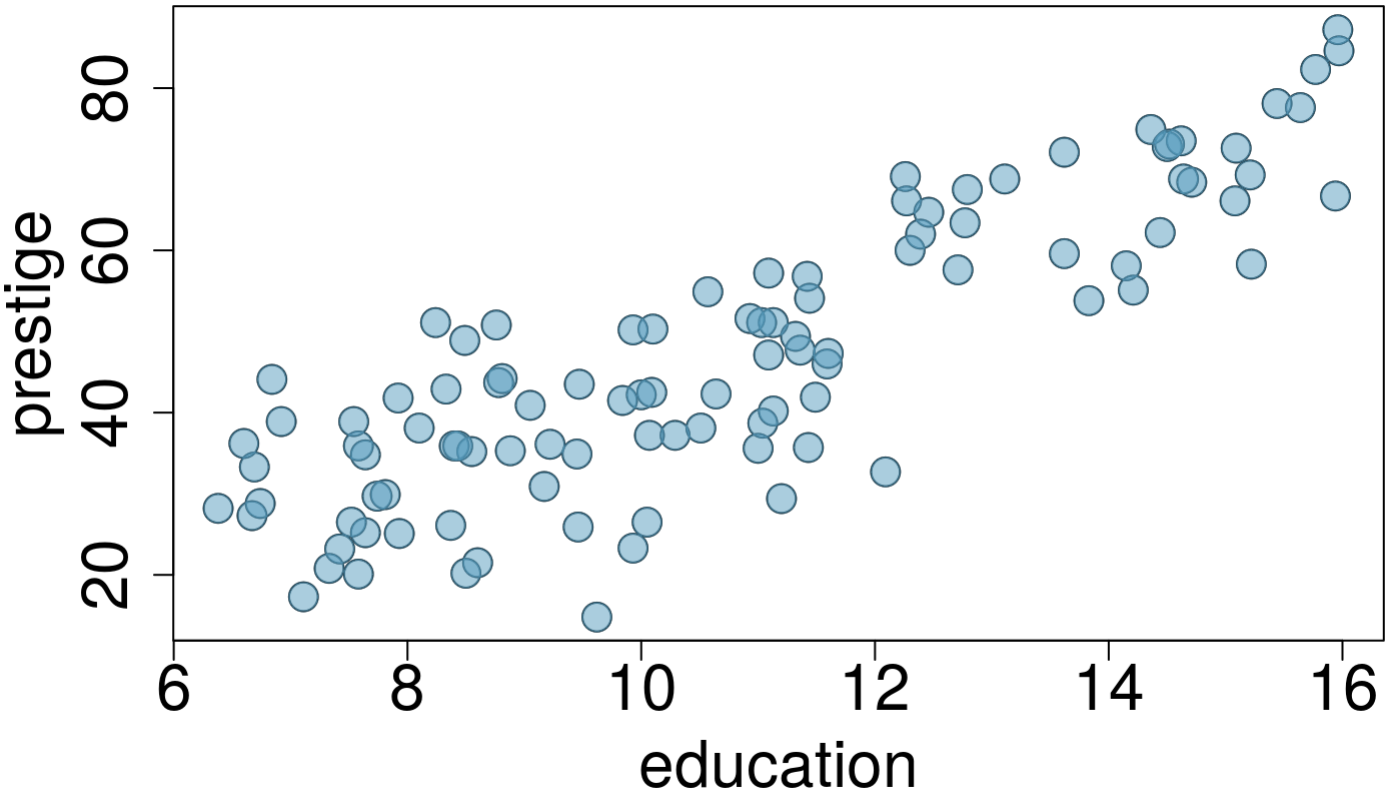
- ~100 occupations
- **education** - avg number of yrs of education beyond grade 4
- **income**
- **women** - % women in occupation
- **prestige** - “Pineo-Porter” score
- **type** - blue collar, white collar, professional

```
library(car)
View(Prestige)
```

Canadian Prestige Data (1971)

	education	income	women	prestige	census	type
physio.therapsts	13.62	5092	82.66	72.1	3137	prof
pharmacists	15.21	10432	24.71	69.3	3151	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
mining.engineers	14.64	11023	0.94	68.8	2153	prof
gov.administrators	13.11	12351	11.16	68.8	1113	prof
osteopaths.chiropractors	14.71	17498	6.91	68.4	3117	prof
medical.technicians	12.79	5180	76.04	67.5	3156	wc
veterinarians	15.94	14558	4.32	66.7	3115	prof
secondary.school.teachers	15.08	8034	46.80	66.1	2733	prof
pilots	12.27	14032	0.58	66.1	9111	prof
nurses	12.46	4614	96.12	64.7	3131	prof
accountants	12.77	9271	15.70	63.4	1171	prof
economists	14.44	8049	57.31	62.2	2311	prof

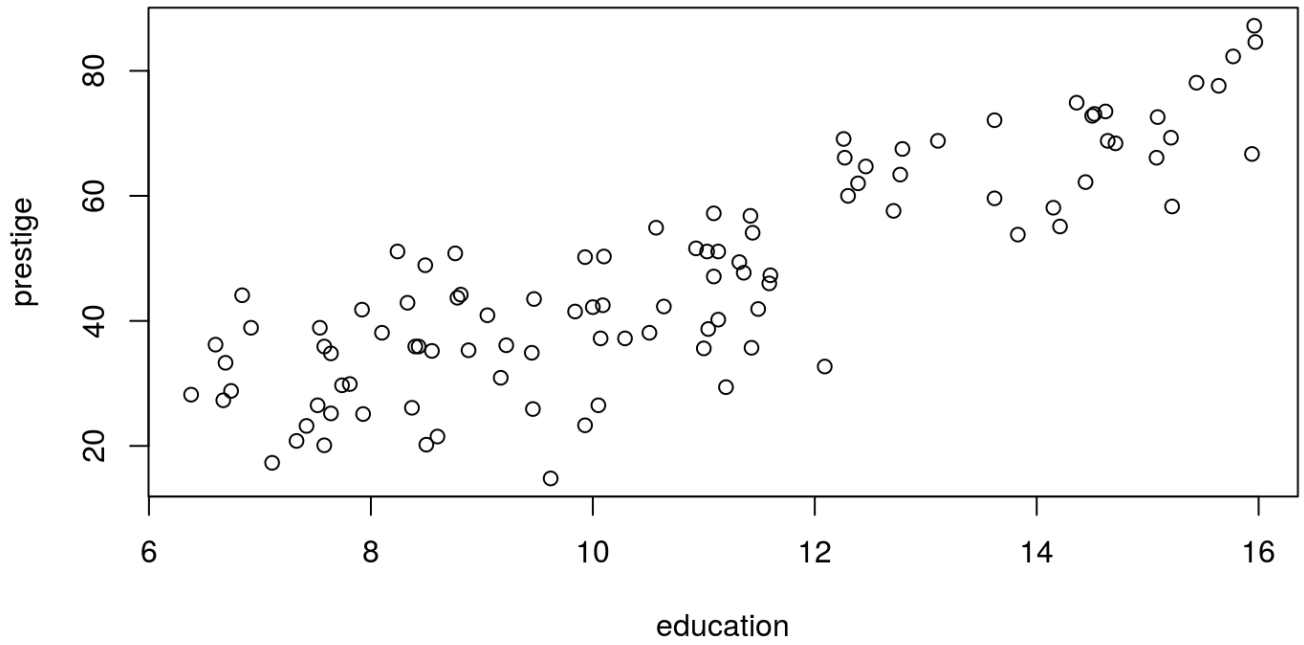
Numerical vs. Numerical



Numerical vs. Numerical

make a scatterplot…

```
plot(prestige~education, data=Prestige)
```



The scatterplot

- best way to visualize **relationship between two quantitative variables**

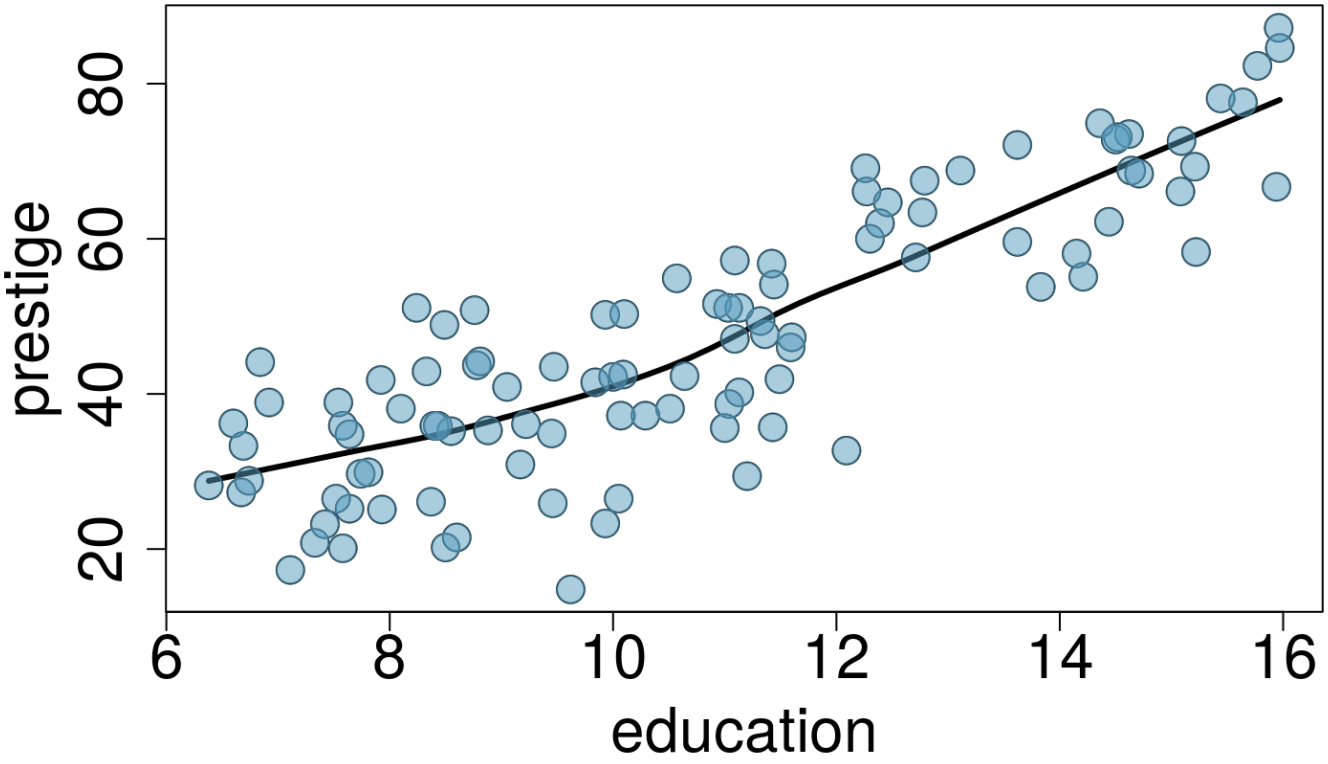
x-axis

- **predictor**
- “explanatory variable”
- “independent variable”

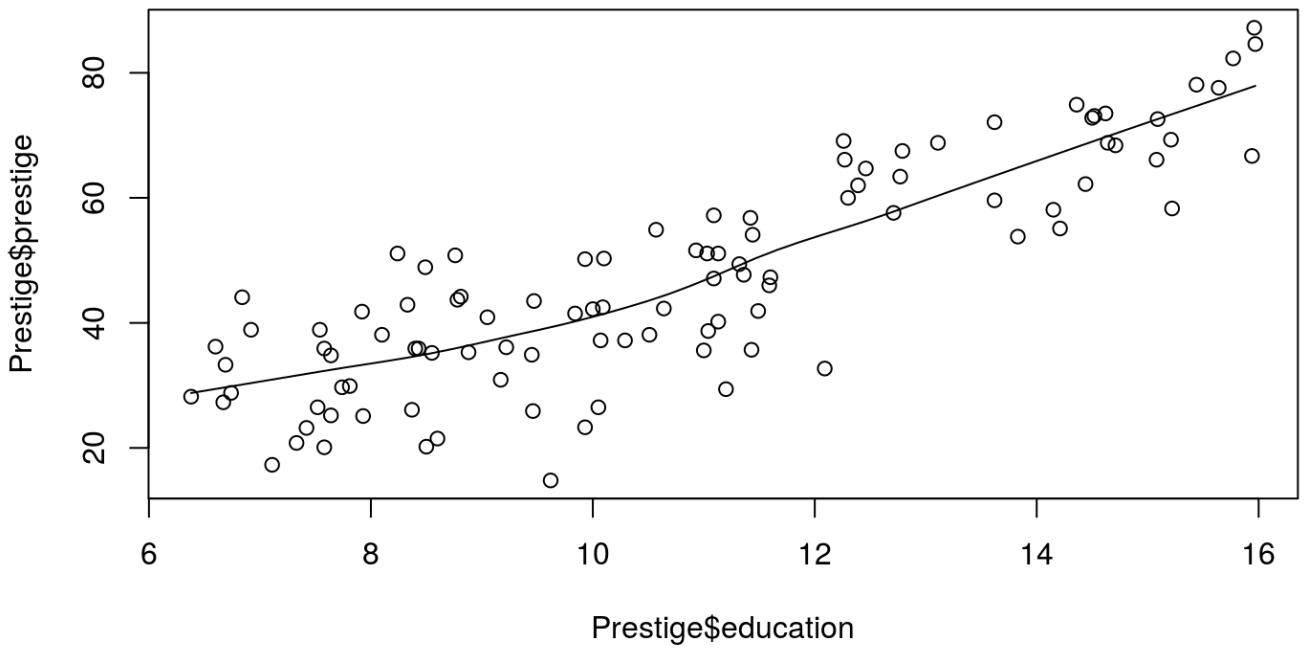
y-axis

- **response**
- “dependent variable”

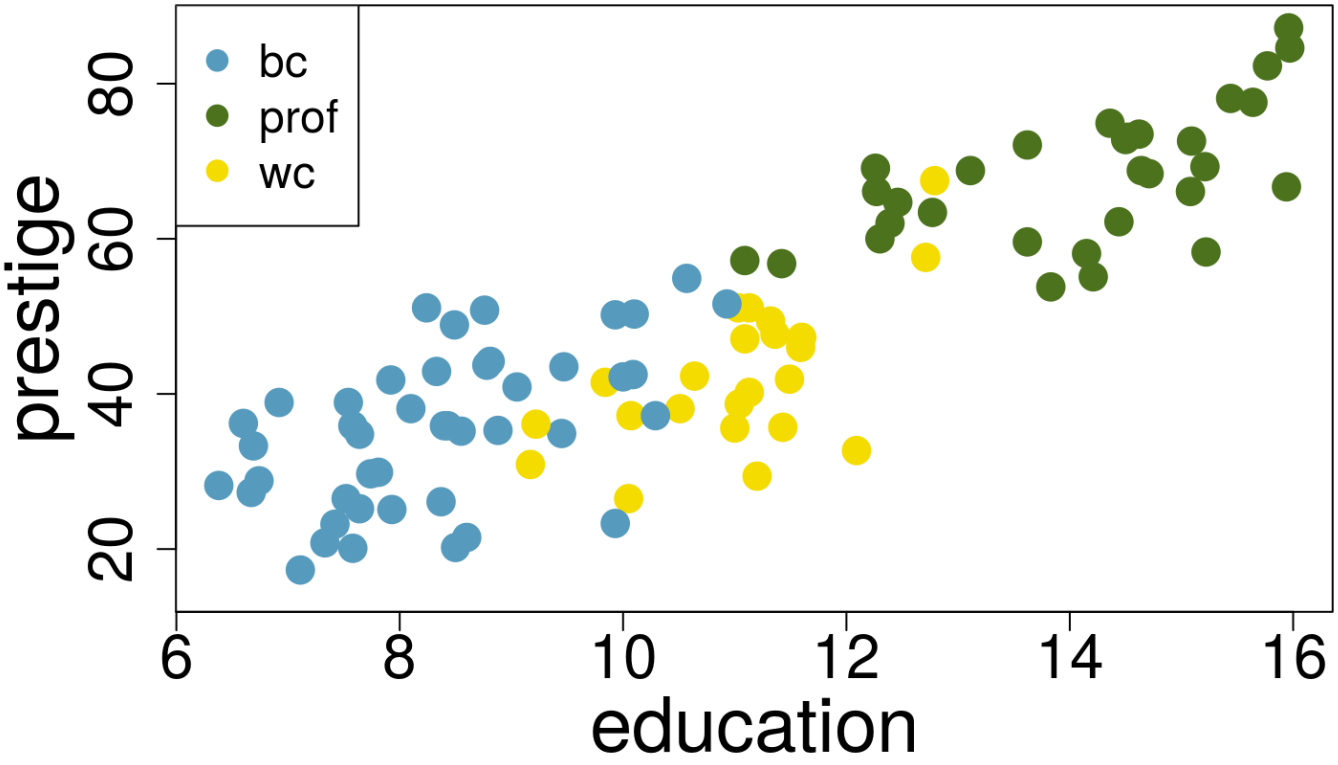
Adding a smoother



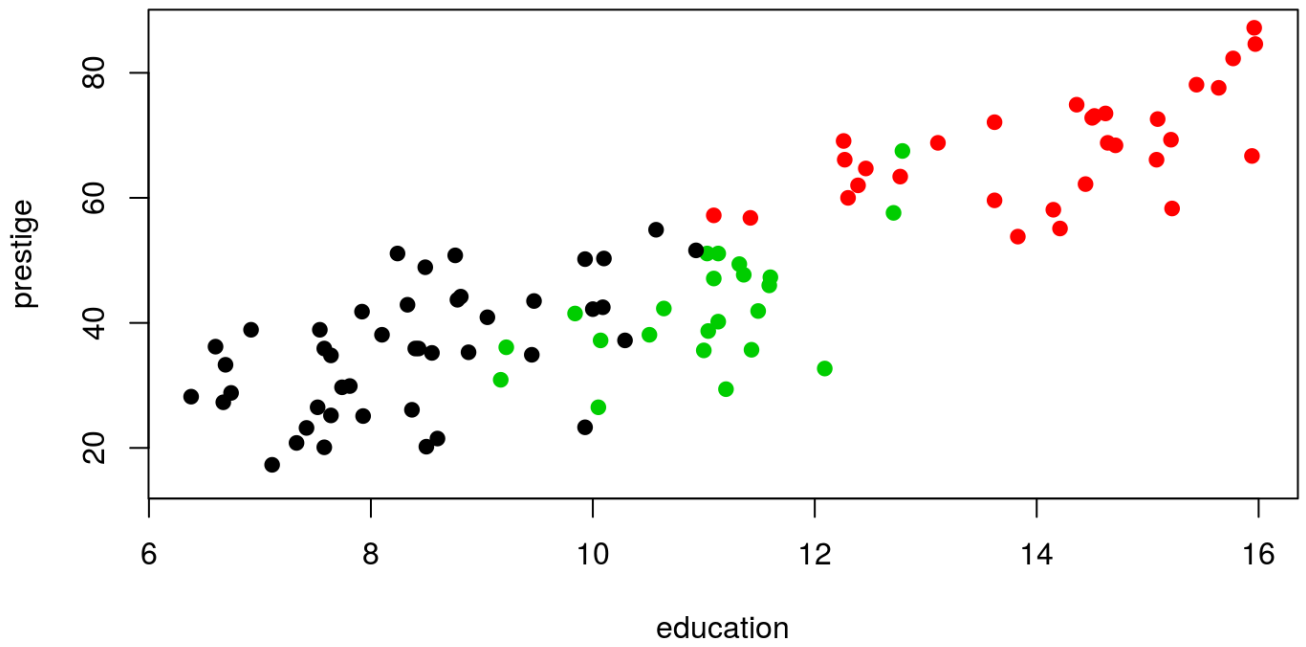
Adding a smoother



Numerical vs. Numerical vs. Categorical

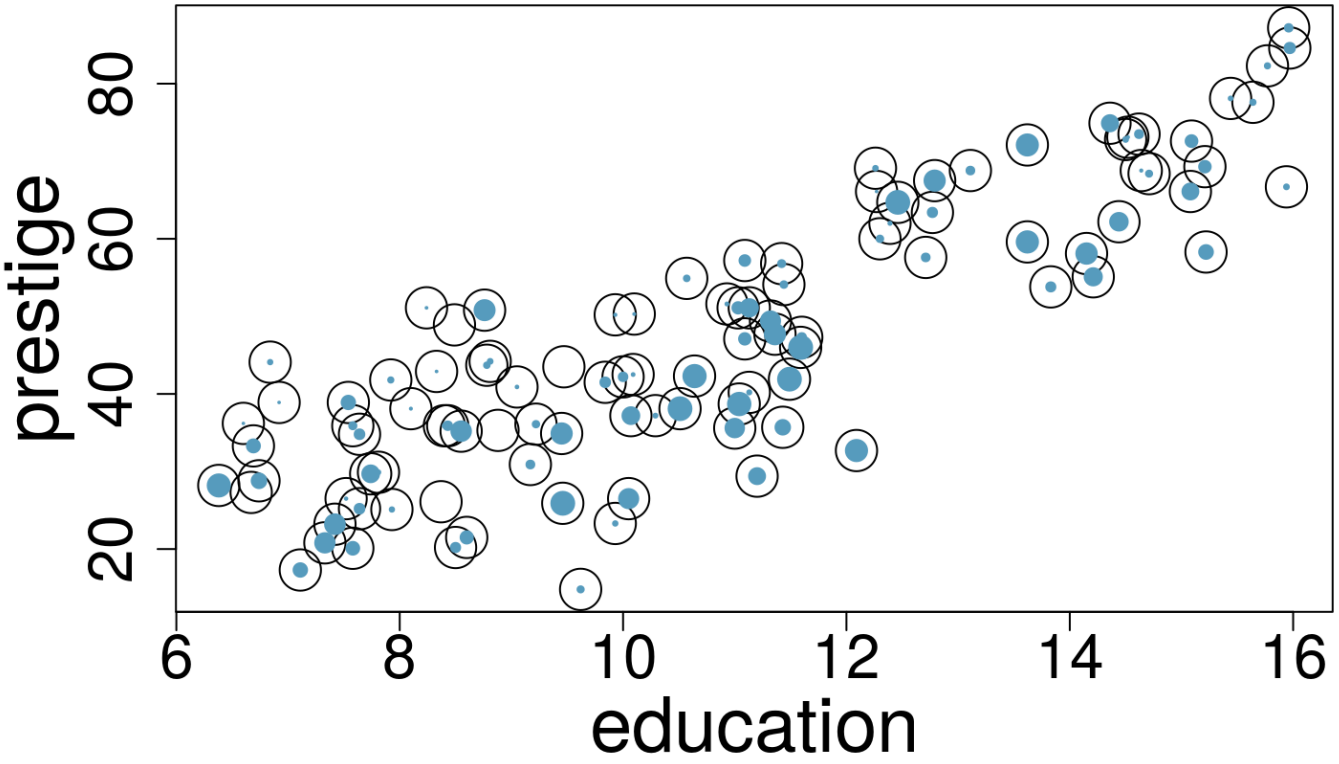


Numerical vs. Numerical vs. Categorical



Numerical vs. Numerical vs. Numerical

Color shows proportion of women in occupation



Visualize all pairs at once

```
pairs(Prestige)
```