Not sure which Statistics course to take first?

STAT 220, STAT 234, or the sequence STAT 244-245?

See a comparison of STAT 220 and 234:
statistics.uchicago.edu/~collins/introSTAT

Read a discussion of all of these courses:
collegecatalog.uchicago.edu/thecollege/statistics/#generalcourseinformation

…and feel free to drop by to talk with Dr. Linda Collins in Jones 205
(not available Tue/Thu afternoons)

Course prerequisites:
STAT 220: MATH 131 (or placement into MATH 151)
STAT 234: MATH 133, 153, or 162 (single-variable calculus)
STAT 244: MATH 195, 200, or analysis (multi-variable calculus)
STAT 245: STAT 244 + linear algebra (STAT 243, MATH 20250 or higher)

Course website:

`https://canvas.uchicago.edu/courses/32398`

Statistics Terminology

Like any field of inquiry,
statistics assigns very specific meaning to some everyday words.

- sample (data), statistic
- population, parameter
- dataset: case, label, variable, value
- variable: quantitative, categorical
- distribution: variance, skew

Example: Wealth Distribution I

```
Rows: 19,674
Columns: 5
$ Y1984 <dbl> -169, 79470, 28660, NA, 5800, NA, NA, NA,...
$ Y1989 <dbl> NA, 150500, 12300, NA, NA, NA, 8685, NA, ...
$ Y1994 <dbl> NA, 145000, 3500, NA, 12000, NA, NA, NA, ...
$ Y2001 <dbl> NA, NA, 113050, 493000, NA, 41000, 148000...
$ Y2017 <dbl> NA, NA, NA, NA, 2000, 83000, NA, 0, NA, 1...
```

This data is from a study conducted by the Institute for Social Research at the University of Michigan, following around 15000 individuals over several decades in terms of a number of income and wealth variables. We are only showing total wealth for several years.

`https://psidonline.isr.umich.edu/`

Terms: popn vs. sample, cases vs. labels, variables vs. values

Variables: quantitative vs. categorical

Example: Wealth Distribution II

What is the distribution of Wealth in 2017?

A sample (or population) **distribution of a variable** has two parts:

1. the set of values observed in the sample
   (or all values possible to observe in the population)
2. the relative frequency of occurrence for those values

Example: Wealth Distribution III

- A **categorical** variable places each case into one of several groups, or categories.

- A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

- The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

Example: Wealth Distribution IV

```r
head(Wealth)
```

```
# A tibble: 5 x 5
  Y1984  Y1989  Y1994  Y2001 Y2017
  <dbl>  <dbl>  <dbl>  <dbl> <dbl>
1  -169     NA     NA     NA    NA
2 79470 150500 145000     NA    NA
3 28660  12300   3500 113050    NA
4    NA     NA     NA 493000    NA
5  5800     NA  12000     NA  2000
```

```r
tail(Wealth)
```

```
# A tibble: 5 x 5
  Y1984 Y1989  Y1994 Y2001  Y2017
  <dbl> <dbl>  <dbl> <dbl>  <dbl>
1    NA  -500 172000 95000 136001
2    NA    NA      0    NA     NA
3    NA    NA     NA    NA   1000
4    NA    NA     NA    NA 149000
5    NA    NA     NA    NA  81000
```

Example: Wealth Distribution V

Why all the NA's?

Do these data constitute a sample or a population?

Example: Wealth Distribution VI

What is the distribution of family wealth in 2017?

Qualitatively describing the distribution of a quantitative variable: center, spread, and shape

```
summary(Wealth$Y2017)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-2320000      285    28000   240328   162000 45745000 
    NA's 
   10067 
```

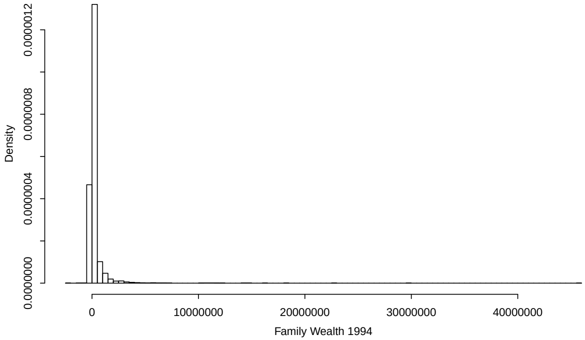
Median$=x$: 50% of data is less than x.

Q1$=x$: 25% of the data is less than x.

p-th percentile$=x$: p-proportion of the data less than x.

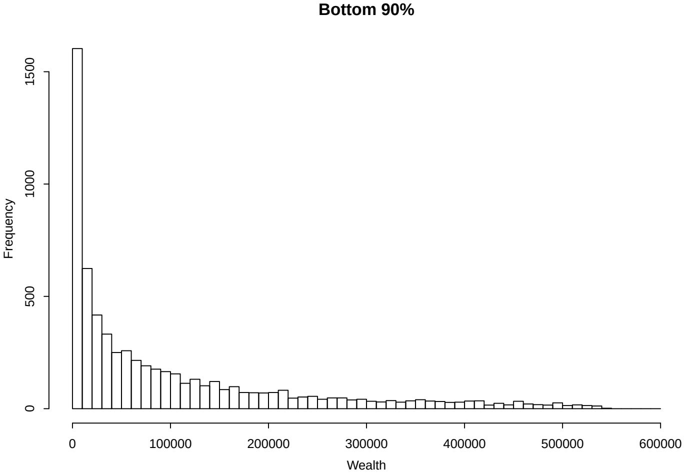
Example: Wealth Distribution VII

```
hist(Wealth$Y2017, freq=FALSE,
        xlab="Family Wealth 1994", breaks=100, main="")
```

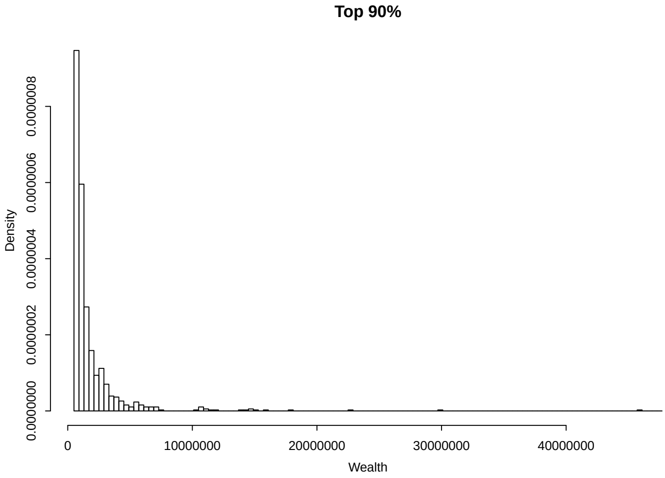


Example: Wealth Distribution VIII

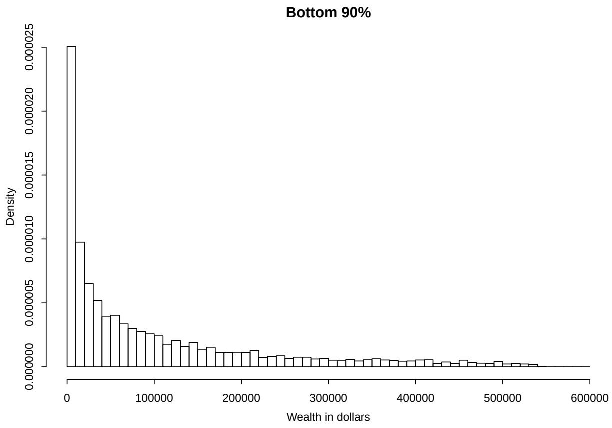
Distribution of wealth for bottom 90% and top 10%

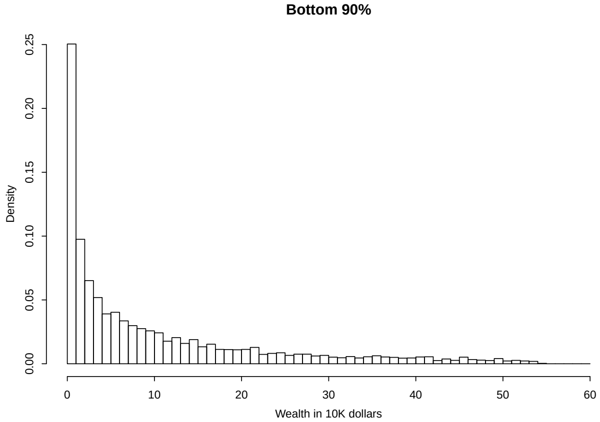


Example: Wealth Distribution IX



Density histogram: Density of bar $\times$ width $=$ percentage




Bottom 90%
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50
60
Wealth in 10K dollars

Software Installation: RStudio (and R) I

RStudio = the work environment

R = the engine (a statistical programming language)

To use the R code suggested for homework, you should install the moisaic package **once** at the start of the quarter.

```
install.packages("mosaic", ...and other packages)
```

Then, **every** time you start RStudio, type

```
require(mosaic)
```

Example: Bicycle weight and commuting time I

```
glimpse(myBikeCommute)
```

```
Rows: 56
Columns: 7
$ Bike     <fct> Steel, Carbon, Steel, Carbon, Carbon, ...
$ Date     <fct> 20/01/10, 21/01/10, 25/01/10, 26/01/10...
$ Distance <dbl> 27.20, 27.46, 27.20, 27.52, 27.51, 27....
$ Minutes  <dbl> 115.1, 115.6, 115.8, 113.9, 119.2, 108...
$ AvgSpeed <dbl> 14.10, 14.25, 14.10, 14.49, 13.84, 14....
$ TopSpeed <dbl> 31.50, 30.64, 30.92, 33.02, 30.92, 32....
$ Month    <fct> 1Jan, 1Jan, 1Jan, 1Jan, 2Feb, 2Feb, 2F...
```

Thanks to Dr. Jeremy Groves for providing his personal data.

`http://www.bmj.com/content/341/bmj.c6801` Groves, J. Bicycle weight and commuting time: randomised trial, *British Medical Journal*, BMJ 2010;341:c6801.

Example: Bicycle weight and commuting time II

```
head(myBikeCommute)
```

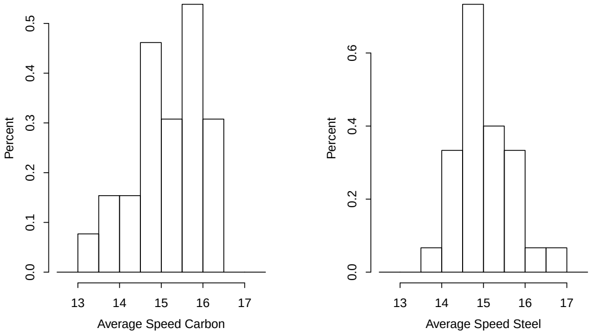
	Bike	Date	Distance	Minutes	AvgSpeed	TopSpeed	Month
1	Steel	20/01/10	27.20	115.1	14.10	31.50	1Jan
2	Carbon	21/01/10	27.46	115.6	14.25	30.64	1Jan
3	Steel	25/01/10	27.20	115.8	14.10	30.92	1Jan
4	Carbon	26/01/10	27.52	113.9	14.49	33.02	1Jan
5	Carbon	27/01/10	27.51	119.2	13.84	30.92	2Feb
6	Steel	01/02/10	27.17	108.7	14.99	32.09	2Feb
7	Steel	03/02/10	27.16	117.7	13.84	32.09	2Feb
8	Carbon	03/02/10	27.49	123.3	13.37	29.58	2Feb
9	Carbon	08/02/10	27.48	112.5	14.65	34.02	2Feb
10	Steel	09/02/10	27.09	112.6	14.43	32.71	2Feb
11	Carbon	11/02/10	27.44	117.7	13.99	32.00	3Mar
12	Carbon	01/03/10	27.49	108.6	15.18	32.71	3Mar
13	Carbon	03/03/10	27.49	110.9	14.82	34.71	3Mar

Why not alternating Steel, Carbon, Steel, Carbon, Steel, etc.?

Terms: popn vs. sample, cases vs. labels, variables vs. values

Variables: quantitative vs. categorical

Example: Bicycle weight and commuting time III



Compare speed distributions: center, spread, shape

Steel: same average?, less spread, right skewed

Carbon: same average?, more spread, left skewed

Summarizing a distribution with a center: Average I

```
by(myBikeCommute$AvgSpeed,myBikeCommute$Bike,mean)
```

```
myBikeCommute$Bike: Carbon
[1] 15.19
----------------------------------------------------
myBikeCommute$Bike: Steel
[1] 15.04
```

Average of average speed is about the same for both frame types.

```
mean(myBikeCommute$Distance)
```

```
[1] 27.16
```

The average distance is close to claimed distance: 27 miles

Summarizing a distribution with a center: Average II

Definition:

sample average $= \overline{x} =$ "x-bar" $= \dfrac{1}{n}\sum_{i=1}^{n} x_i$

$n =$ sample size

Reminder: The average is the balancing point of the data

For **any** sample of size n, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

Start on the left side: $\sum_{i=1}^{n}(x_i - \overline{x})$

Summarizing a distribution with a center: Average III

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \text{rewritten expression}$$
$$= \text{rewritten again}$$
$$= \text{and rewritten again}$$
$$= \text{until arriving at the right side} = 0$$
$$\sum_{i-1}^{n} x_i - n\bar{x} = \ldots$$

Summarizing a distribution with a center: Median

Median: Splits the data in half, half above half below.

```
by(myBikeCommute$Distance,myBikeCommute$Bike,median)

myBikeCommute$Bike: Carbon
[1] 27.38
---------------------------------------------
myBikeCommute$Bike: Steel
[1] 27.01

median(myBikeCommute$Distance)

[1] 27.19
```

The median distance is close to claimed distance: 27 miles

```
sort(myBikeCommute$Distance)

 [1] 25.86 26.60 26.74 26.88 26.90 26.91 26.91 26.91 26.92
[10] 26.94 26.94 26.94 26.95 26.99 27.00 27.00 27.01 27.02
[19] 27.02 27.03 27.03 27.05 27.06 27.09 27.10 27.16 27.16
[28] 27.17 27.20 27.20 27.27 27.29 27.31 27.31 27.32 27.32
[37] 27.33 27.34 27.34 27.36 27.36 27.38 27.39 27.40 27.40
[46] 27.43 27.44 27.45 27.46 27.48 27.49 27.49 27.49 27.51
[55] 27.52 27.52
```

Measuring Spread of Data Distribution

The average devation $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$ **always** $= 0$!

So, we need a different measure for distance (spread)

There are many measures of spread:

- mean squared deviation (MSD or "variance"),
- mean absolute deviation (MAD),
- standard deviation (SD) $=$ root $MSD = RMSD = \sqrt{MSD}$,
- interquartile range ($IQR =$ range of middle 50% of data)
- range,
- ...and more (not covered in this course).

Loss Functions I

No matter which one number we might choose to measure center, we are summarizing an entire distribution with one number.

- There is a cost.
- We lose information.
- We should measure that loss and be aware of its magnitude.
- Statisticians measure loss numerically with a **loss function**
- A loss function measures the distance of the data from the one-number summary (the "center").

A loss function is a measure of distance (spread).

Loss Functions II

Let's consider two common loss functions

- The sum (or mean) of absolute deviations:

$$SAD(w) = \sum_{i=1}^{n} |x_i - w| \qquad MAD(w) = \frac{1}{n}\sum_{i=1}^{n} |x_i - w|$$

- The sum (or mean) of squared deviations:

$$SSD(w) = \sum_{i=1}^{n} (x_i - w)^2 \qquad MSD(w) = \frac{1}{n}\sum_{i=1}^{n} (x_i - w)^2$$

Loss Functions III

What value of w should we choose if loss is SAD? If SSD?

It seems reasonable that w should be in the "center" of the data for each measure. But which value in the middle would be best?

One optimality criteria: Choose w that minimizes SAD or SSD.

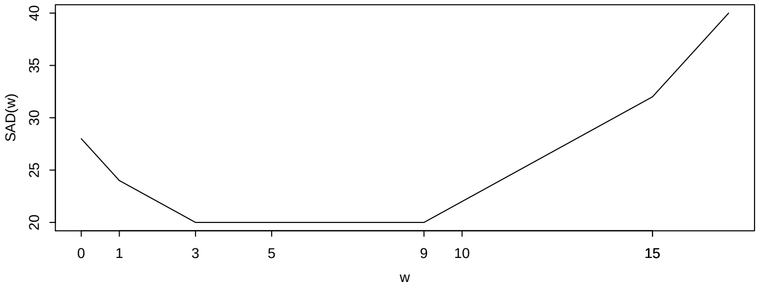
What is so special about the median? I

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

What does the *SAD*(w) function look like for these data?

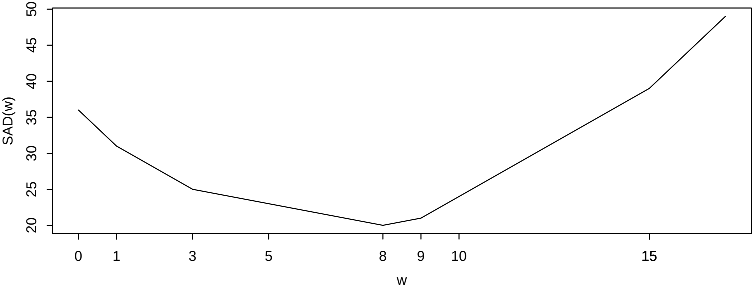
```
ww=seq(0,17,.1)
SAD <- function(w) { sum( abs(x-w) ) }


plot(ww,sapply(ww, SAD), type=c('l'), xlab="w", ylab="SAD(w)")
axis(1,at=x)
```



What is so special about the median? II

```
y <- c(9,3,15,8,1)
SAD1=function(w){sum(abs(y-w))}
plot(ww,sapply(ww, SAD1), type=c('l'), xlab="w", ylab="SAD(w)")
axis(1,at=y)
```

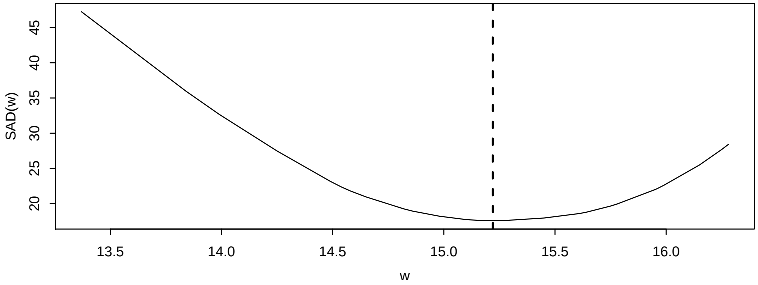


Where is the function $SAD(w)$ smallest (minimized)?

What is so special about the median? III

Looking at the data: Carbon frame `AvgSpeed`

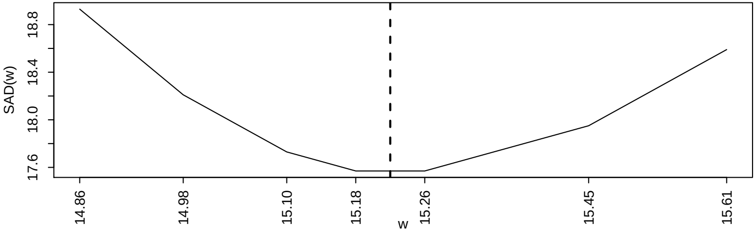
What does the $SAD(w)$ function look like for these data?



Where is the function $SAD(w)$ smallest (minimized)?

What is so special about the median? IV

Zooming in:




```
sort(carbonSpeed)

 [1] 13.37 13.84 13.99 14.25 14.49 14.54 14.58 14.65 14.82
[10] 14.86 14.98 15.10 15.18 15.26 15.45 15.61 15.64 15.75
[19] 15.78 15.95 15.96 15.99 16.15 16.15 16.25 16.28

median(carbonSpeed)

[1] 15.22
```

What is so special about the average? I

Consider again the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

What is the $SSD(w)$ function for these data?

$$\begin{aligned}\sum_{i=1}^{4}(x_i - w)^2 &= (9 - w)^2 + (3 - w)^2 + (15 - w)^2 + (1 - w)^2 \\ &= 4w^2 - 56w + 316\end{aligned}$$

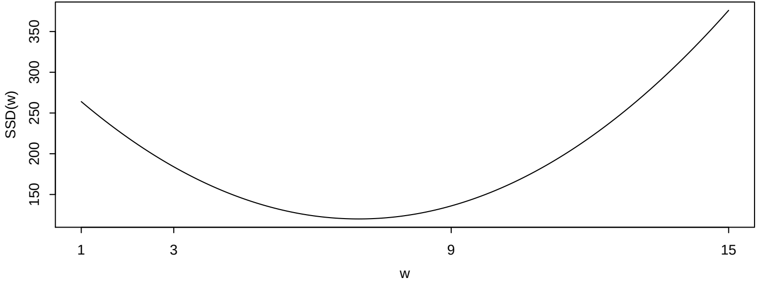
So, as ugly as $\sum_{i=1}^{n}(x_i - w)^2$ originally looks

it's just a smooth quadratic function (convex, opening up).

What is so special about the average? II

What does the $SSD(w)$ function look like for these data?

```
SSD <- function(w) { sum( (x-w)^2 ) }
```



What value w minimizes $SSD(w) = 4w^2 - 56w + 316$?
$\frac{d}{dw} SSD(w) = 8w - 56$.
Set the derivative $= 0$ and solve for w: $w = 7$.

What is so special about the average? III

What value w minimizes $SSD(w)$ for **any** sample: $x_1, x_2, \ldots, x_n$?

We want to minimize the following function with respect to w:

$$f(w) = SSD(w) = \sum_{i=1}^{n} (x_i - w)^2$$

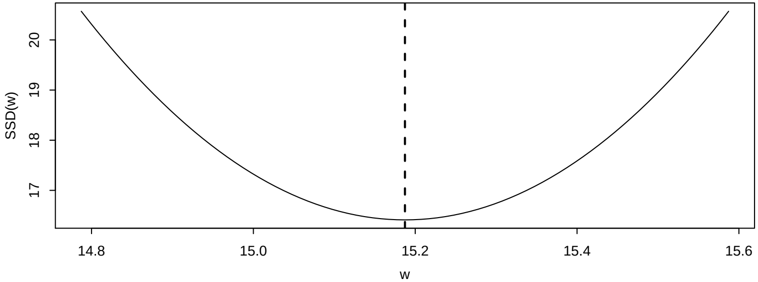
On your own: Show that minimizer is $w = \overline{x}$ (average).
Then, check that the average is the *unique* minimum (not just one of several values that attain the minimum, as for the median).

We say that $\overline{x}$ is a **"least squares"** statistic since it minimizes the sum of squared deviations.

What is so special about the average? IV

Consider again the bike commute data: Carbon frame `AvgSpeed`

What does the $SSD(w)$ function look like for these data?



Where is the function $SSD(w)$ smallest (minimized)?

```
mean(carbonSpeed)
```

```
[1] 15.19
```

Formulas for Sample Average, Variance, SD

$$\text{sample average} = \overline{x} = \text{“x-bar”} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{sample variance} = s^2 = \text{“s-squared”} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$= \text{“typical” distance from the average}$$

Why divide by $(n-1)$ instead of n for sample variance and SD?