Introduction to Descriptive Statistics

17.871

Spring 2012

Reasons for paying attention to data description

- Double-check data acquisition
- Data exploration
- Data explanation

Key measures

Describing data

	Moment	**Non-moment based location parameters**
Center	Mean	Mode, median
Spread	Variance (standard deviation)	Range, Interquartile range
Skew	Skewness	--
Peaked	Kurtosis	--

Key distinction

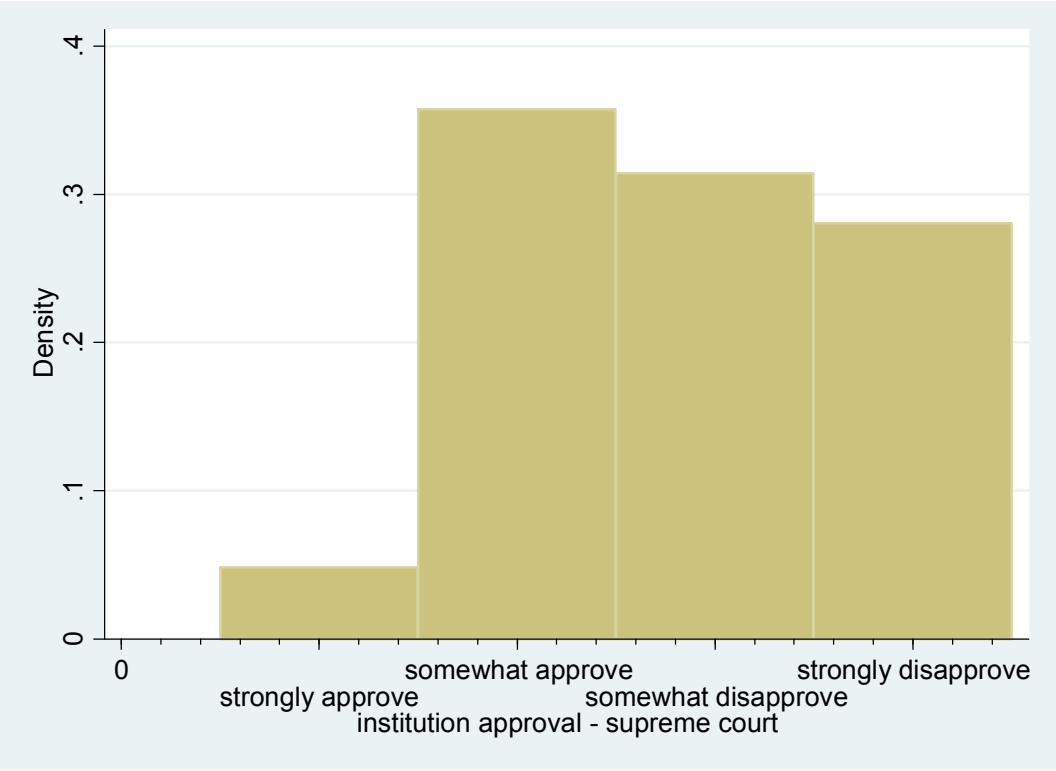
Population vs. Sample Notation

Population	***vs.***	**Sample**
Greeks		Romans
μ, σ, β		s, b

Mean

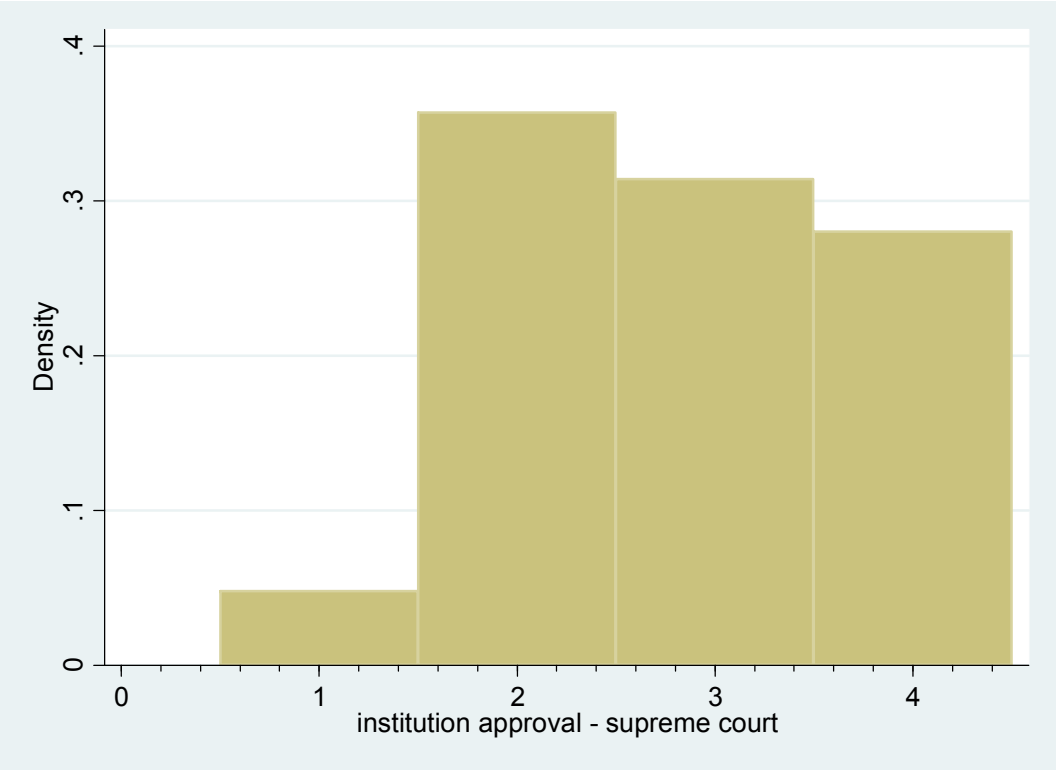
$$\frac{\sum_{i=1}^{n} x_i}{n} \equiv \mu \equiv \overline{X}$$

Guess the Mean



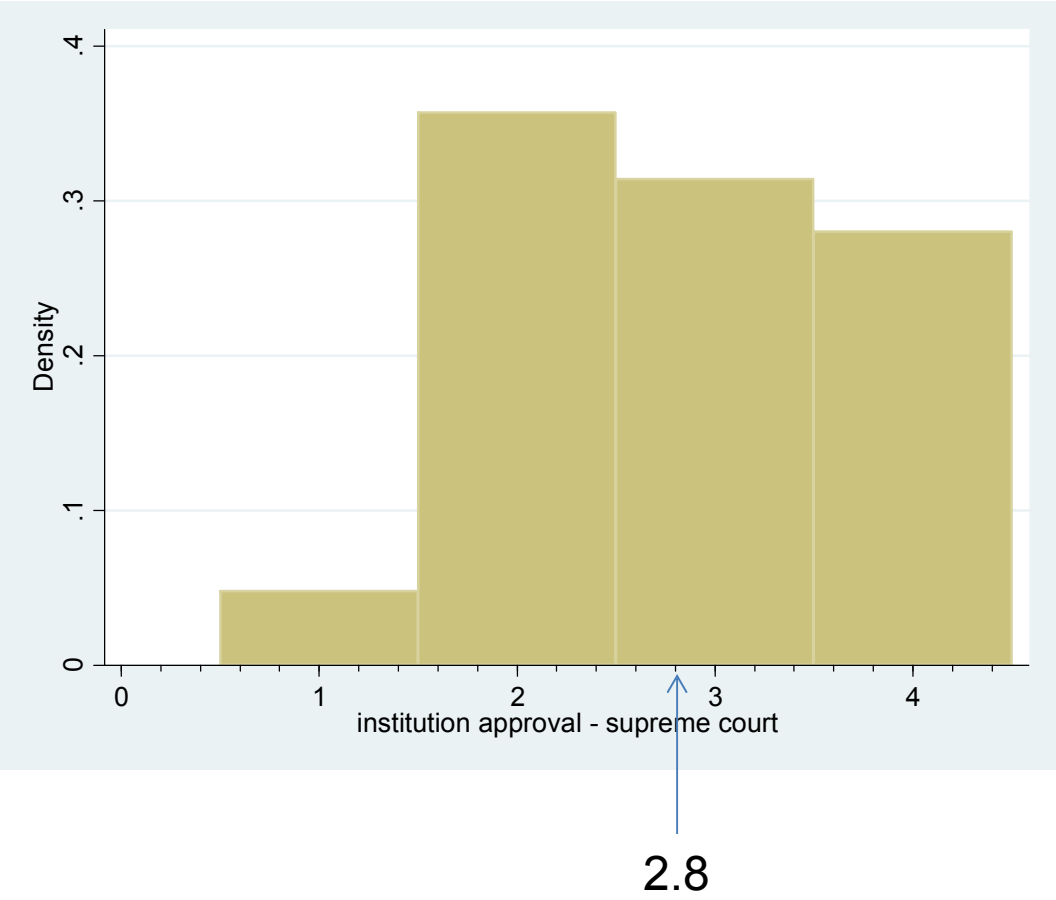
Source: CCES

Guess the Mean



Source: CCES

Guess the Mean



Source: CCES

Variance, Standard Deviation of a Population

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

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

Variance, S.D. of a Sample

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1} \equiv s^2,$$

Degrees of freedom

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

Guess

What was the mean and standard deviation of the MIT undergraduate population on Registration Day, Fall 2012?

Guess

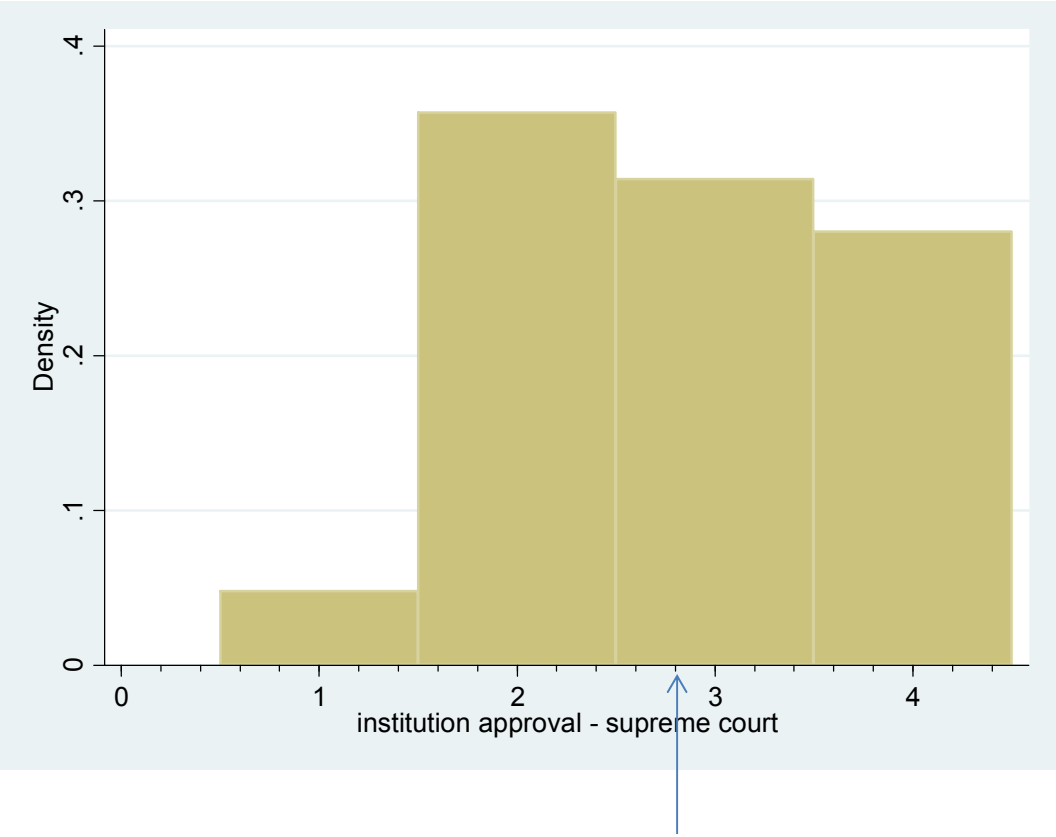
What was the mean and standard deviation of the MIT undergraduate population on Registration Day, Fall 2012?

My guess:

Mean probably ~ 19.5 (if everyone is 18, 19, 20, or 21, and they are evenly distributed.

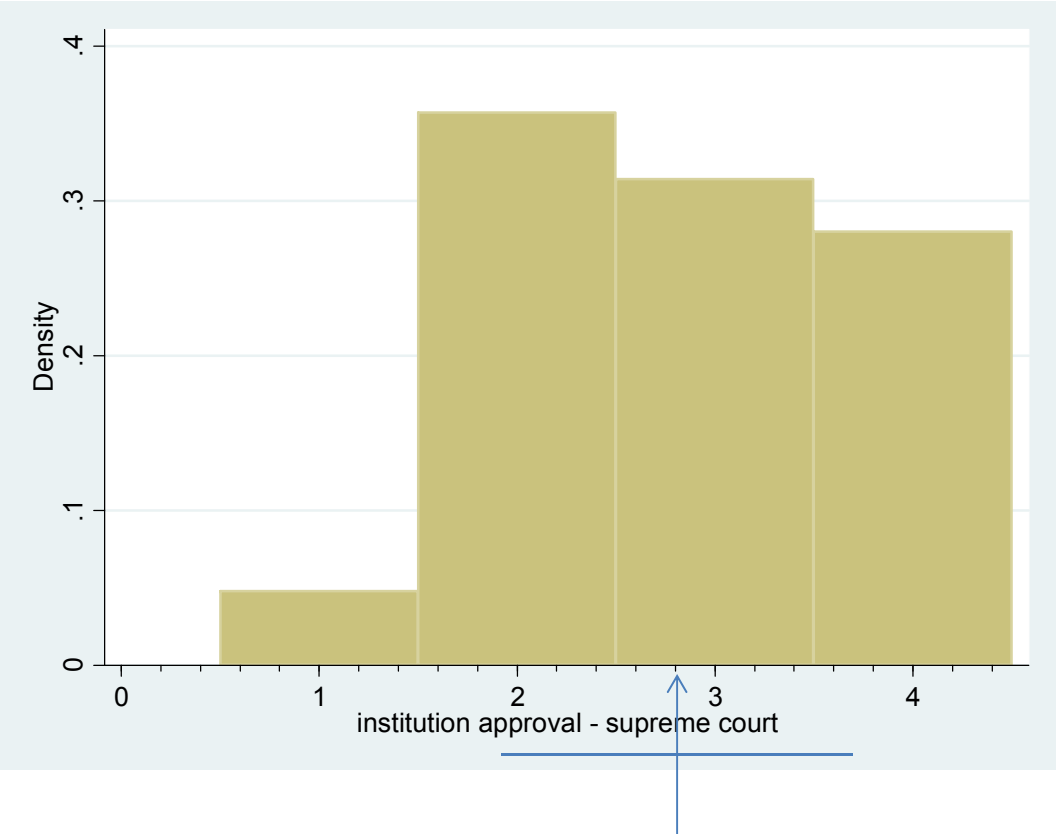
s.d. probably ~ 1

Guess the Standard Deviation



Source: CCES

Guess the Standard Deviation



σ = 0.89

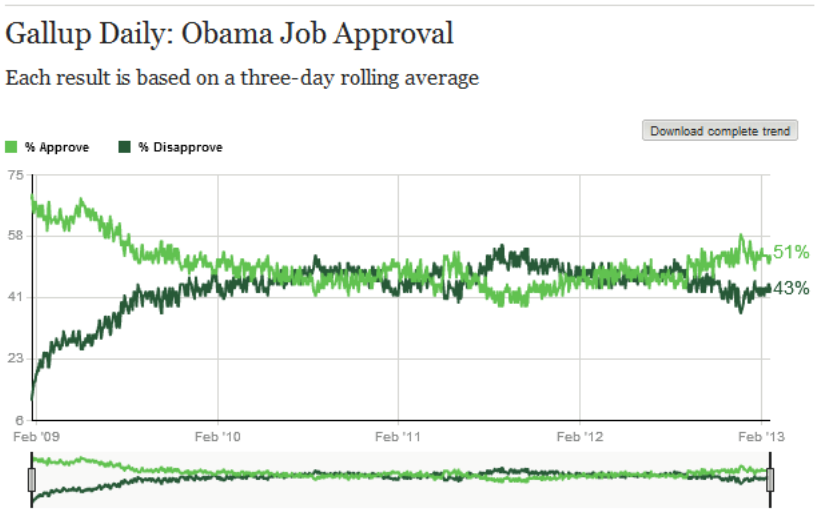
Source: CCES

Binary data

$$\overline{X} = prob(X) = 1 = \text{proportion of time } x = 1$$

$$s_x^2 = \bar{x}(1-\bar{x}) \Rightarrow s_x = \sqrt{\bar{x}(1-\bar{x})}$$

Example of this, using the most recent Gallup approval rating of Pres. Obama



- `gen o_approve = 1 if gallup==“Approve”`
- `replace o_approve = 0 if gallup==“Disapprove”`
- the command `summ o_approve` produces
- Mean = 0.51
- Var = 0.51(1-0.51)=.2499
- S.d. = .49989999

Therefore, reporting the standard deviation (or variance) of a binary variable is redundant information. **Don't do it** for papers written for 17.871.

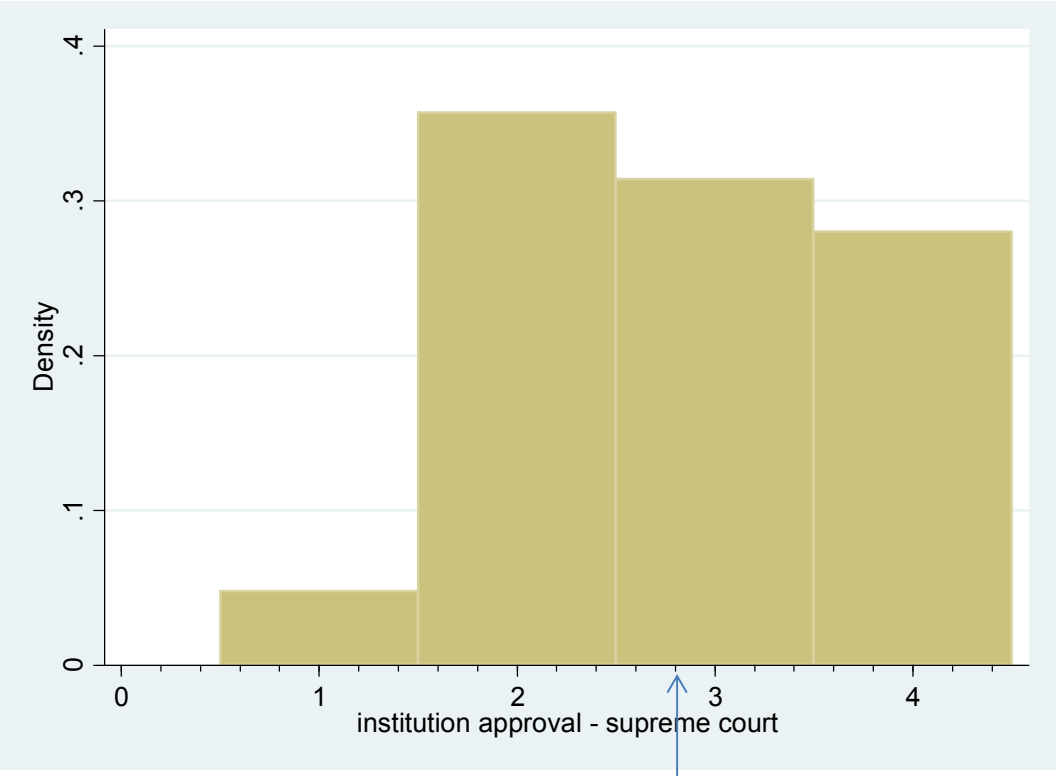
Non-moment base measures of center or spread

- Central tendency
  - Mode
  - Median
- Spread
  - Range
  - Interquartile range

Mode

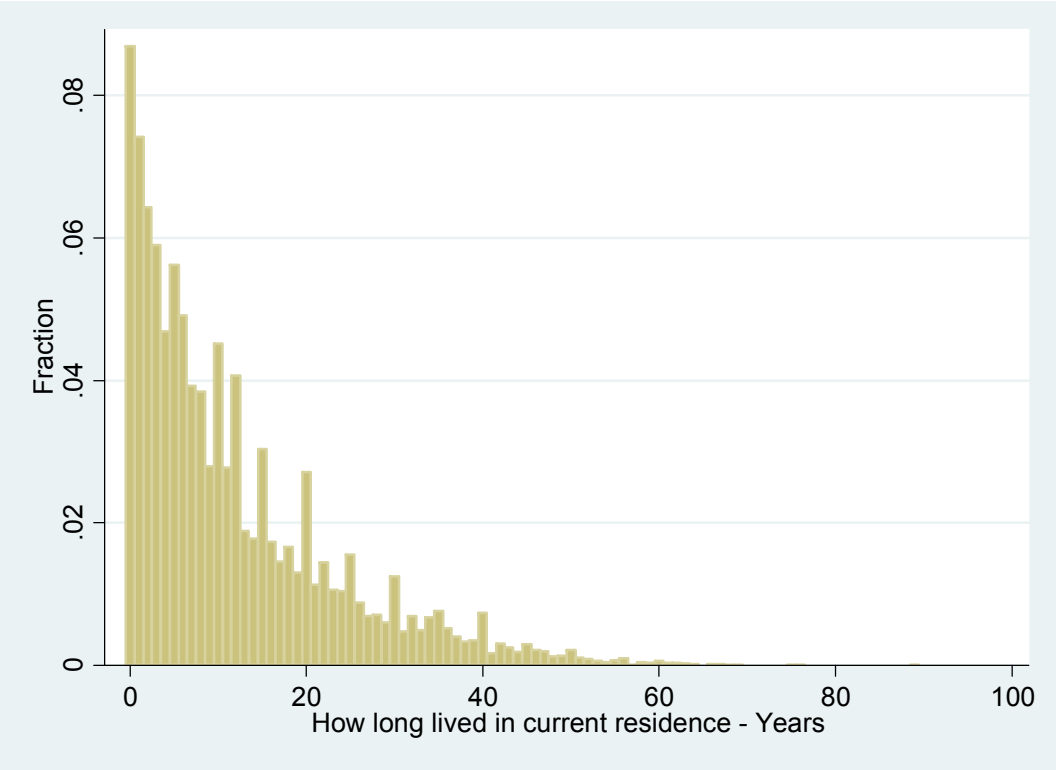
- The most common value

Guess the Mode



Source: CCES

Guess the Mode



Number of years the respondent has lived in his/her current home

Guess the Mode

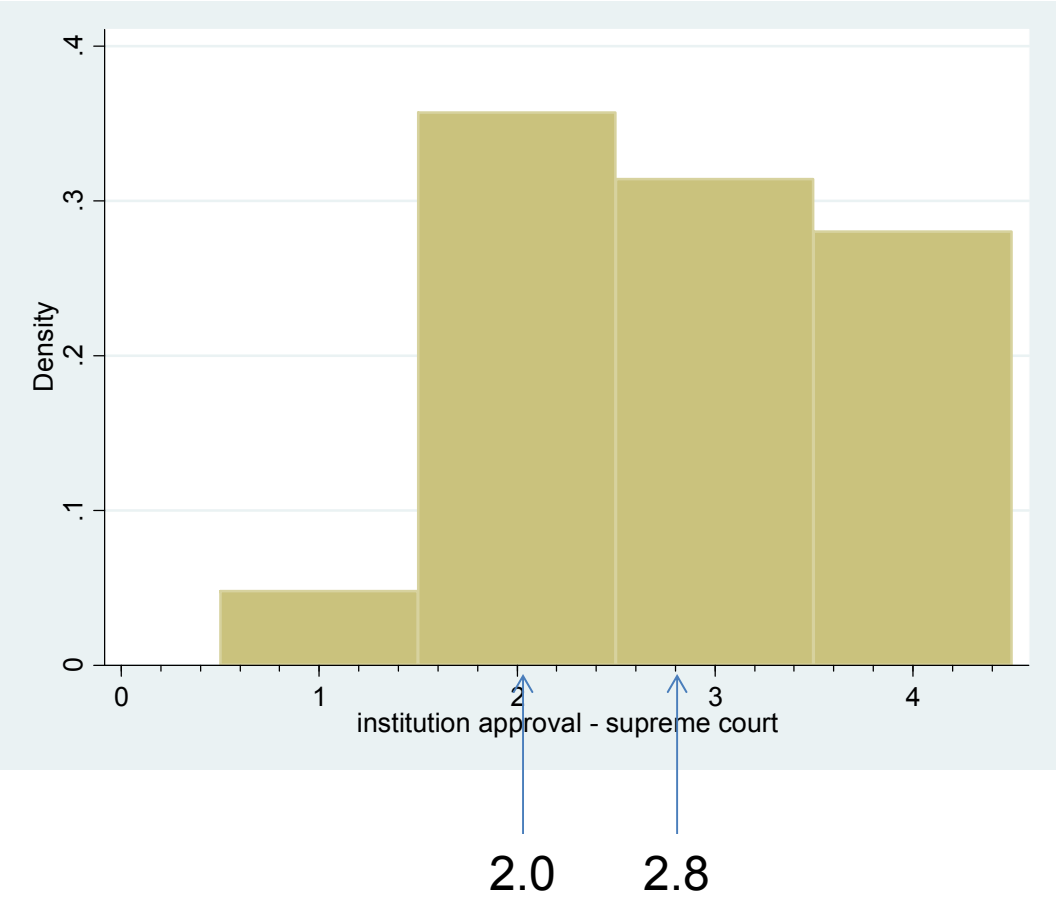
pew religion	Freq.	Percent	Cum.
protestant	26,241	47.40	47.40
roman catholic	12,348	22.30	69.70
mormon	931	1.68	71.38
eastern or greek orthodox	275	0.50	71.88
jewish	1,678	3.03	74.91
muslim	164	0.30	75.21
buddhist	445	0.80	76.01
hindu	89	0.16	76.17
agnostic	2,885	5.21	81.38
nothing in particular	7,641	13.80	95.18
something else	2,667	4.82	100.00
Total	55,364	100.00	

The mode is rarely an informative statistic about the central tendency of the data.
It’s most useful in describing the “typical” observation of a categorical variable

Median

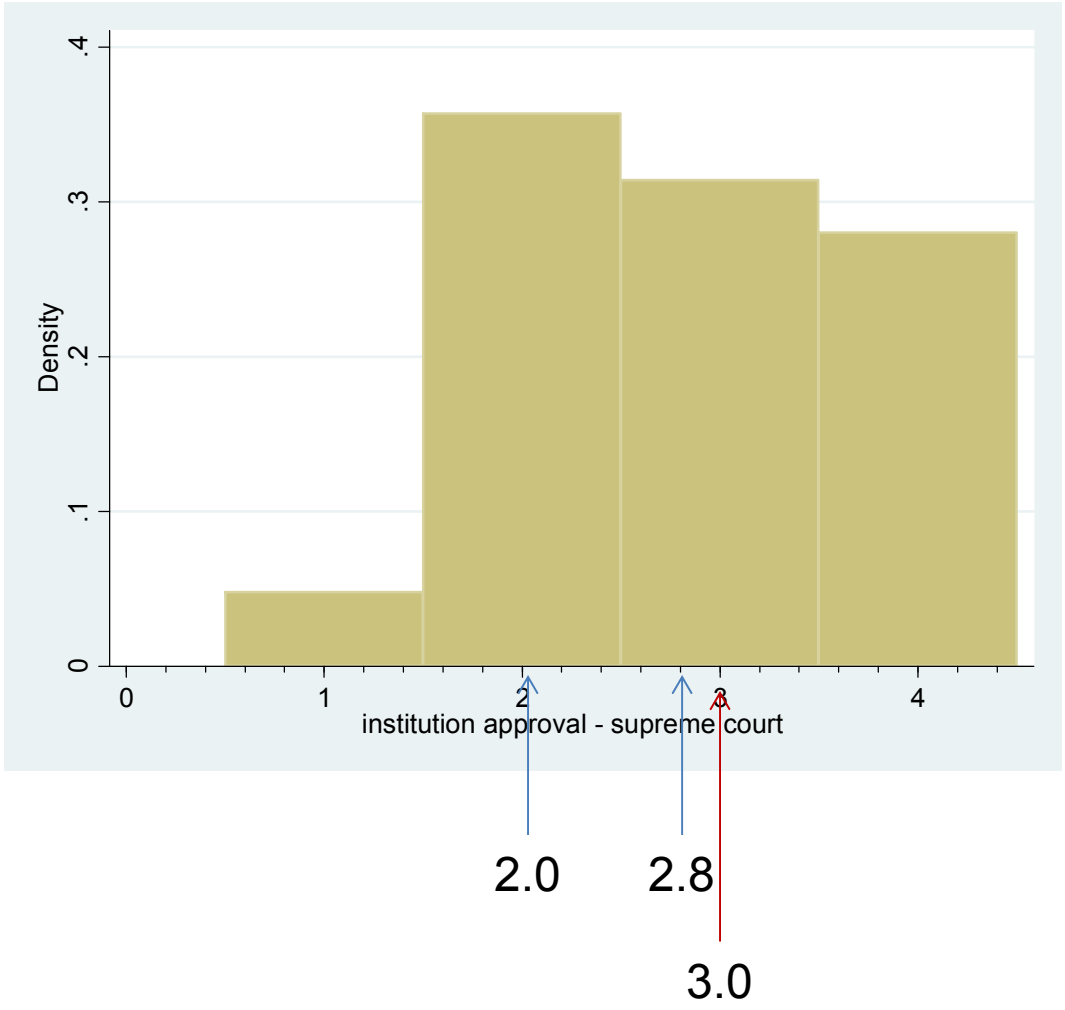
- The numerical value separating the upper half of a distribution from the lower half of the distribution
  - If N is odd, there is a unique median
  - If N is event, there is no unique median --- the convention is to average the two middle values

Guess the Median



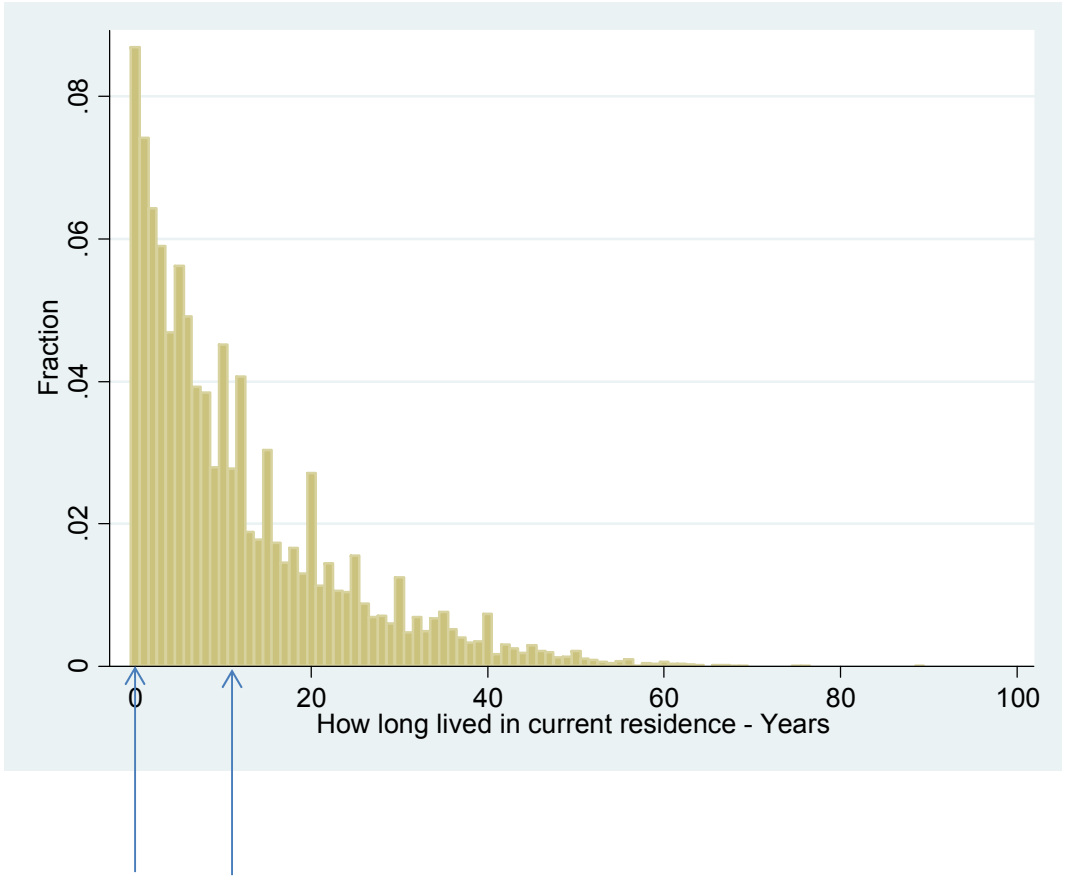
Source: CCES

Guess the Median



Source: CCES

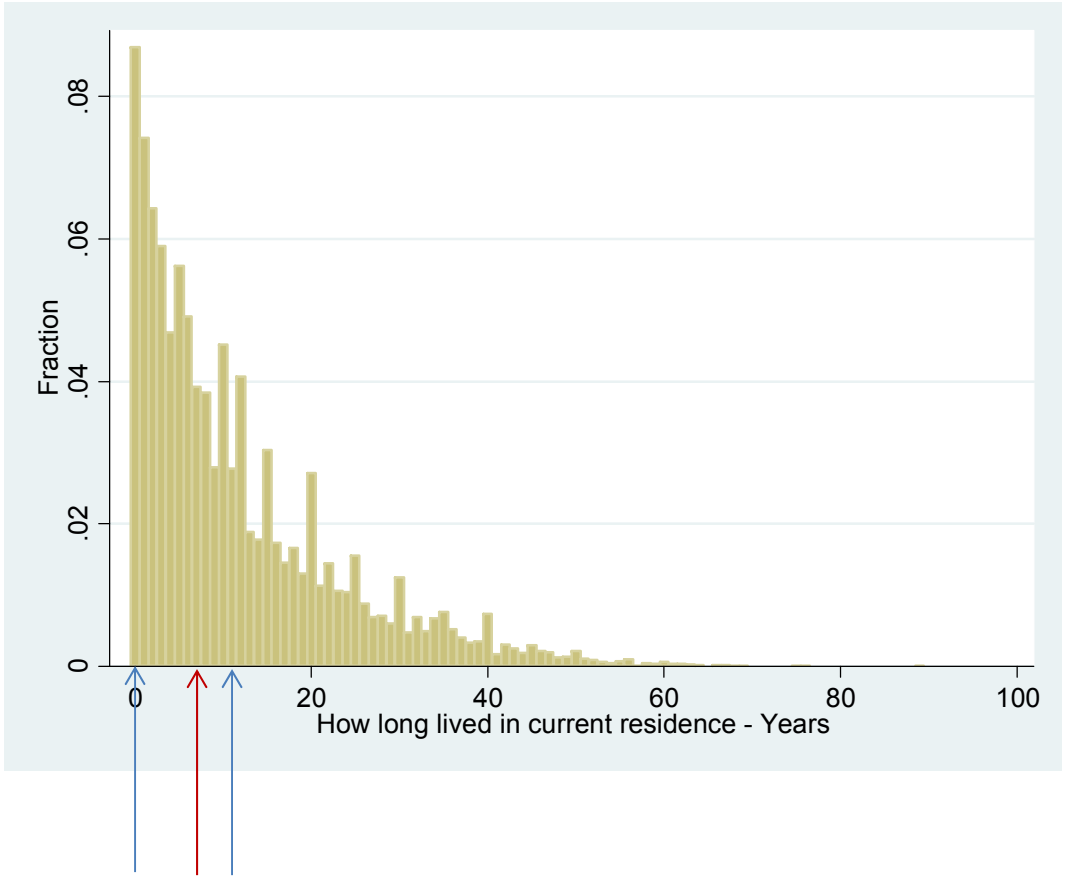
Guess the Median



Mode = 0
Mean = 11.8

Number of years the respondent has lived in his/her current home

Guess the Median



Mode = 0

Mean = 11.8

Median = 8

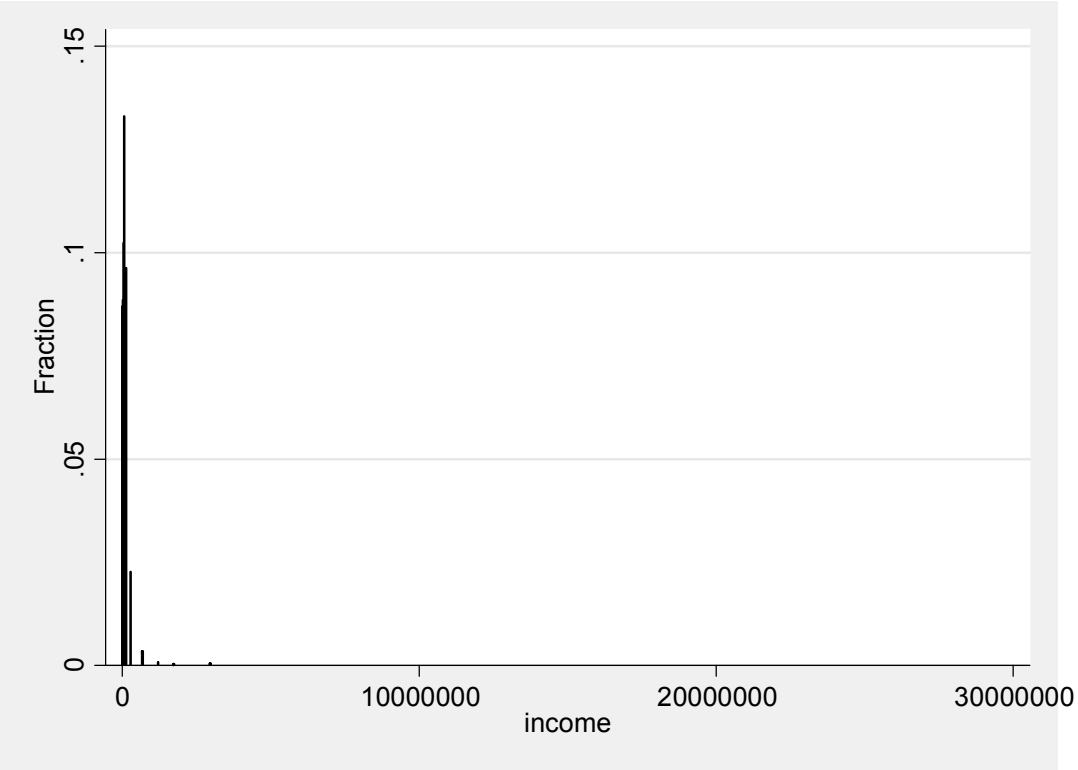
Number of years the respondent has lived in his/her current home

Median frequently preferred for income data

Table 1.1 Selected Income and Tax Items, by Size and Accumulated Size of Adjusted Gross Income, Tax Year 2009
(All figures are estimates based on samples—money amounts are in thousands of dollars except as indicated)

Size and accumulated size of adjusted gross income	Number of returns	Percent of total
	(1)	(2)
All returns	**140,494,127**	**100.0**
No adjusted gross income	2,511,925	1.8
$1 under $5,000	10,447,635	7.4
$5,000 under $10,000	12,220,335	8.7
$10,000 under $15,000	12,444,512	8.9
$15,000 under $20,000	11,400,228	8.1
$20,000 under $25,000	10,033,887	7.1
$25,000 under $30,000	8,662,392	6.2
$30,000 under $40,000	14,371,647	10.2
$40,000 under $50,000	10,796,412	7.7
$50,000 under $75,000	18,694,893	13.3
$75,000 under $100,000	11,463,725	8.2
$100,000 under $200,000	13,522,048	9.6
$200,000 under $500,000	3,195,039	2.3
$500,000 under $1,000,000	492,568	0.4
$1,000,000 under $1,500,000	108,096	0.1
$1,500,000 under $2,000,000	44,273	[2]
$2,000,000 under $5,000,000	61,918	[2]
$5,000,000 under $10,000,000	14,322	[2]
$10,000,000 or more	8,274	[2]

The (uninformative) graph



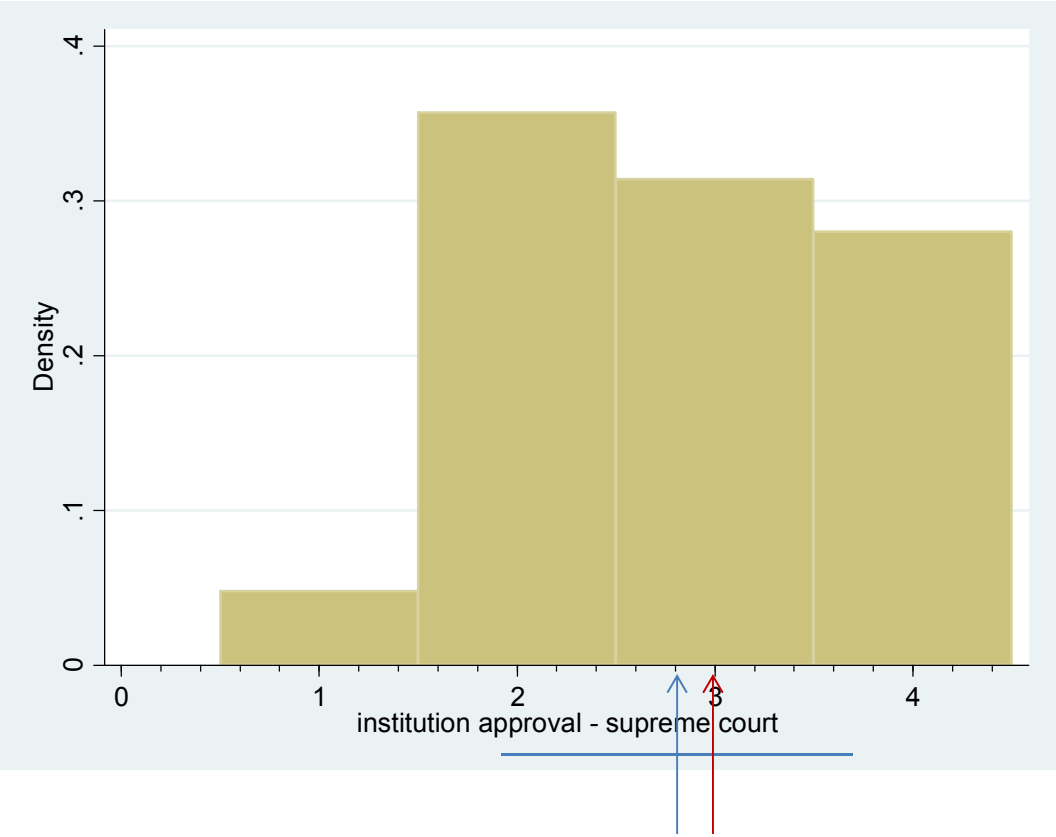
Mean = 55,699
Median = 34,738
Mode = 0 (probably)

Spread

- Range
  - Max(x) – Min(x)
- Interquartile range (IQR)
  - $Q_3(x) – Q_1(x)$

$Q_1 = CDF^{-1}(.25)$
$Q_3 = CDF^{-1}(.75)$

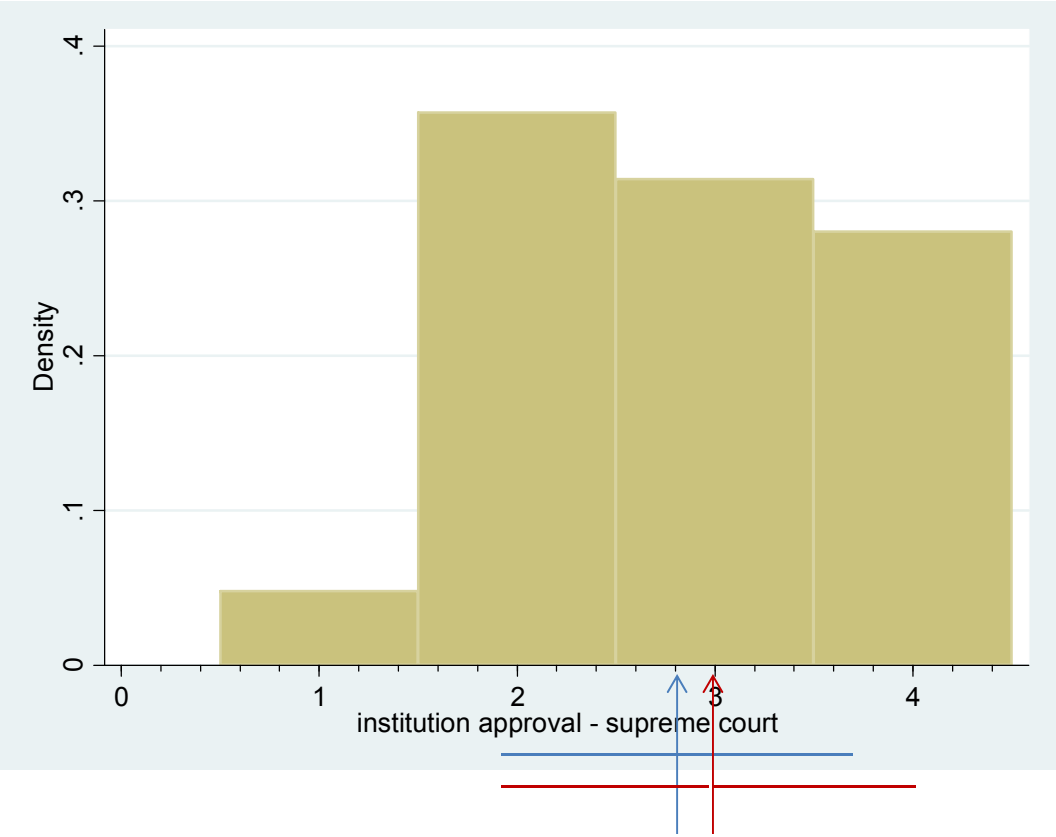
Guess the IQR



σ = 0.89

Source: CCES

Guess the IQR

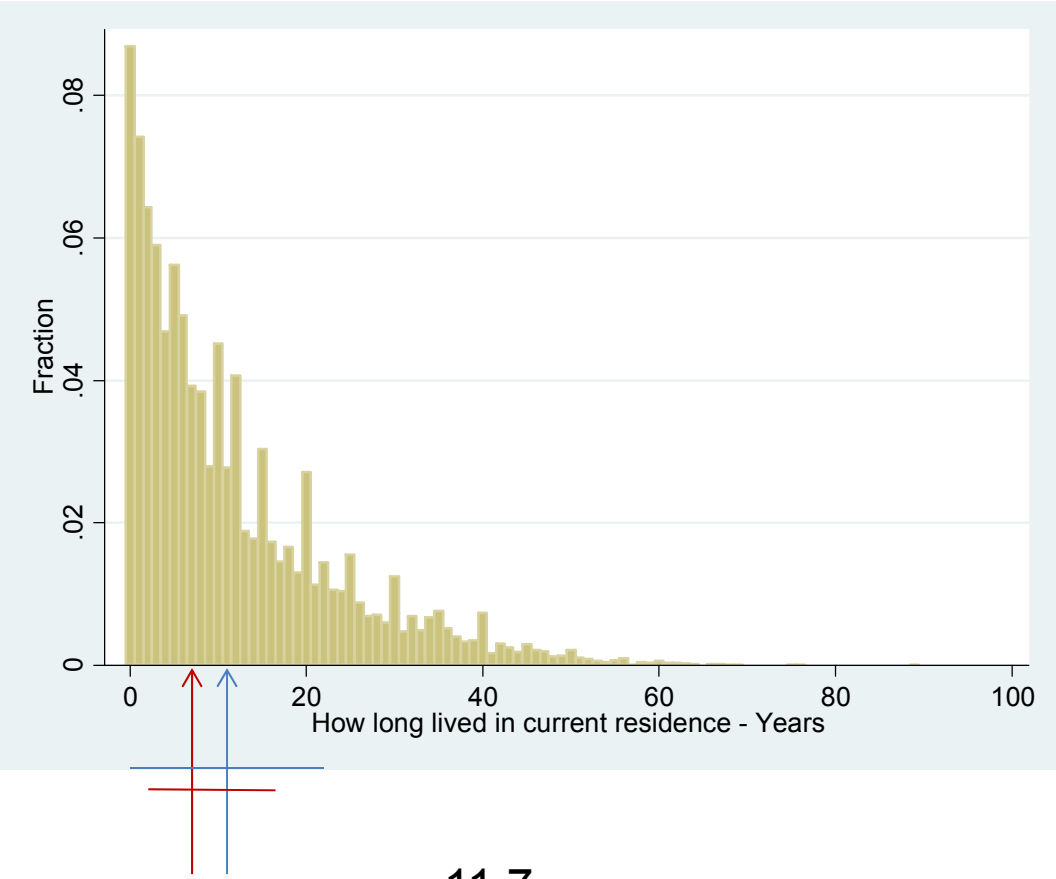


σ = 0.89

IQR = 2

Source: CCES

Guess the IQR

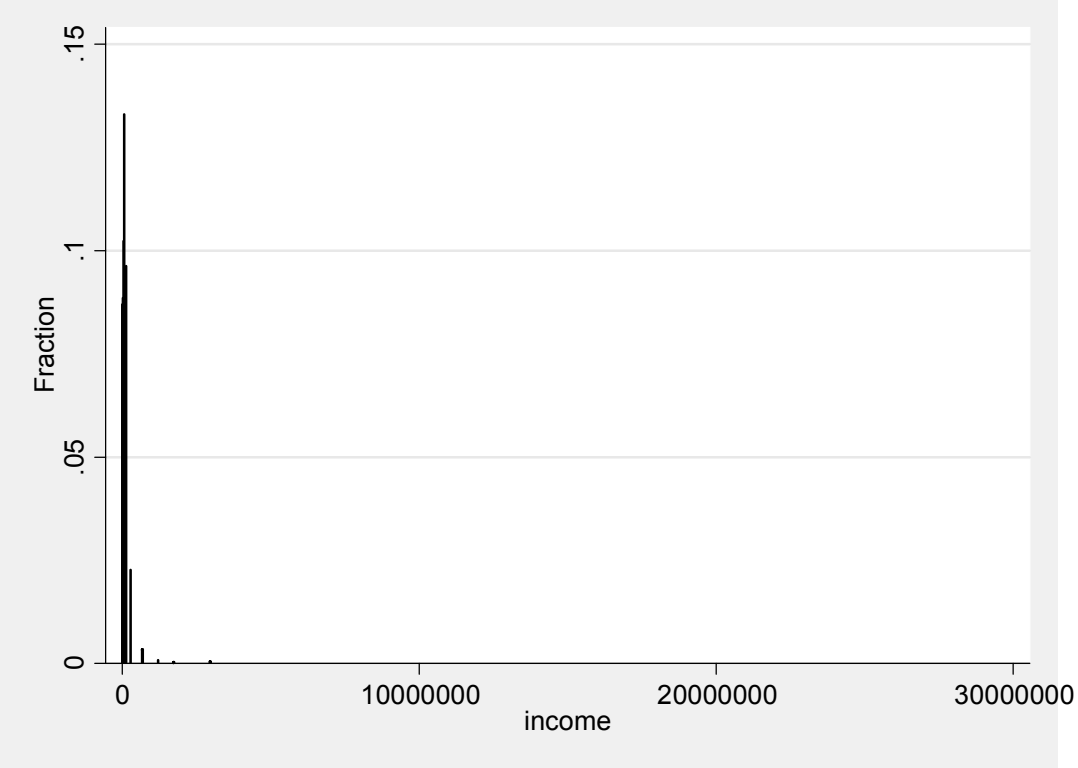


Mode = 0
Mean = 11.8
Median = 8

σ = 11.7
IQR = 14 (17-3)

Number of years the respondent has lived in his/her current home

Guess the IQR

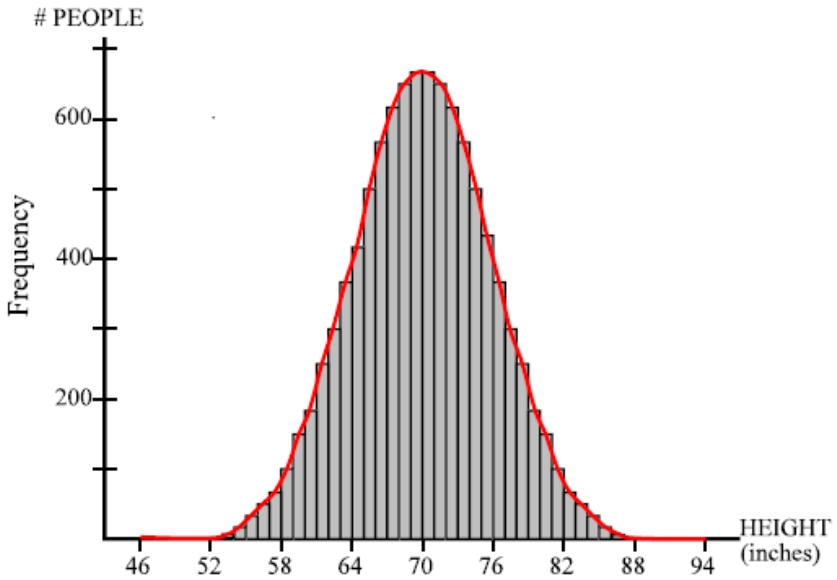


Mean = 55,699
Median = 34,738
Mode = 0 (probably)

σ = 252,522
IQR = 48,971 (61,464 – 12,493)

Lopsidedness and peakedness

Normal distribution example



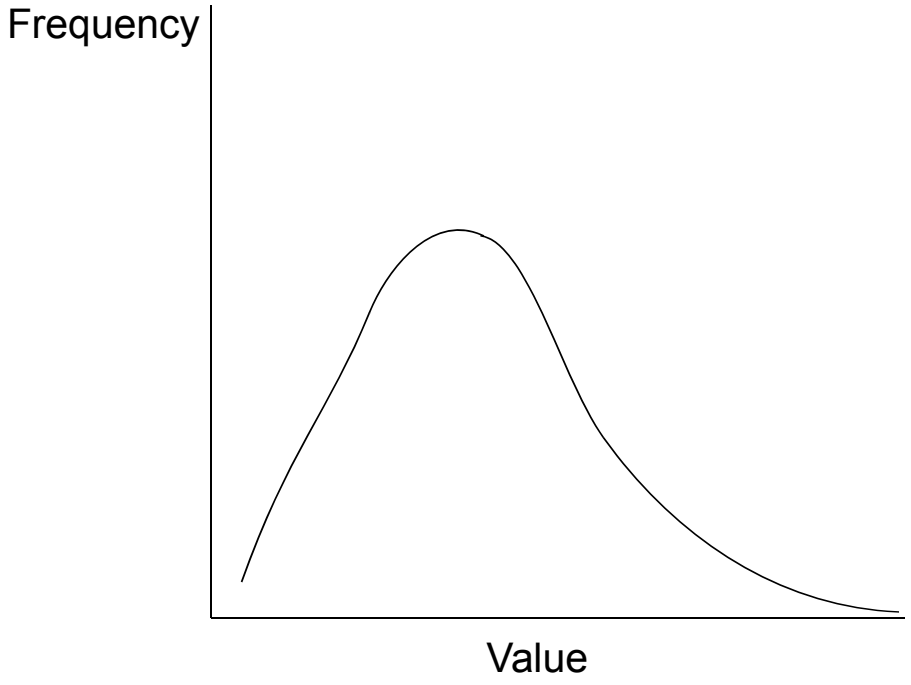
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

- IQ
- SAT
- Height

- Symmetrical
- Mean = median = mode

Skewness

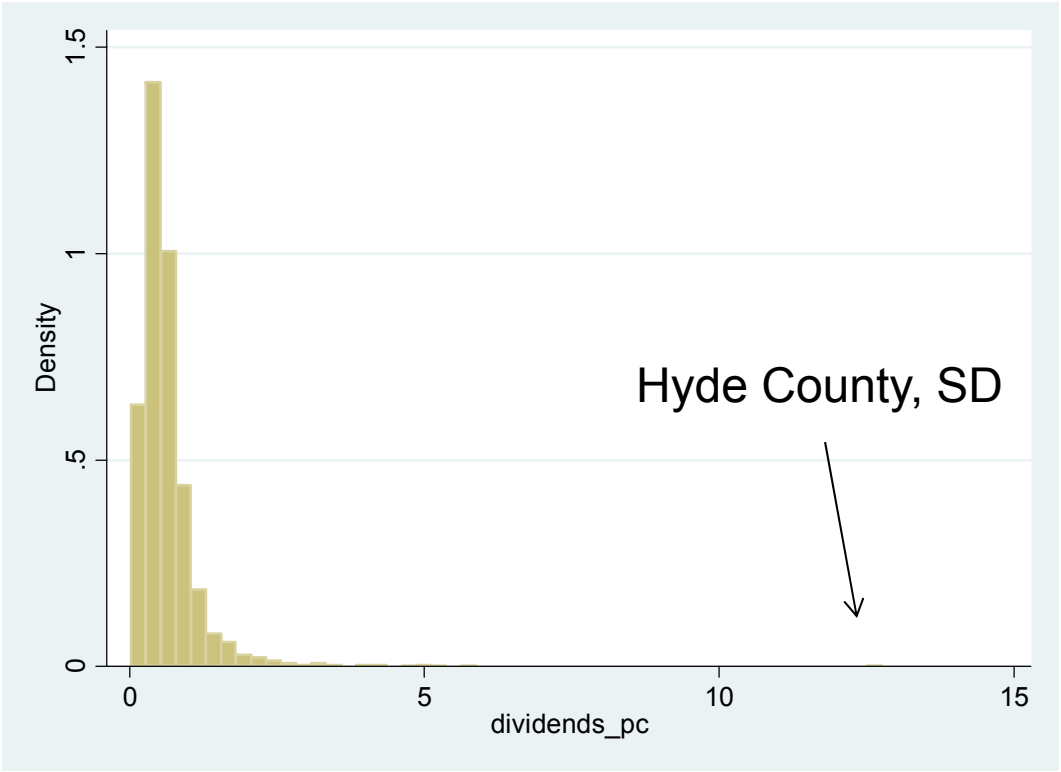
Asymmetrical distribution



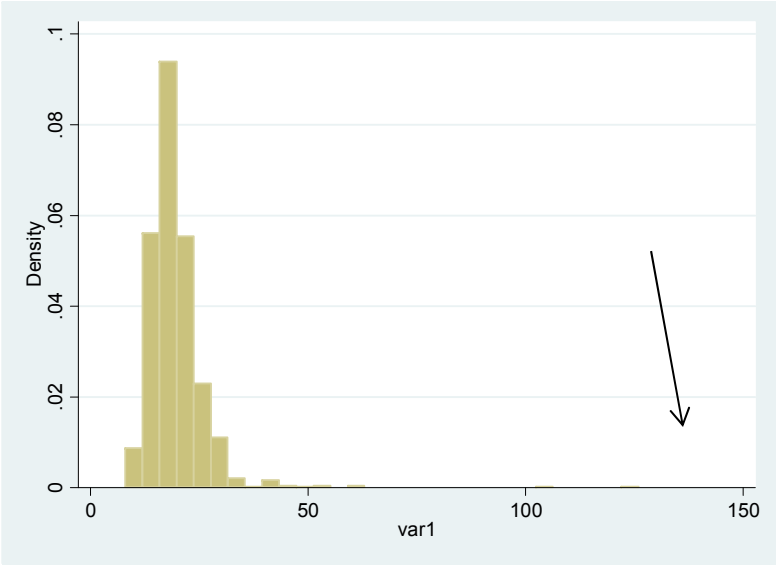
- Income
- Contribution to candidates
- Populations of countries
- “Residual vote” rates

- “Positive skew”
- “Right skew”

Distribution of the average $$ of dividends/tax return (in K's)

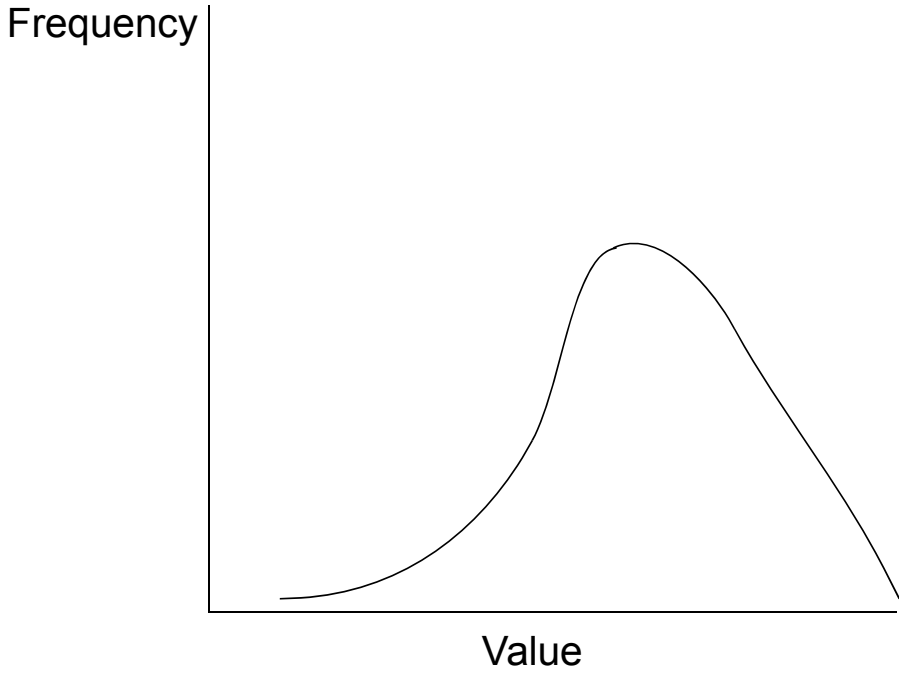


Mitsubishi i-MiEV
(which is supposed to be all electric)



Fuel economy of cars for sale in the US

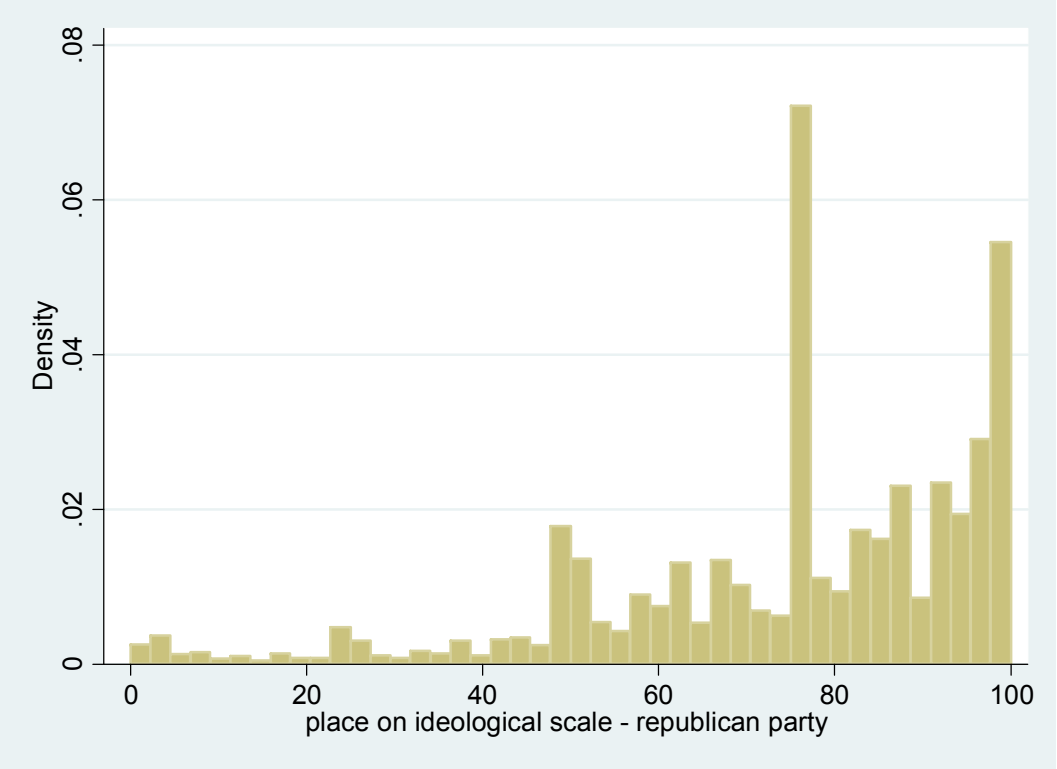
Skewness
Asymmetrical distribution



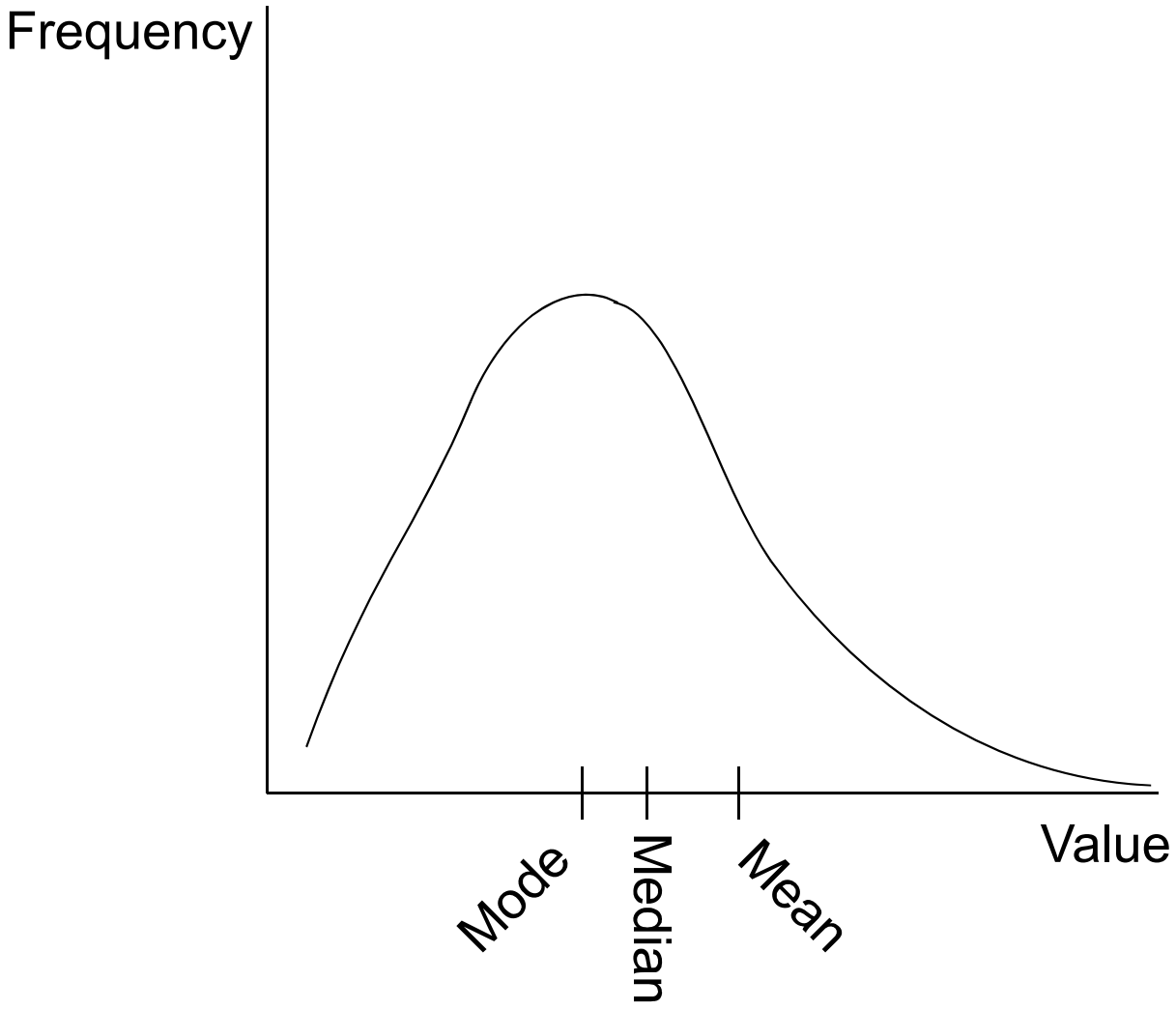
- GPA of MIT students

- “Negative skew”
- “Left skew”

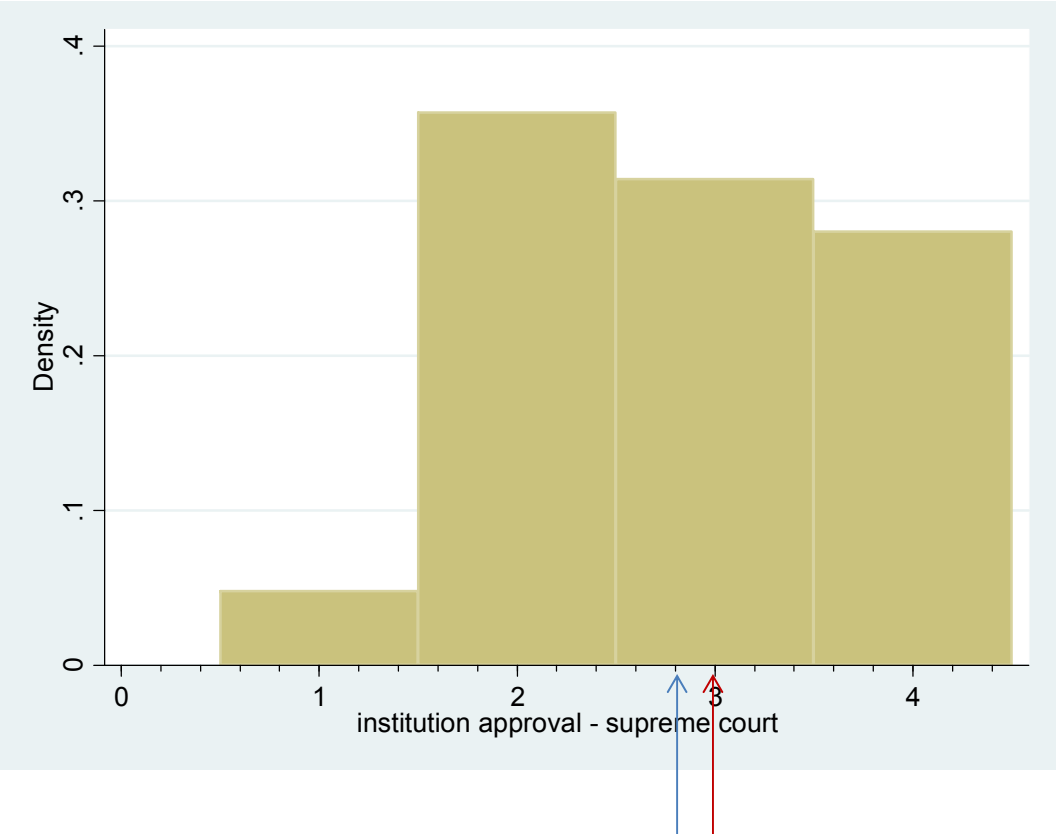
Placement of Republican Party on 100-point scale



Skewness



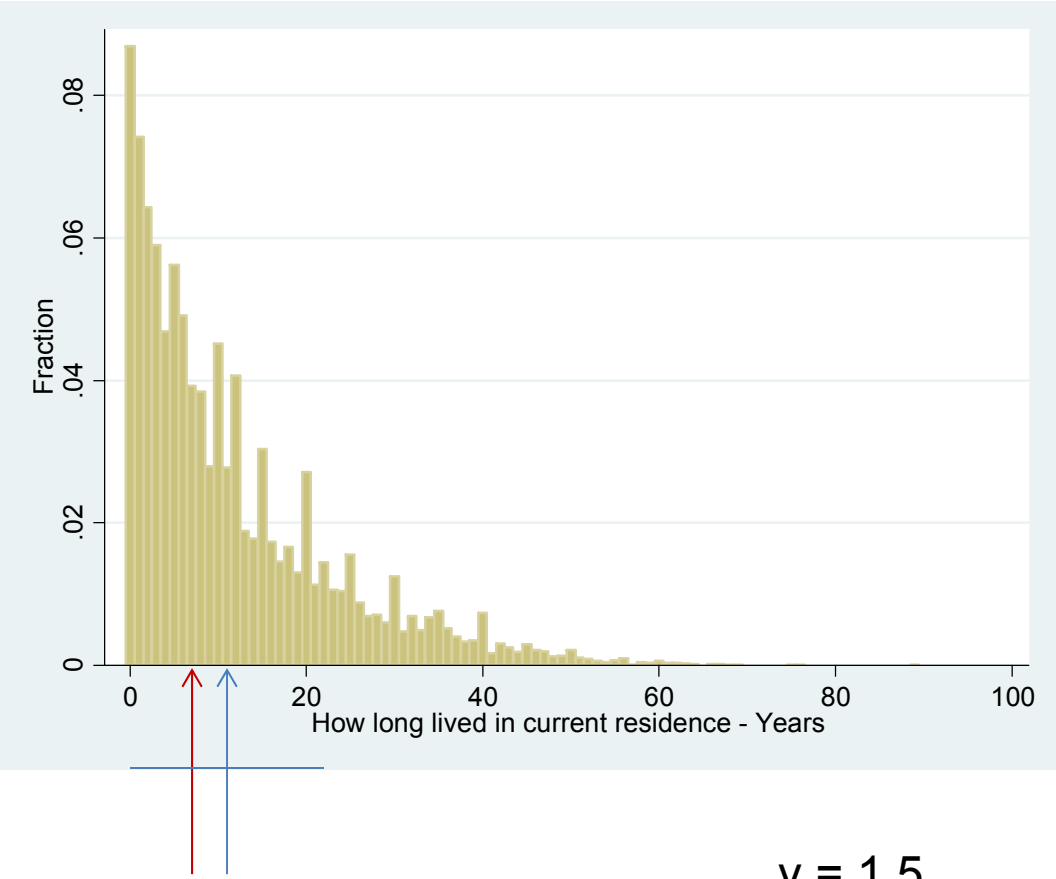
Guess the skew



$\gamma = 0.89$

Source: CCES

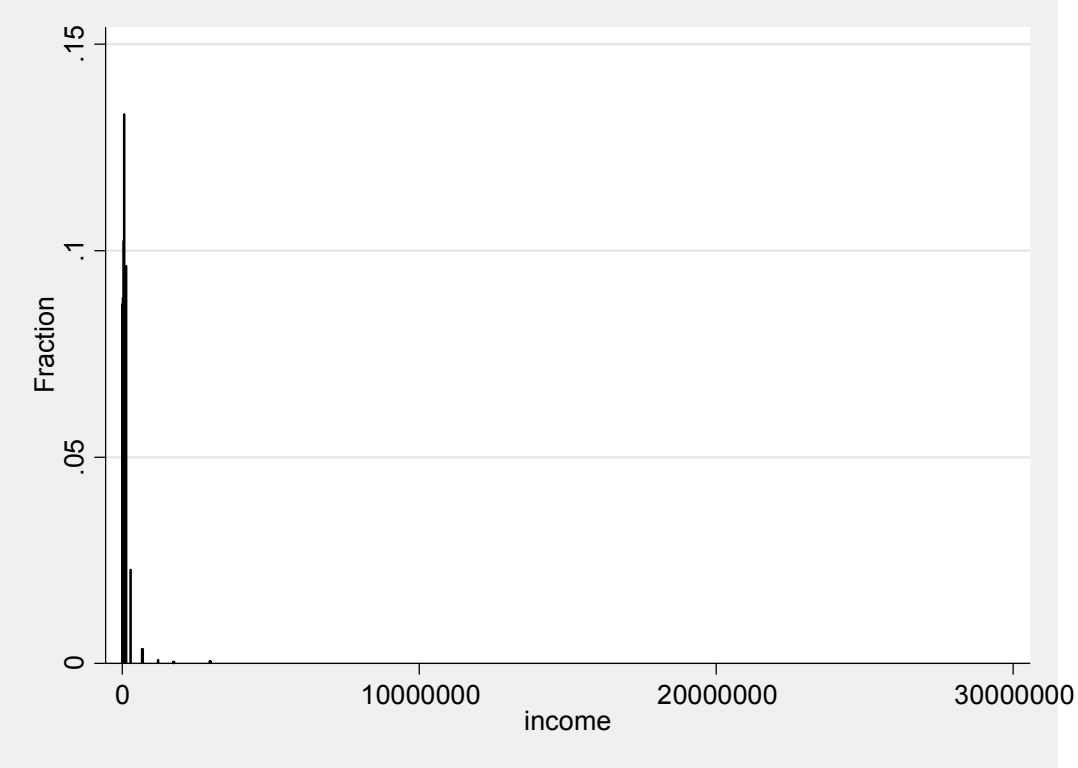
Guess the skew



$\gamma = 1.5$

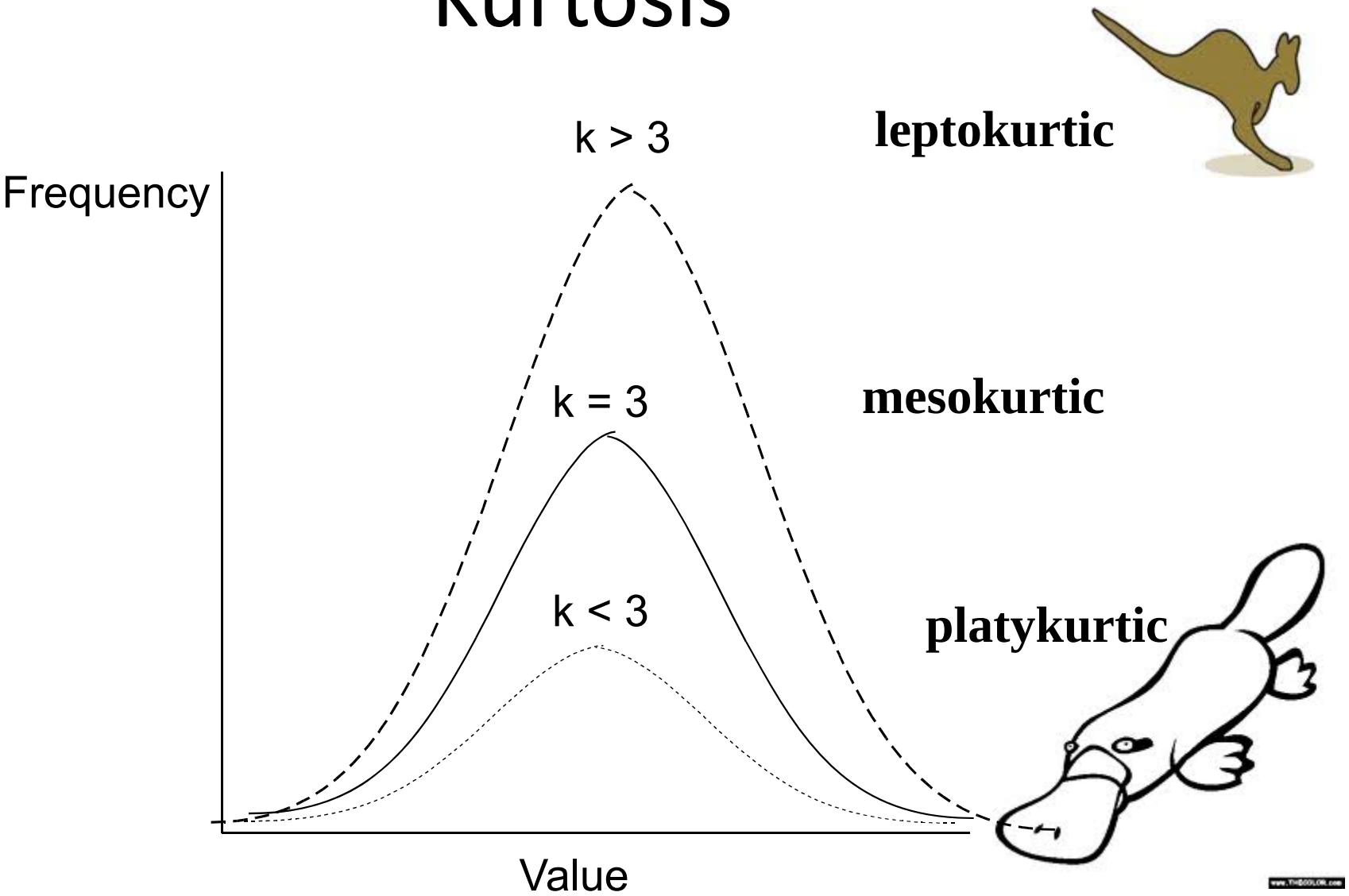
Number of years the respondent has lived in his/her current home

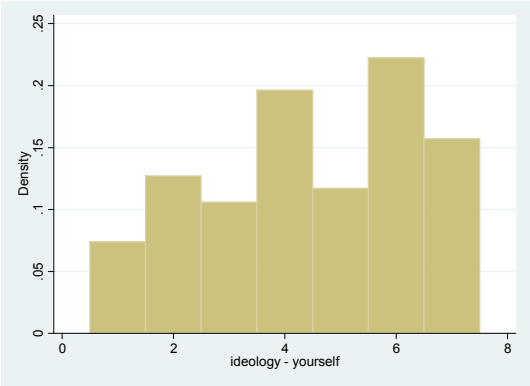
Guess the IQR

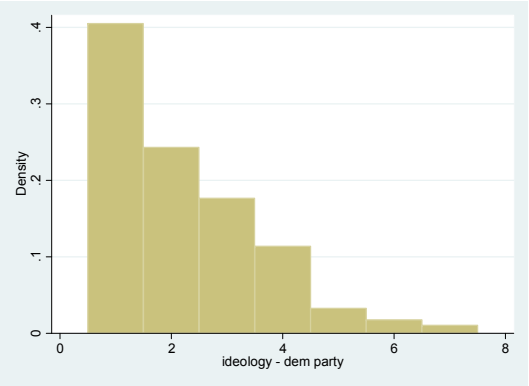


γ = 91.7

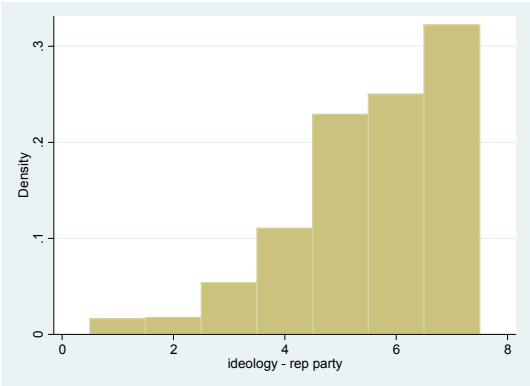
Kurtosis


k > 3
leptokurtic
Frequency
k = 3
mesokurtic
k < 3
platykurtic
Value

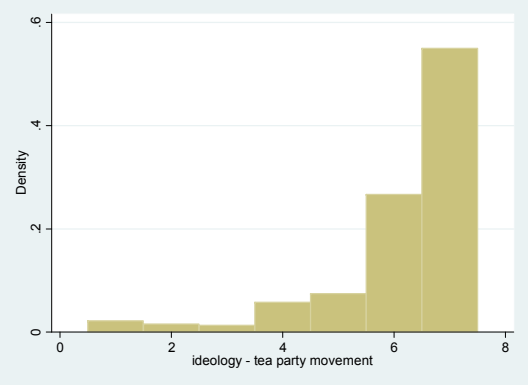




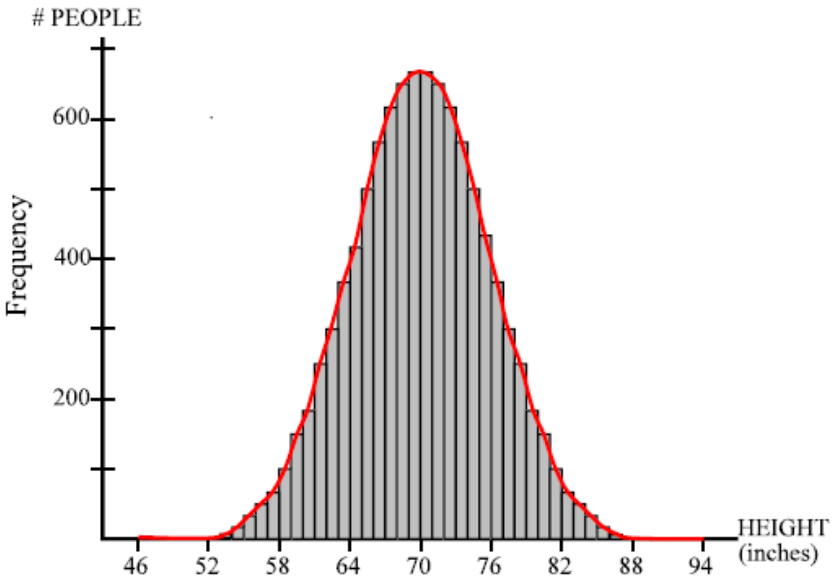
	Mean	s.d.	Skew.	Kurt.
Self-placement	4.5	1.9	-0.28	1.9
Dem. pty	2.2	1.4	1.1	3.9
Rep. pty	5.6	1.4	-0.98	3.7
Tea party	6.1	1.3	-2.1	7.5



Source: CCES, 2010



Normal distribution



- Skewness = 0
- Kurtosis = 3

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

Commands in STATA for univariate statistics

- summarize *varlist*
- summarize *varlist*, detail
- histogram *varname*, *bin() start() width() density/fraction/frequency normal discrete*
- table *varname*,contents(*clist)*
- tabstat *varlist,*statistics(*statname…*)

- tabulate

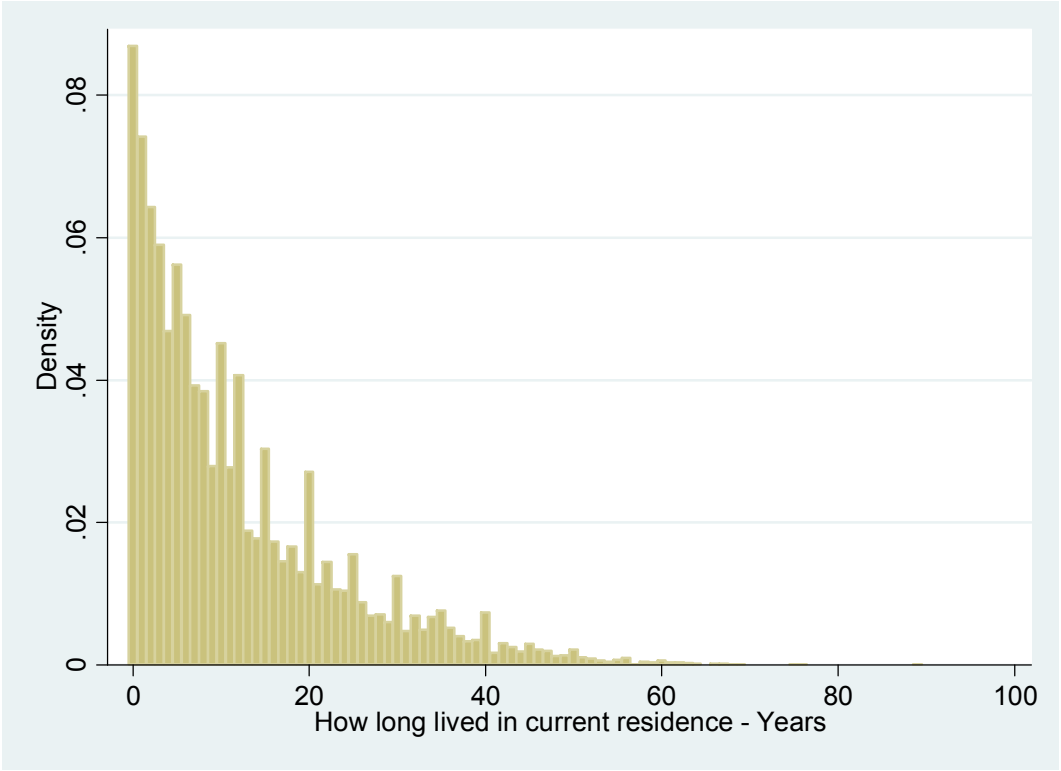
```
. summ time_1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      time_1 |     10153    11.78371    11.70837           0         89

. summ time_1,det

         How long lived in current residence - Years
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            1              0       Obs                 10153
25%            3              0       Sum of Wgt.         10153

50%            8                      Mean           11.78371
                      Largest         Std. Dev.      11.70837
75%           17             69
90%           29             75       Variance        137.086
95%           36             76       Skewness       1.470977
99%           50             89       Kurtosis        5.21861
```




```
. hist time_1,discrete
(start=0, width=1)
```

```
. table pid3

-------------------------
3 point      |
party ID     |      Freq.
-------------+-----------
   Democrat  |      3,808
 Republican  |      3,036
Independent  |      2,825
      Other  |        234
   Not sure  |        297
-------------------------
```

```
. tabstat time_1 age

   stats |    time_1       age
---------+--------------------
    mean |  11.78371  49.33363
------------------------------

. tabstat time_1 age,stats(mean sd skew kurt)

   stats |    time_1       age
---------+--------------------
    mean |  11.78371  49.33363
      sd |  11.70837  15.89716
skewness |  1.470977 -.0152461
kurtosis |   5.21861  2.177523
------------------------------
```

```
. tabstat time_1 age,by(pid3) s(mean sd)

Summary statistics: mean, sd
  by categories of: pid3 (3 point party ID)

       pid3 |    time_1       age
------------+--------------------
   Democrat |  11.28602  47.72348
            |   11.7268  15.88458
------------+--------------------
 Republican |   13.1379  52.27569
            |  12.17941  15.69504
------------+--------------------
Independent |  11.66335  50.07646
            |  11.35228  15.51778
------------+--------------------
      Other |  8.457265  42.52991
            |  9.328546  13.84282
------------+--------------------
   Not sure |  8.084459  38.19865
            |  9.559606  14.14754
------------+--------------------
      Total |  11.78371  49.33363
            |  11.70837  15.89716
----------------------------------
```

```
. table pid3,c(mean time_1 sd time_1)

-------------------------------------
3 point      |
party ID     | mean(time_1)    sd(time_1)
-------------+-----------------------
   Democrat  |       11.286       11.7268
 Republican  |      13.1379      12.17941
Independent  |      11.6634      11.35228
      Other  |      8.45726      9.328546
   Not sure  |      8.08446      9.559606
-------------------------------------
```

Univariate graphs