Introduction to Descriptive Statistics

17.871

Spring 2012

Key measures

Describing data

	Moment	**Non-mean based measure**
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

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

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

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

Binary data

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

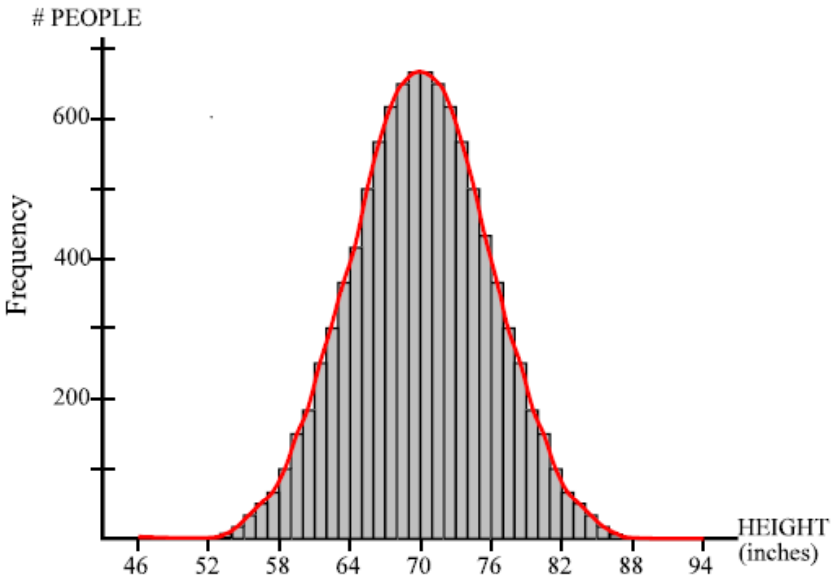
$$s_x^2 = \bar{x}(1-\bar{x}) \Rightarrow s_x = \sqrt{\bar{x}(1-\bar{x})}$$

Example of this, using today's NBC News/Marist Poll in Michigan

Candidate	Pct.
Santorum	35
Romney	37
Paul	13
Gingrich	8
[Unaccounted for]	[7]

- gen santorum = 1 if candidate=="Santorum"
- replace santorum = 0 if candidate~="Santorum"
- the command `summ santorum` produces
- Mean = .35
- Var = .35(1-.35)=.2275
- S.d. = . 4769696

Normal distribution example



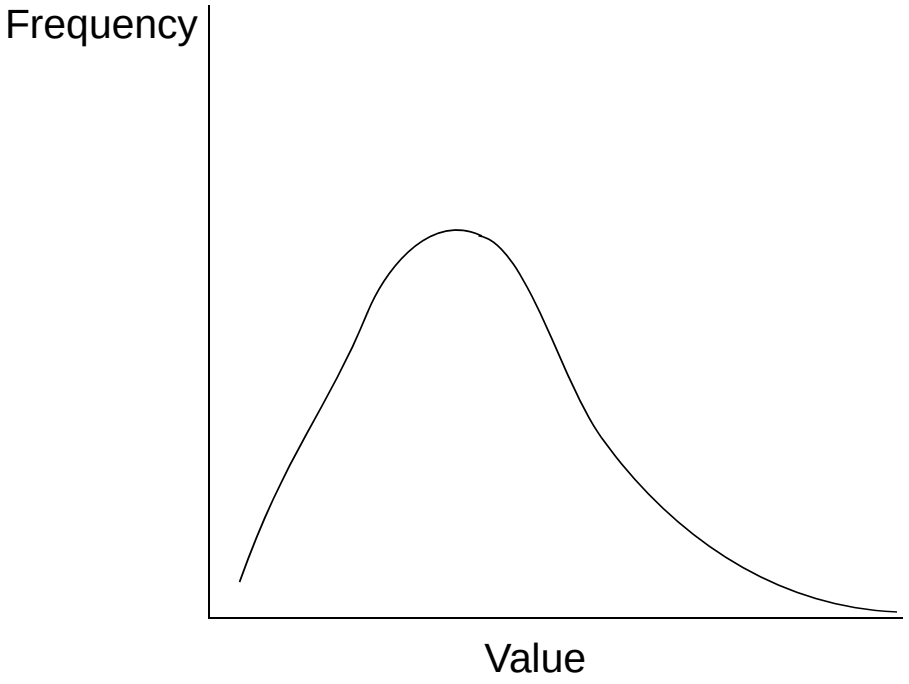
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

- IQ
- SAT
- Height

- “No skew”
- “Zero skew”
- Symmetrical
- Mean = median = mode

Skewness

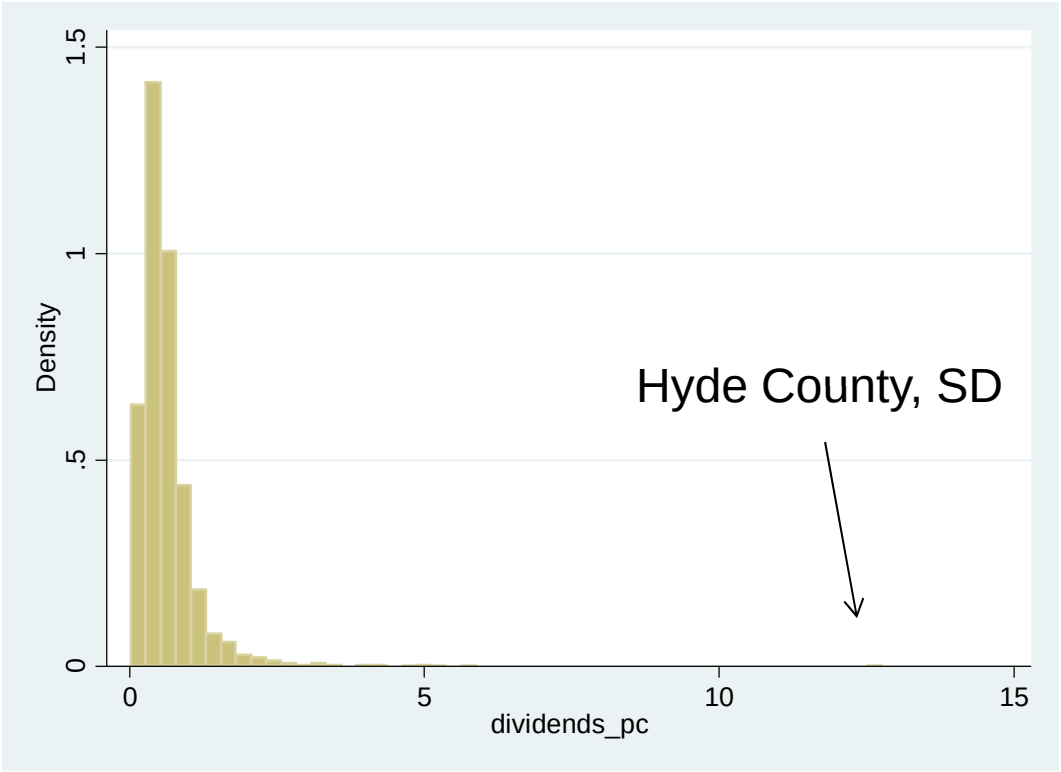
Asymmetrical distribution



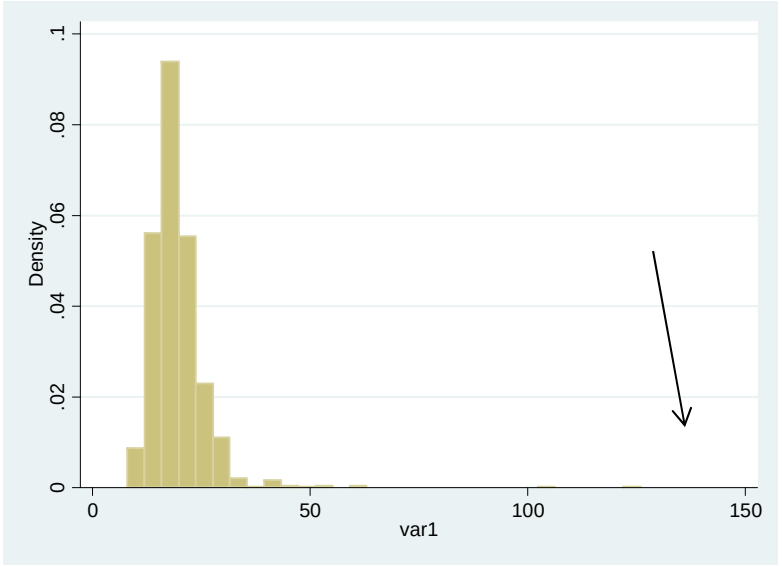
- Income
- Contribution to candidates
- Populations of countries
- “Residual vote” rates

- “Positive skew”
- “Right skew”

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



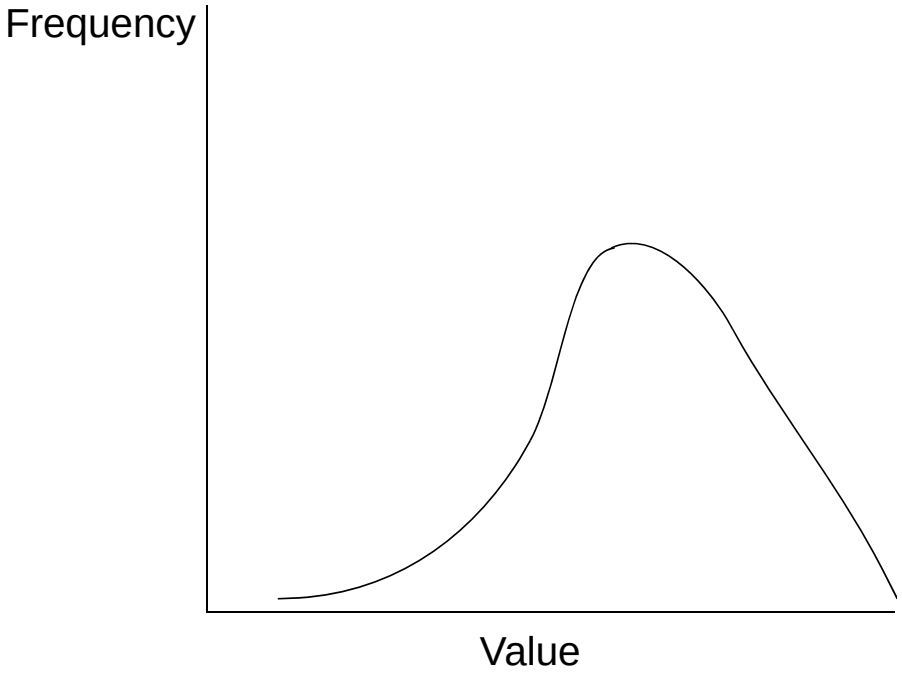
Mitsubishi i-MiEV
(which is supposed to be all electric)



Fuel economy of cars for sale in the US

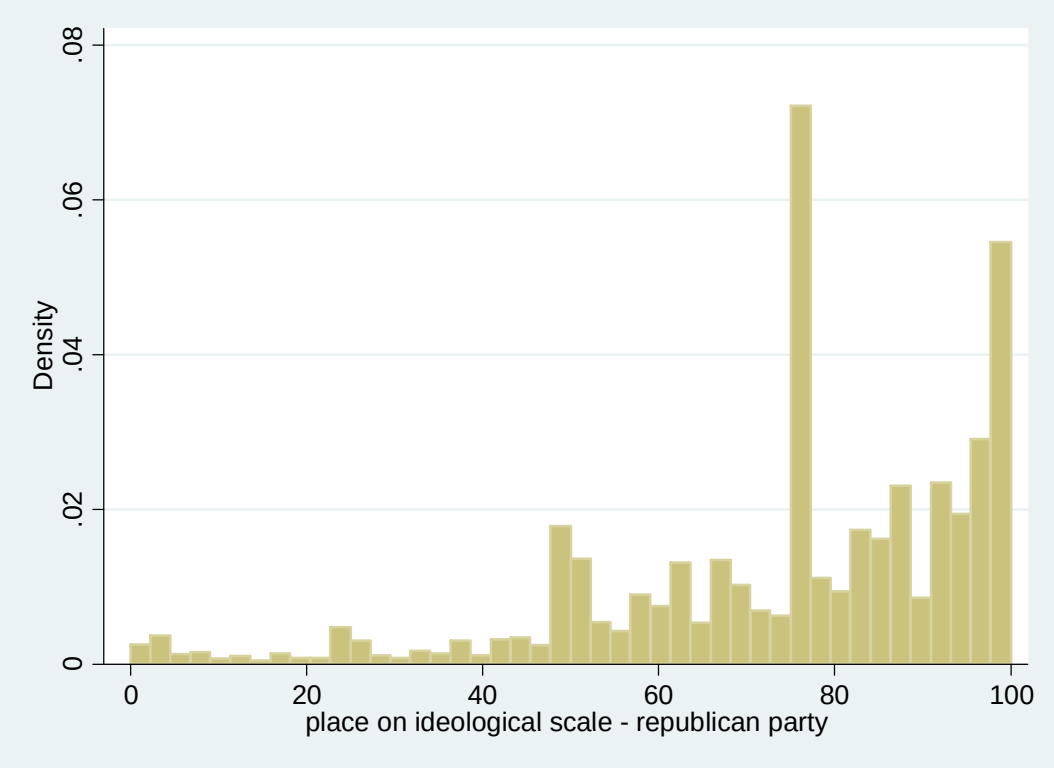
Skewness

Asymmetrical distribution

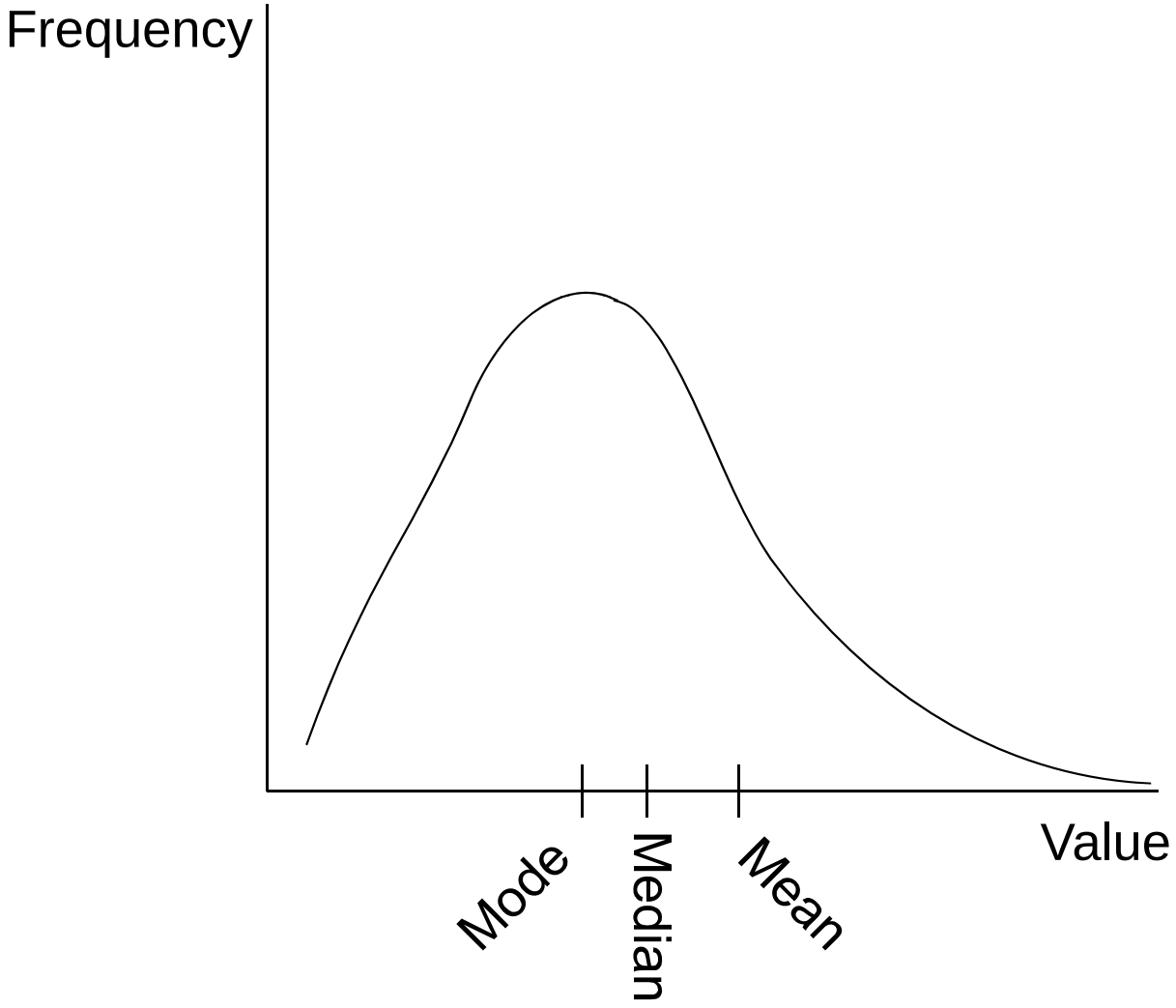


- GPA of MIT students
- “Negative skew”
- “Left skew”

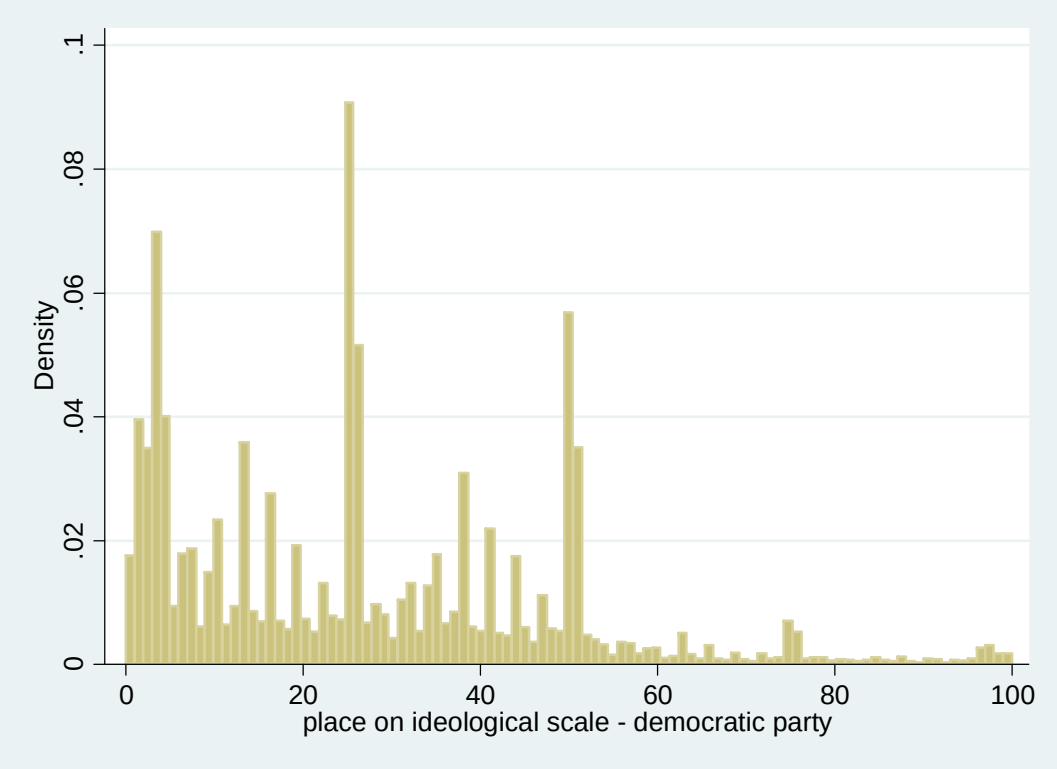
Placement of Republican Party on 100-point scale



Skewness

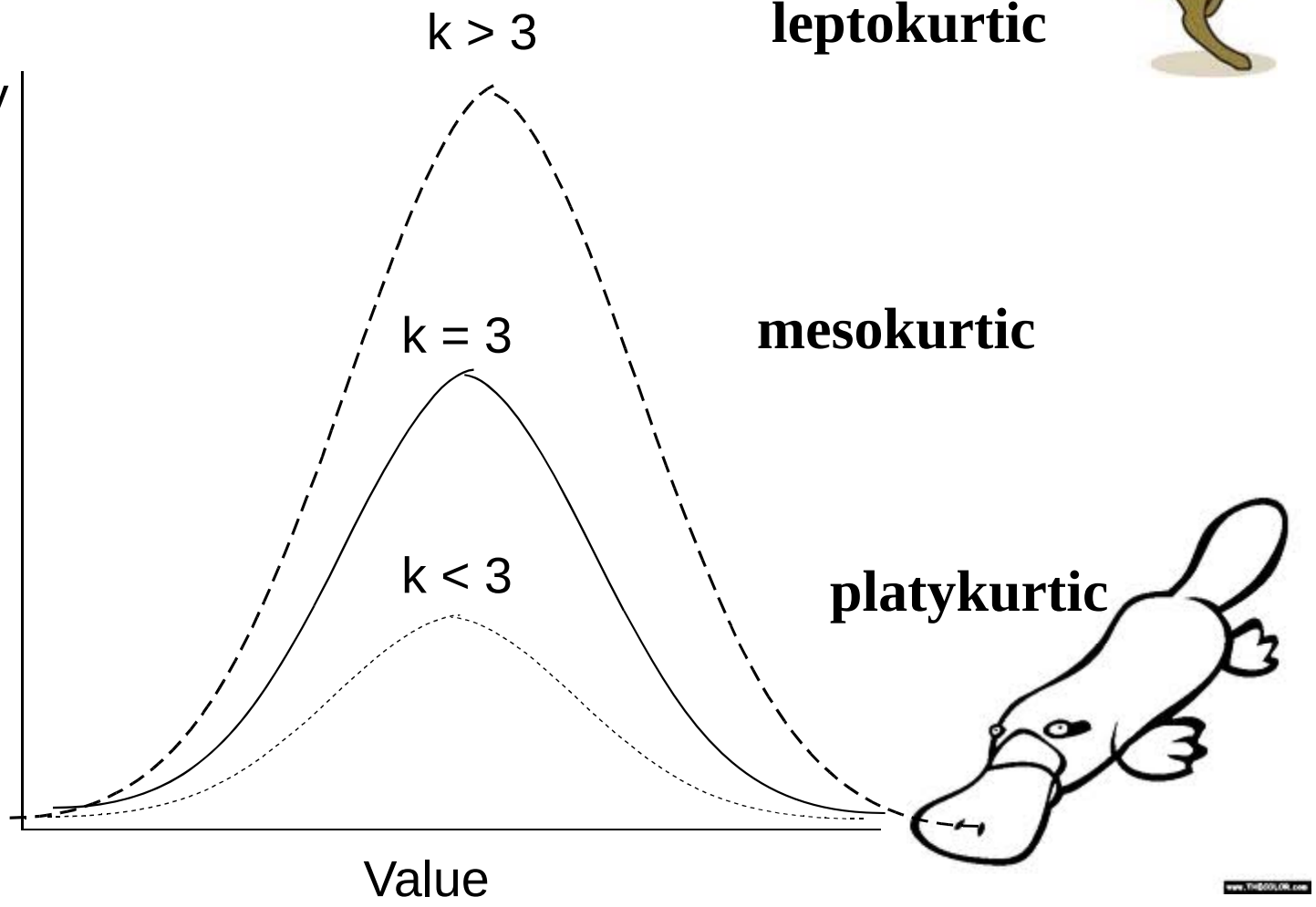


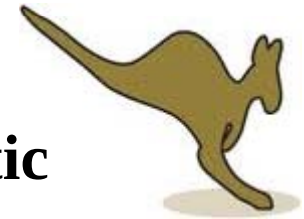
Placement of Republican Party on 100-point scale



Mean = 26.8; median = 25; mode = 25

Kurtosis

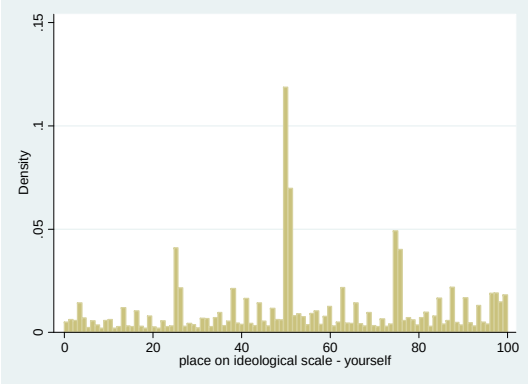


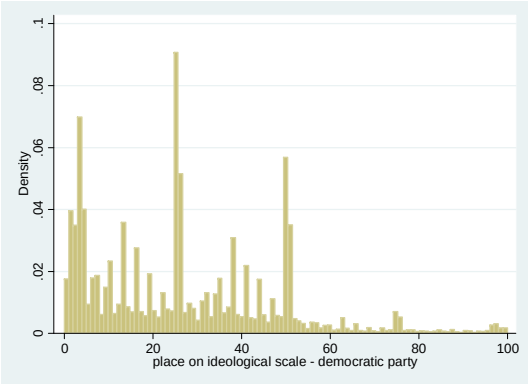


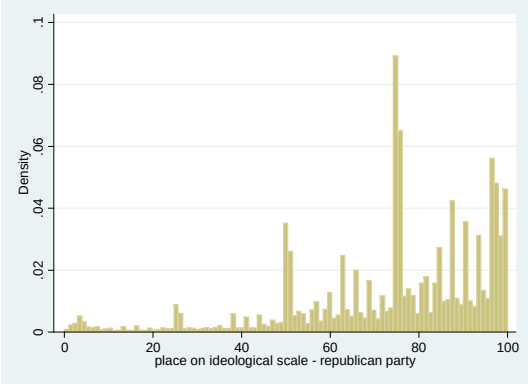
leptokurtic

mesokurtic

platykurtic



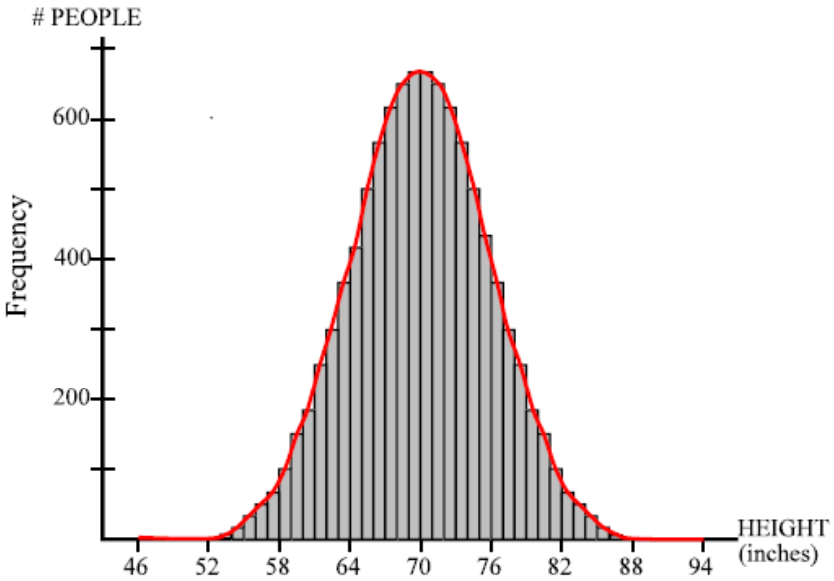




	Mean	s.d.	Skew.	Kurt.
Self-placement	55.1	26.4	-0.14	2.21
Rep. pty.	26.8	21.2	0.87	3.59
Dem. pty	74.7	21.8	-1.18	4.29

Source: Cooperative Congressional Election Study, 2008

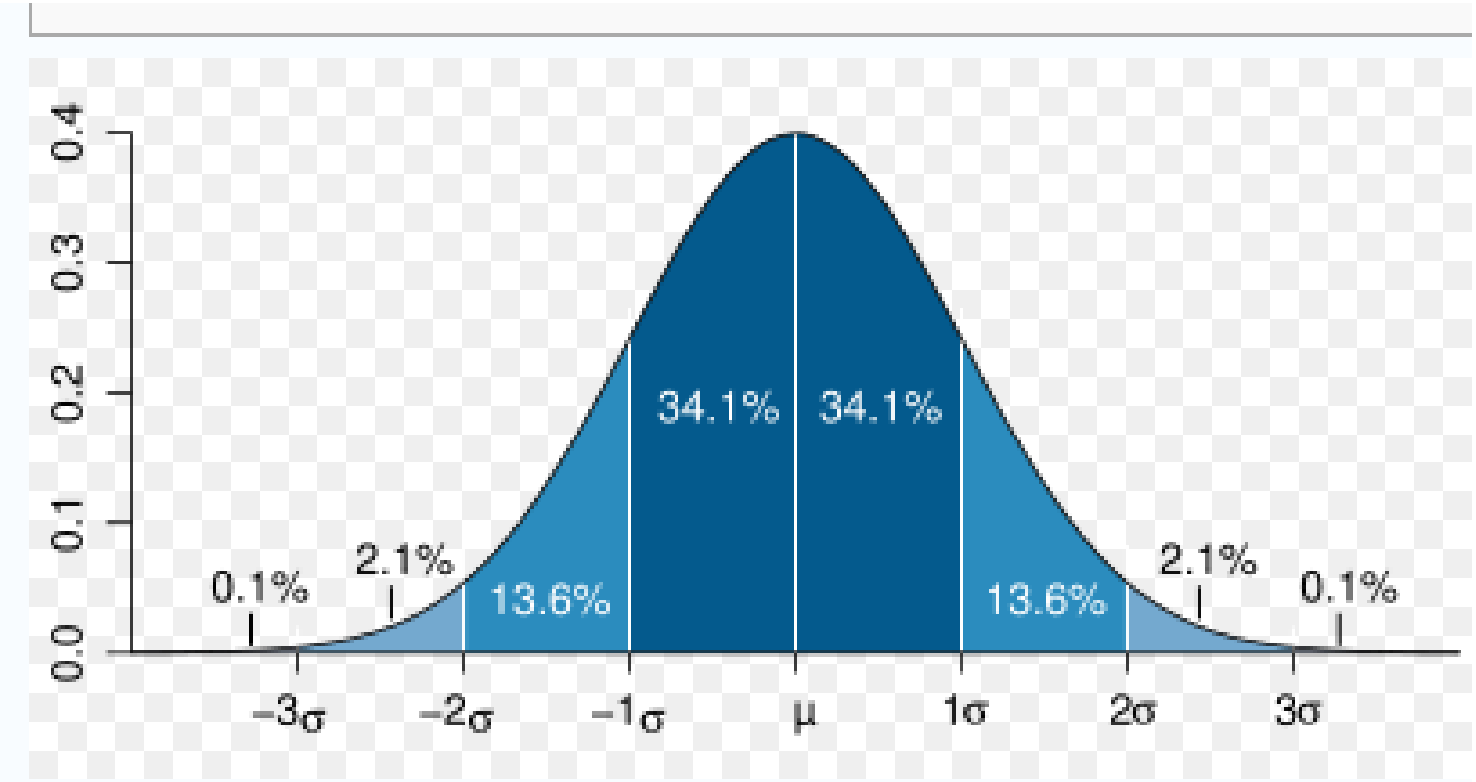
Normal distribution



- Skewness = 0
- Kurtosis = 3

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

More words about the normal curve



The *z*-score
or the
"standardized score"

$$Z = \frac{x - \bar{x}}{\sigma_x}$$

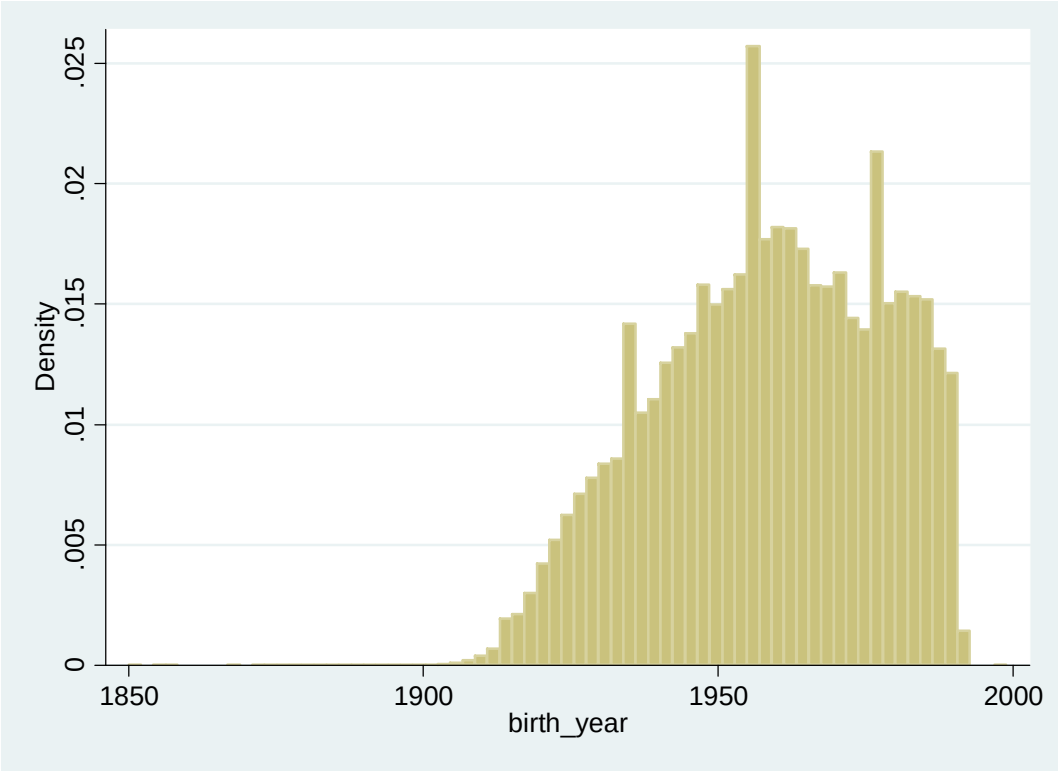
Commands in STATA for univariate statistics

- summarize *varname*
- summarize *varname*, detail
- histogram *varname*, *bin() start() width() density/fraction/frequency normal*
- graph box *varnames*
- tabulate

Example of Florida voters

- Question: does the age of voters vary by race?
- Combine Florida voter extract files, 2008
- gen new_birth_date=date(birth_date,"MDY")
- gen birth_year=year(new_b)
- gen age= 2010-birth_year

Look at distribution of birth year

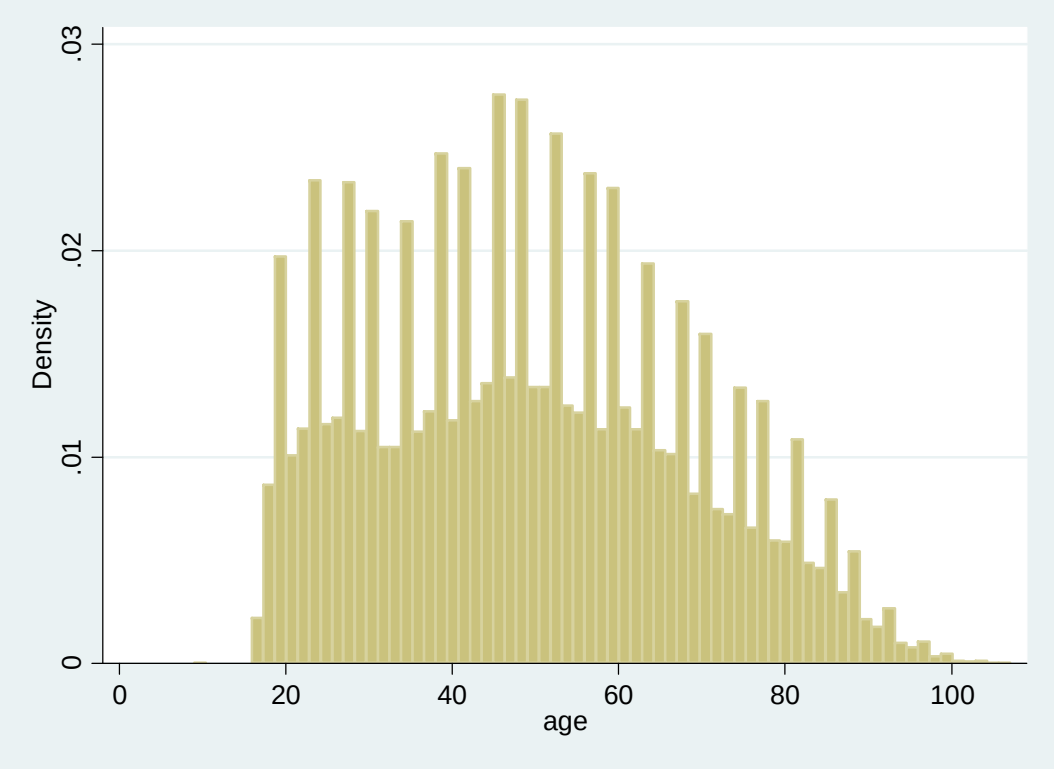


Explore age by voting mode

```
. table race if birth_year>1900,c(mean age)

----------------------
     race |  mean(age)
----------+-----------
        1 |   45.61229
        2 |   42.89916
        3 |    42.6952
        4 |   45.09718
        5 |   52.08628
        6 |   44.77392
        9 |   40.86704
----------------------
```

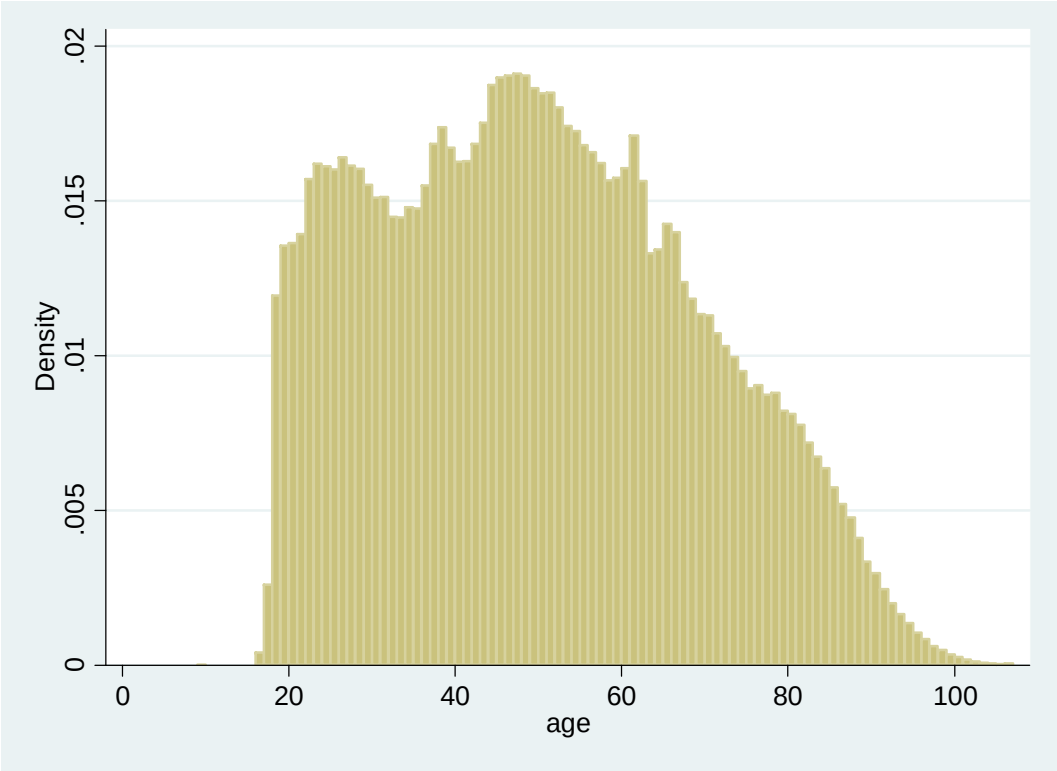
3 = Black
4 = Hispanic
5 = White

Graph birth year




```
. hist age if birth_year>1900
(bin=71, start=9, width=1.3802817)
```

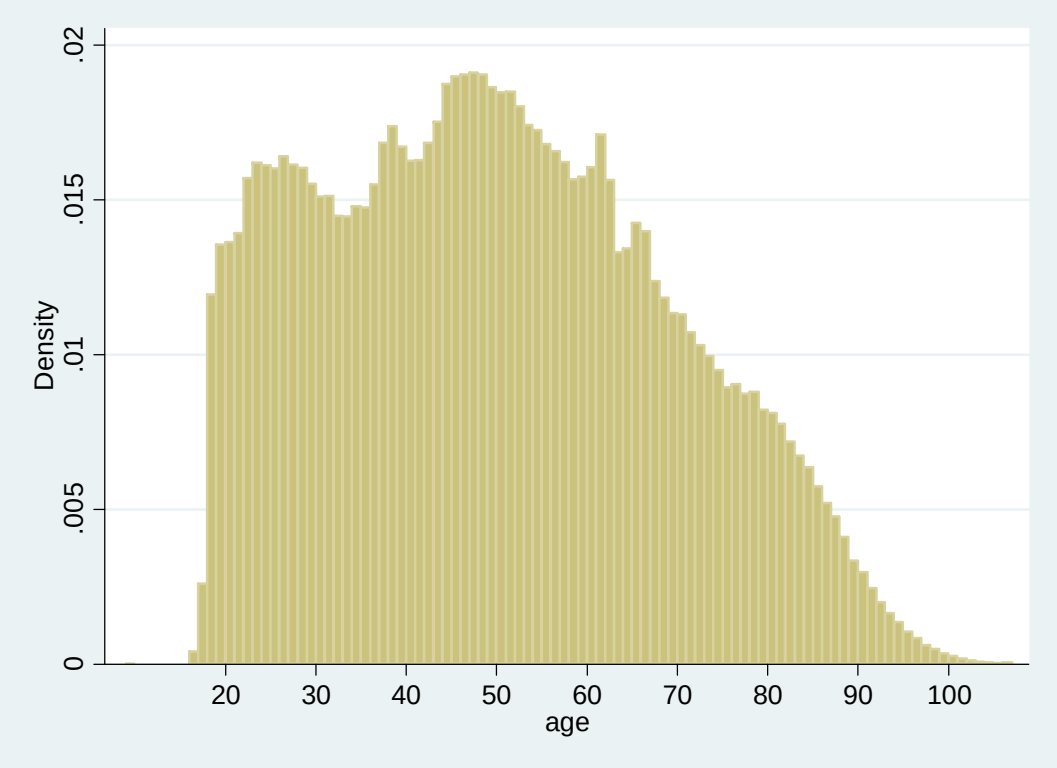
Divide into “bins” so that each bar represents 1 year




```
. hist age if birth_year>1900,width(1)
```

Add ticks at 10-year intervals

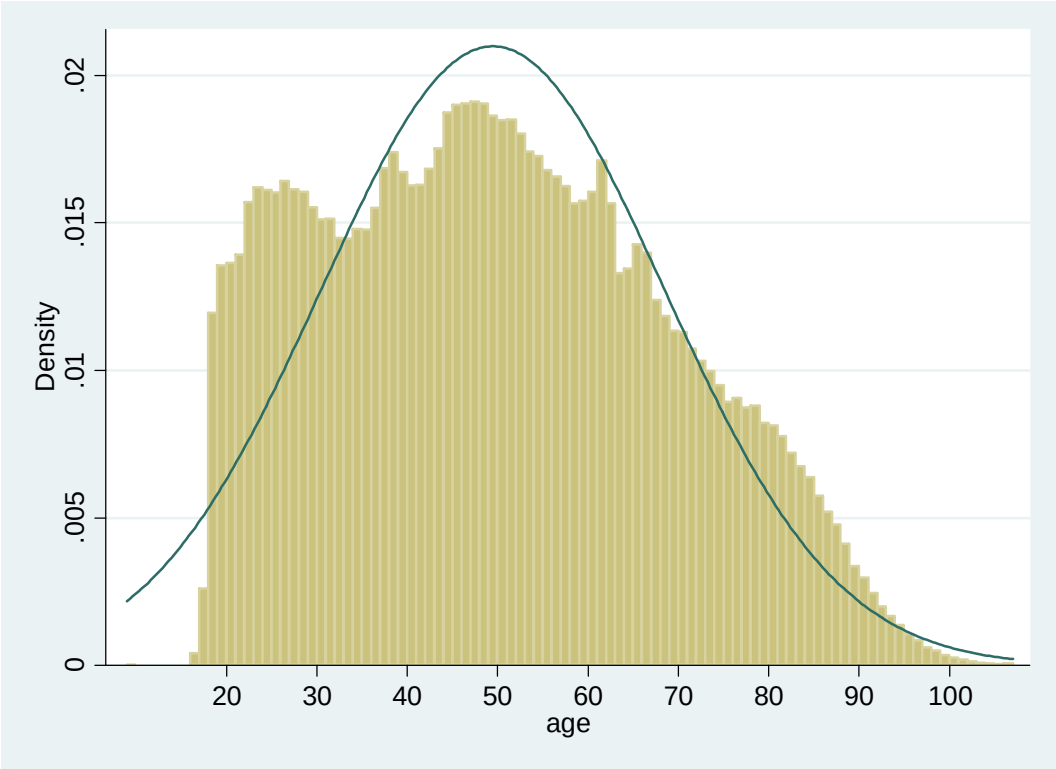
```
histogram totalscore, width(1) xlabel(-.2 (.1) 1)
```



Superimpose the normal curve

(with the same mean and s.d. as the empirical distribution)

```
hist age if birth_year>1900,wid(1) xlabel(20 (10) 100)
   normal
```

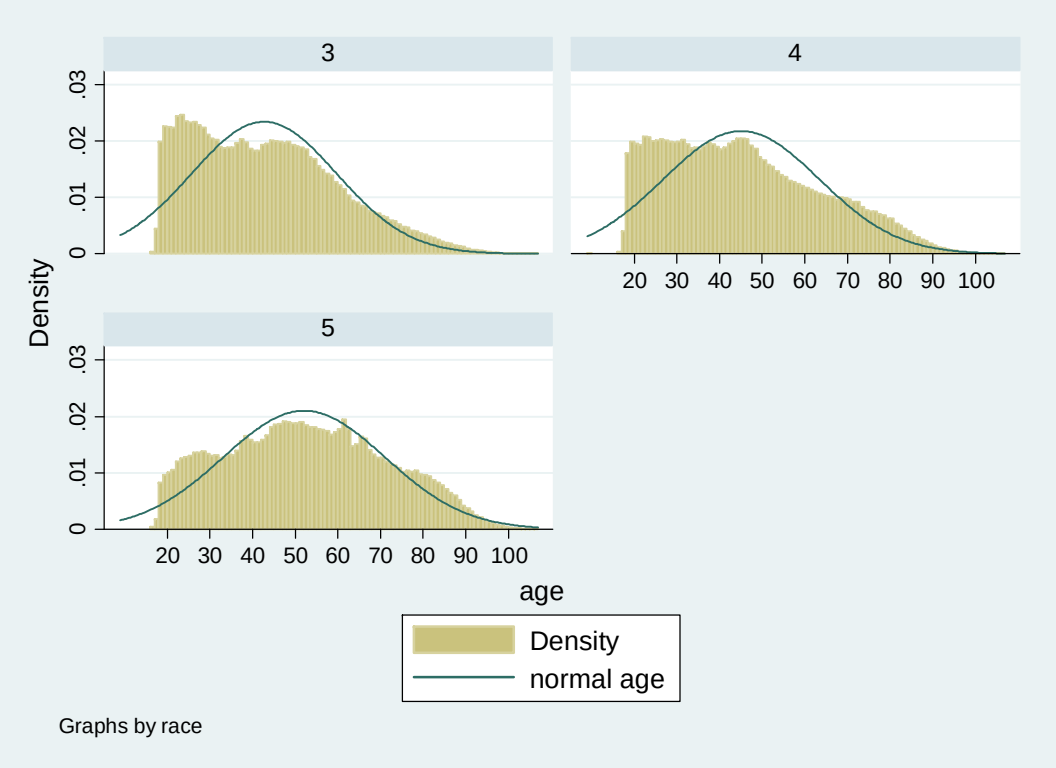


```
. summ age if birth_year>1900,det

                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18             9
 5%           21            16
10%           24            16       Obs            12612114
25%           34            16       Sum of Wgt.    12612114

50%           48                     Mean           49.47549
                      Largest        Std. Dev.      19.01049
75%           63           107
90%           77           107       Variance       361.3986
95%           83           107       Skewness       .2629496
99%           91           107       Kurtosis       2.222442
```

Histograms by race

```
hist age if birth_year>1900&race>=3&race<=5,wid(1)
   xlabel(20 (10) 100) normal by(race)
```

3 = Black
4 = Hispanic
5 = White

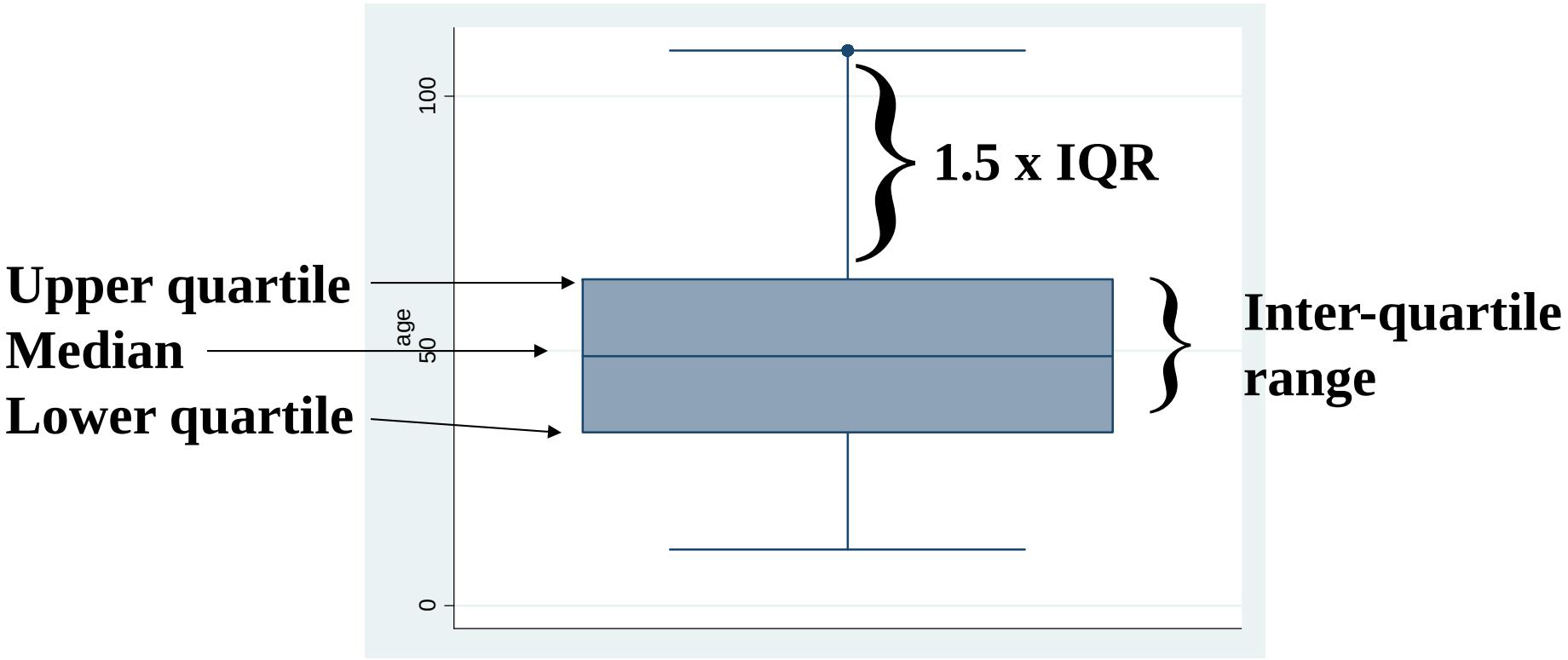


Main issues with histograms

- Proper level of aggregation
- Non-regular data categories

Draw the previous graph with a box plot

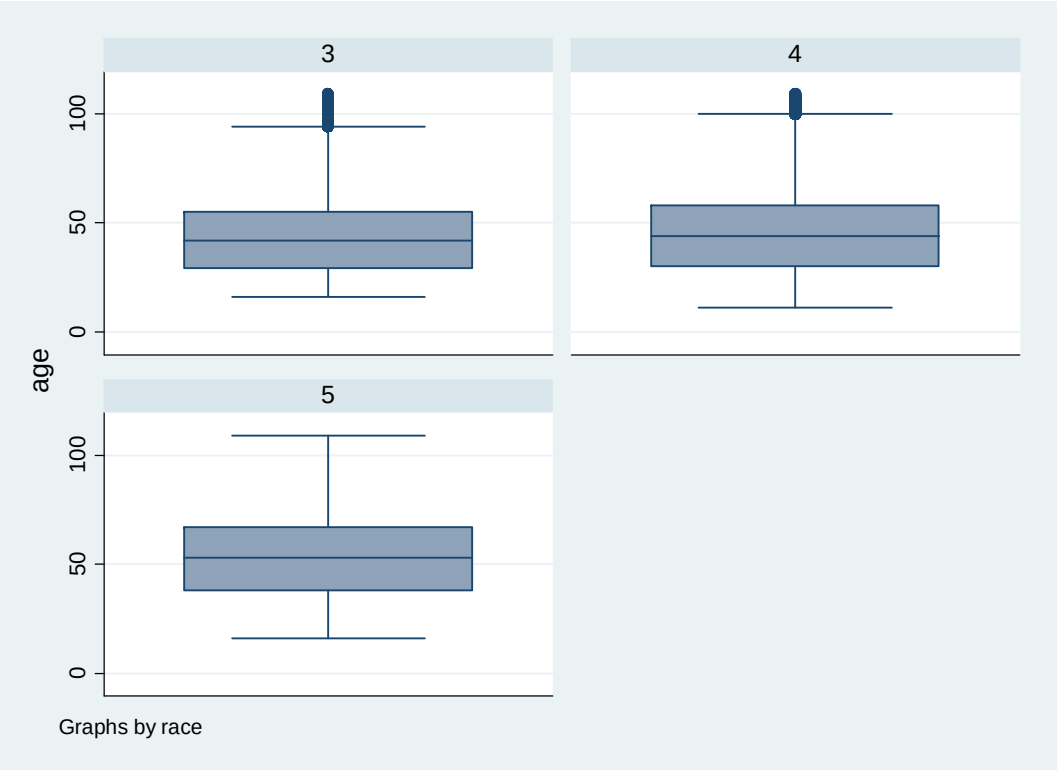
```
graph box age if birth_year>1900
```



Draw the box plots for the different races

```
graph box age if birth_year>1900&race>=3&race<=5,by(race)
```

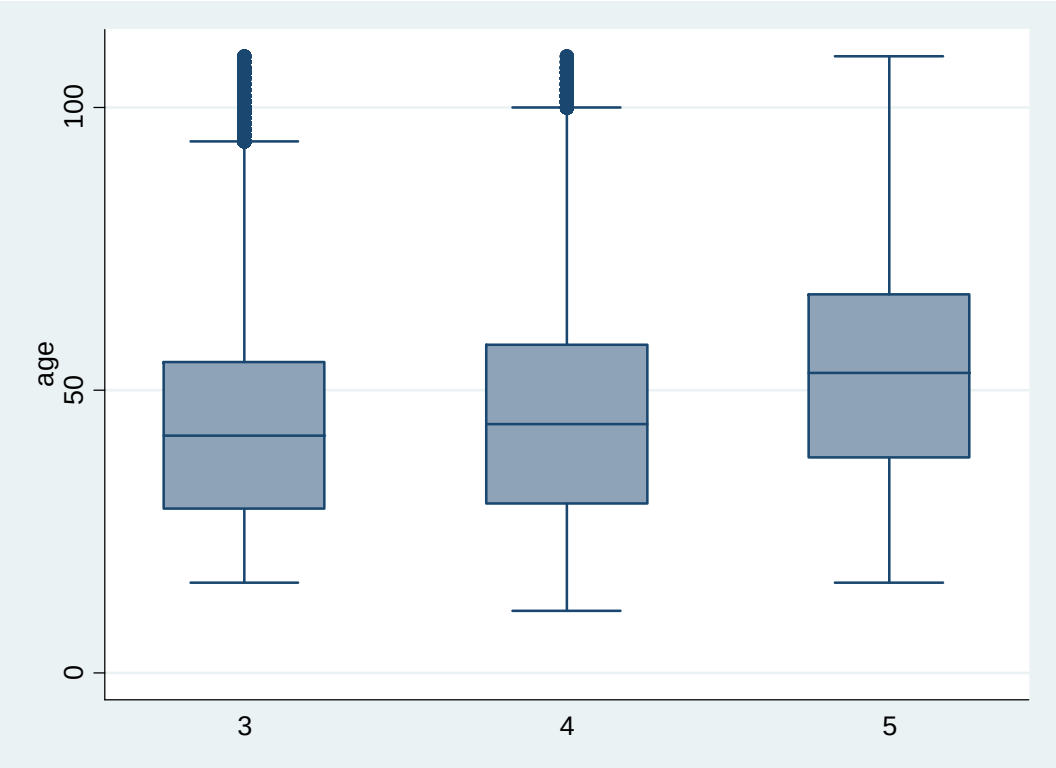
3 = Black
4 = Hispanic
5 = White



Draw the box plots for the different races using “over” option

```
graph box age if birth_year>1900&race>=3&race<=5,over(race)
```

3 = Black
4 = Hispanic
5 = White



A note about histograms with unnatural categories

From the Current Population Survey (2000), Voter and Registration Survey

How long (have you/has name) lived at this address?

-9 No Response
-3 Refused
-2 Don't know
-1 Not in universe
1 Less than 1 month
2 1-6 months
3 7-11 months
4 1-2 years
5 3-4 years
6 5 years or longer

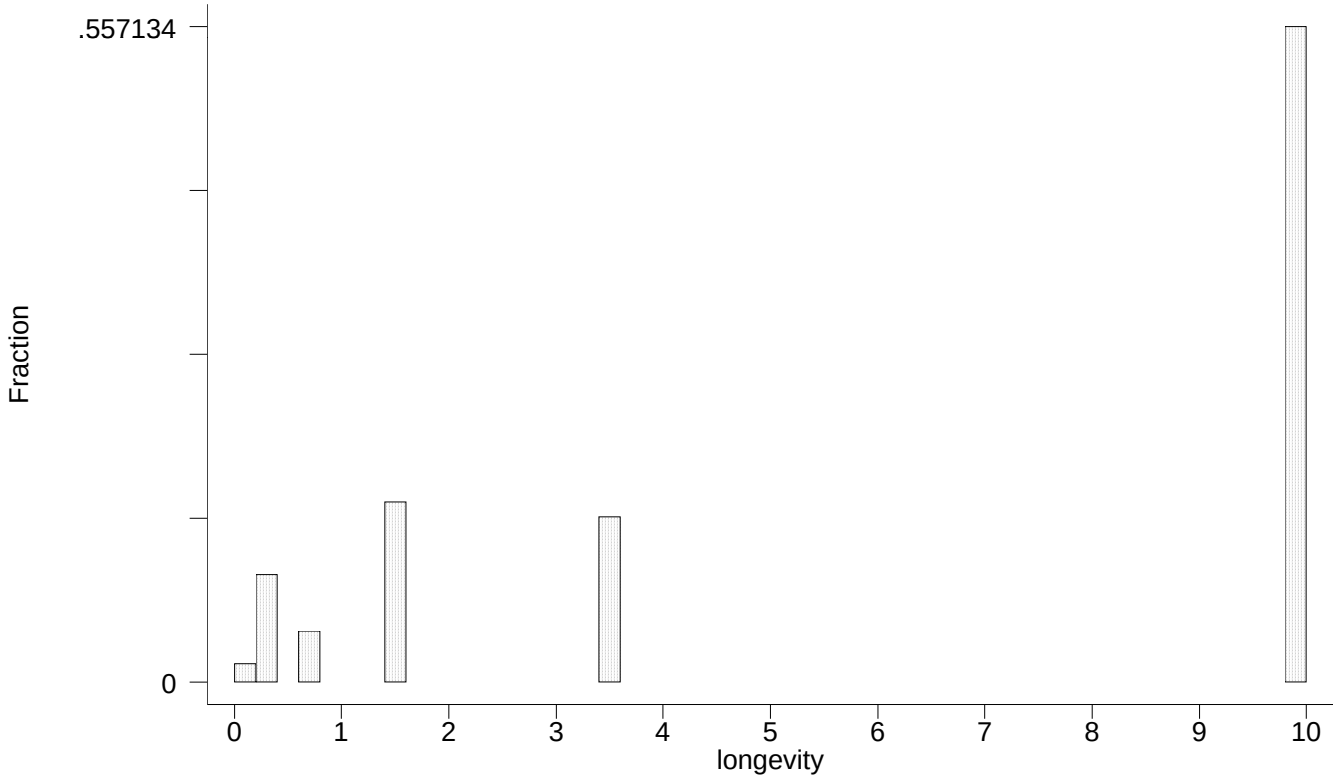
Solution, Step 1
Map artificial category onto "natural" midpoint

-9 No Response → missing
-3 Refused → missing
-2 Don't know → missing
-1 Not in universe → missing
1 Less than 1 month → 1/24 = 0.042
2 1-6 months → 3.5/12 = 0.29
3 7-11 months → 9/12 = 0.75
4 1-2 years → 1.5
5 3-4 years → 3.5
6 5 years or longer → 10 (arbitrary)

recode live_length (min/-1 =.)(1=.042)(2=.29)(3=.75)(4=1.5)(5=3.5)(6=10)

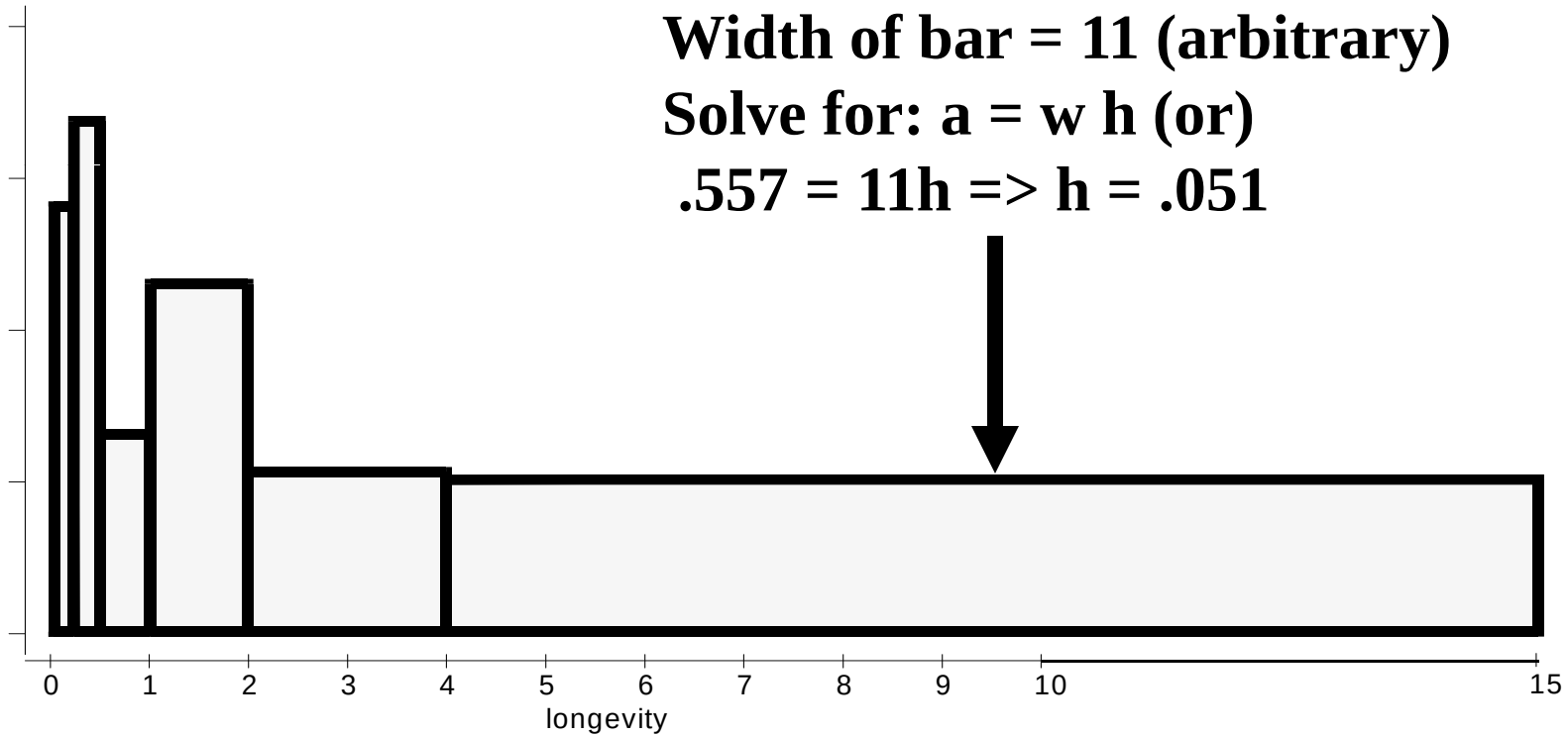
Graph of recoded data

```
histogram longevity, fraction
```



Density plot of data

Total area of last bar = .557
Width of bar = 11 (arbitrary)
Solve for: a = w h (or)
.557 = 11h => h = .051

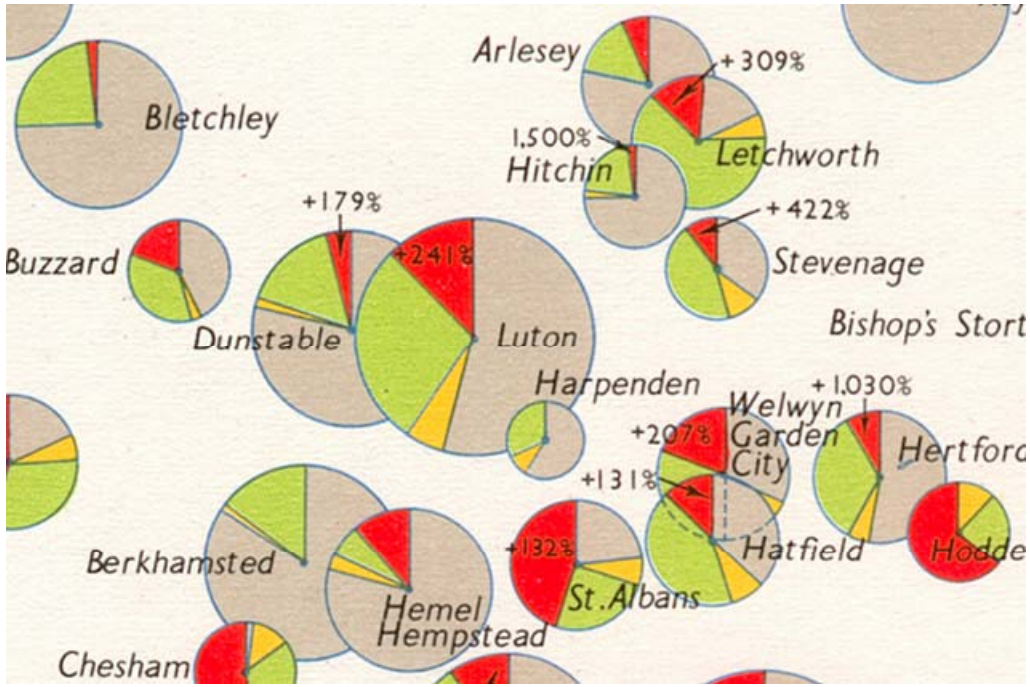


Density plot template

Category	Fraction	X-min	X-max	X-length	Height (density)
< 1 mo.	.0156	0	1/12	.082	.19*
1-6 mo.	.0909	1/12	½	.417	.22
7-11 mo.	.0430	½	1	.500	.09
1-2 yr.	.1529	1	2	1	.15
3-4 yr.	.1404	2	4	2	.07
5+ yr.	.5571	4	15	11	.05

*** = .0156/.082**

Three words about pie charts: don’t use them



So, what’s wrong with them

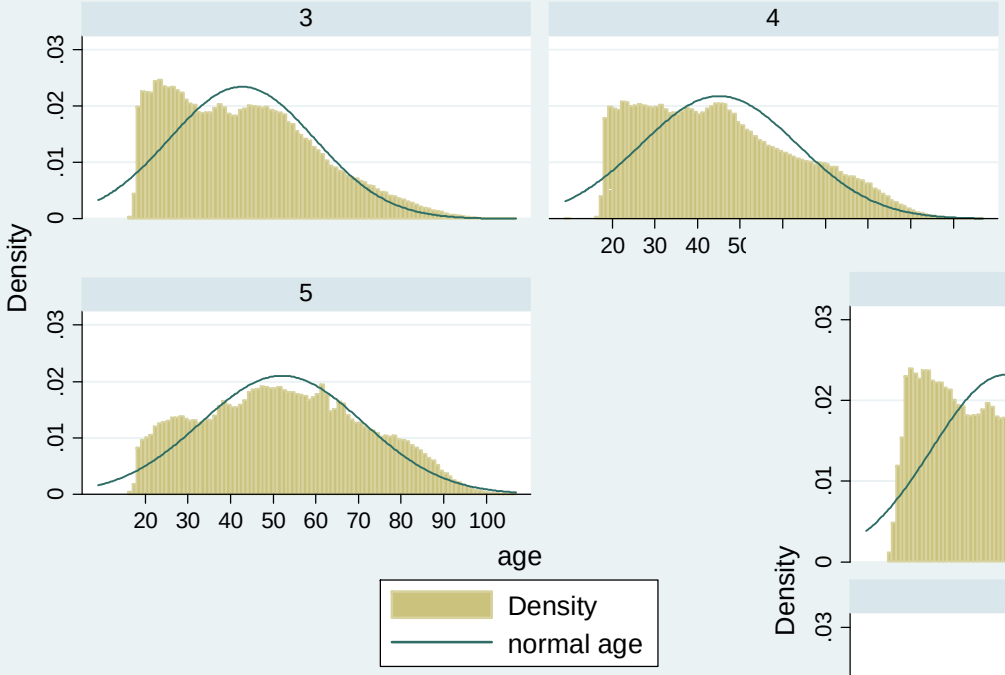
- For non-time series data, hard to get a comparison among groups; the eye is very bad in judging relative size of circle slices
- For time series, data, hard to grasp cross-time comparisons

Some words about graphical presentation

- Aspects of graphical integrity (following Edward Tufte, *Visual Display of Quantitative Information*)
  - Main point should be readily apparent
  - Show as much data as possible
  - Write clear labels on the graph
  - Show data variation, not design variation

There is a difference between graphs in research and publication

Publish OK



Do not publish