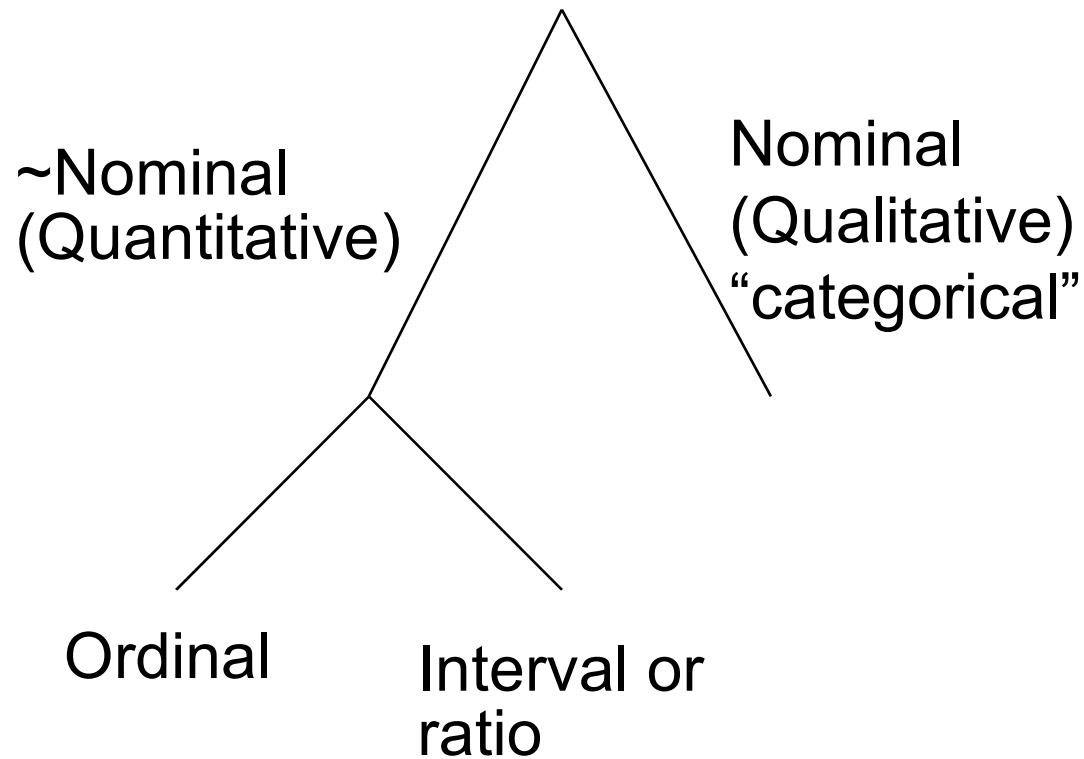




Introduction to Descriptive Statistics

17.871

Types of Variables



Describing data

	Moment	Non-mean based measure
Center	Mean	Mode, median
Spread	Variance (standard deviation)	Range, Interquartile range
Skew	Skewness	--
Peaked	Kurtosis	--

Population vs. Sample Notation

Population	Vs	Sample
Greeks		Romans
μ, σ, β		s, b

Mean

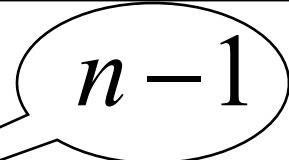
$$\frac{\sum_{i=1}^n x_i}{n} \equiv \mu \equiv \bar{X}$$

Variance, Standard Deviation

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

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

Variance, S.D. of a Sample

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


Degrees of freedom

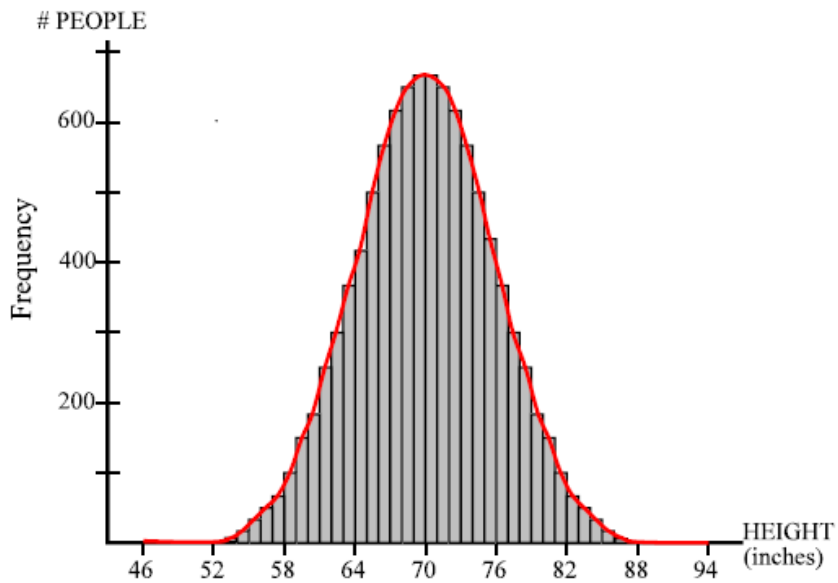
$$\sqrt{\sum_{i=1}^n \frac{(x_i - \mu)^2}{n-1}} \equiv s$$

Binary data

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

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

Example

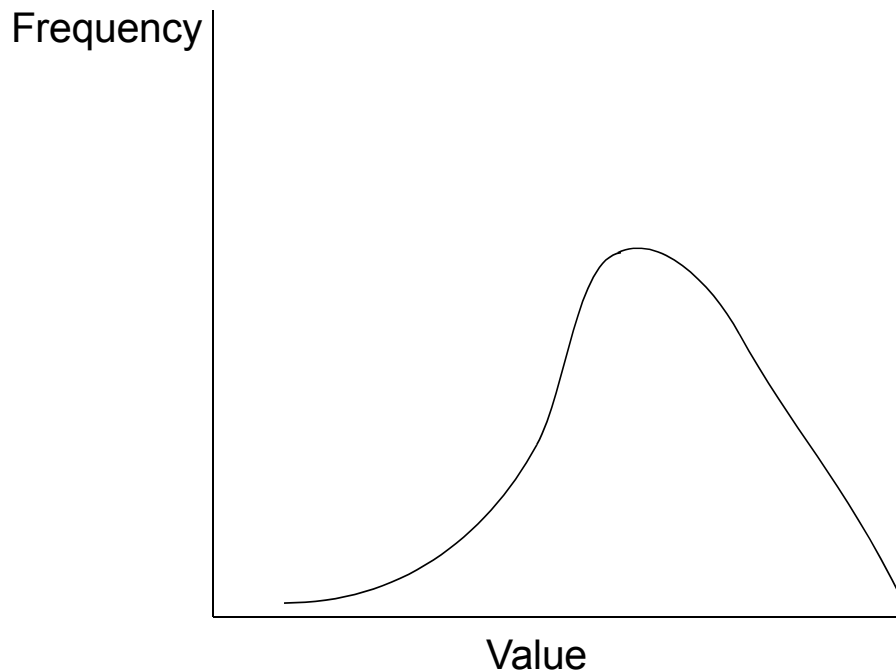


- IQ
- SAT
- Height

- “No skew”
- “Zero skew”
- Symmetrical

Skewness

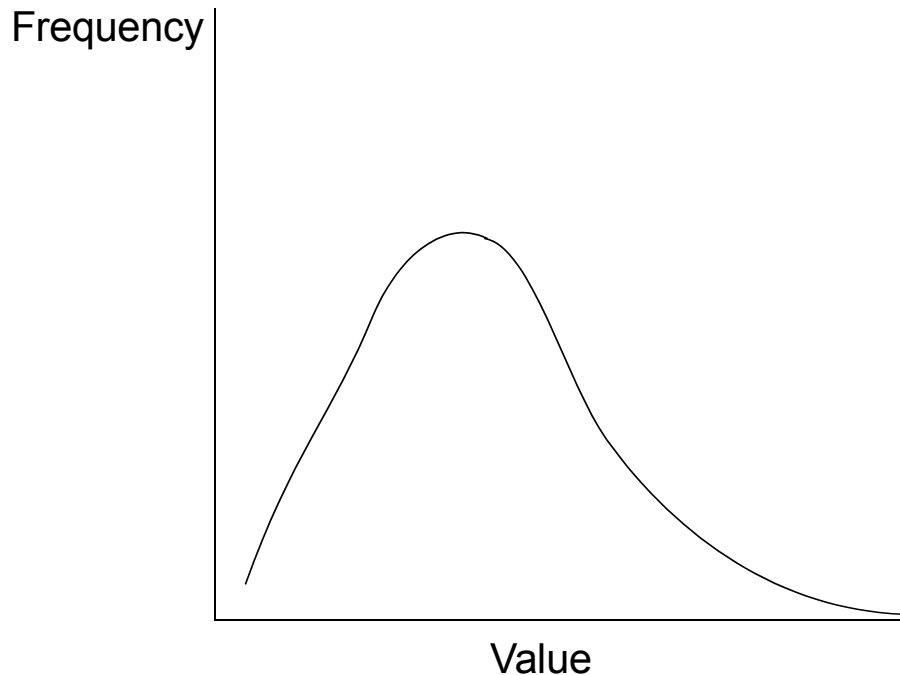
Asymmetrical distribution



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

Skewness

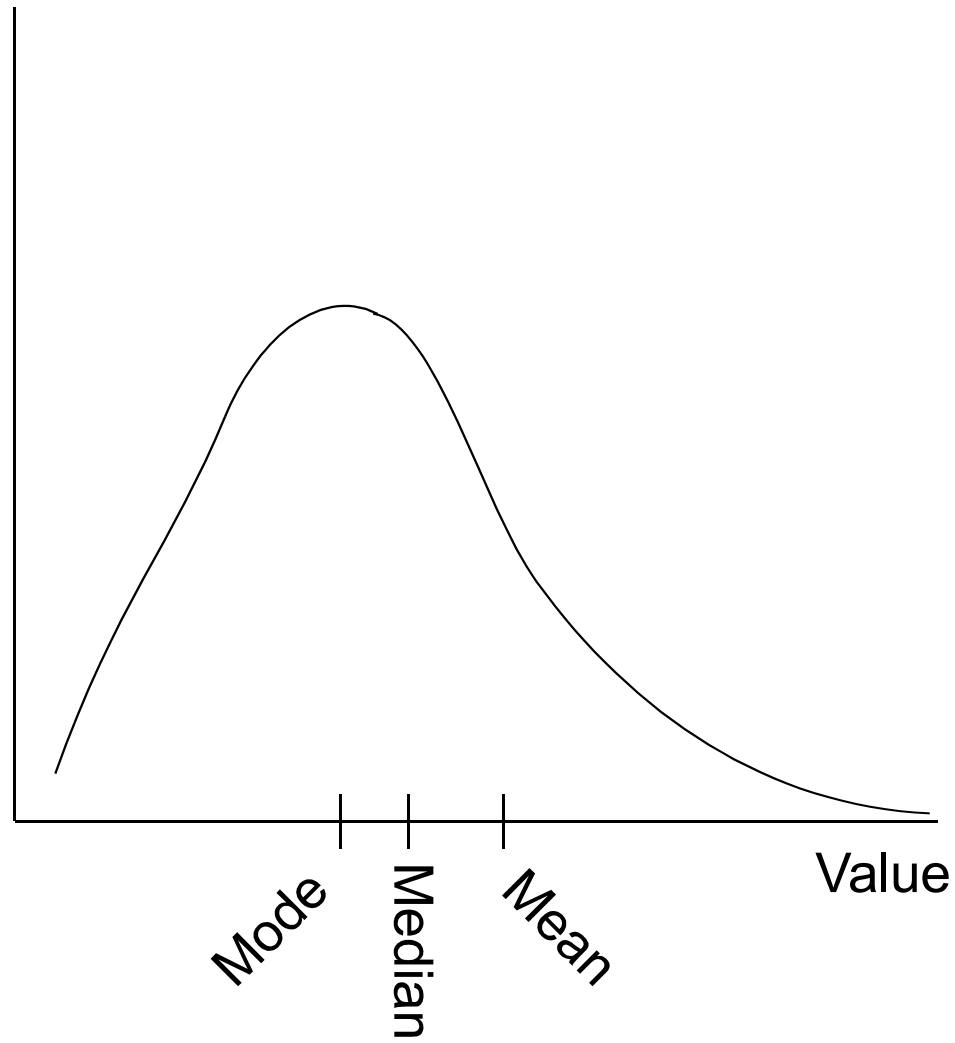
(Asymmetrical distribution)



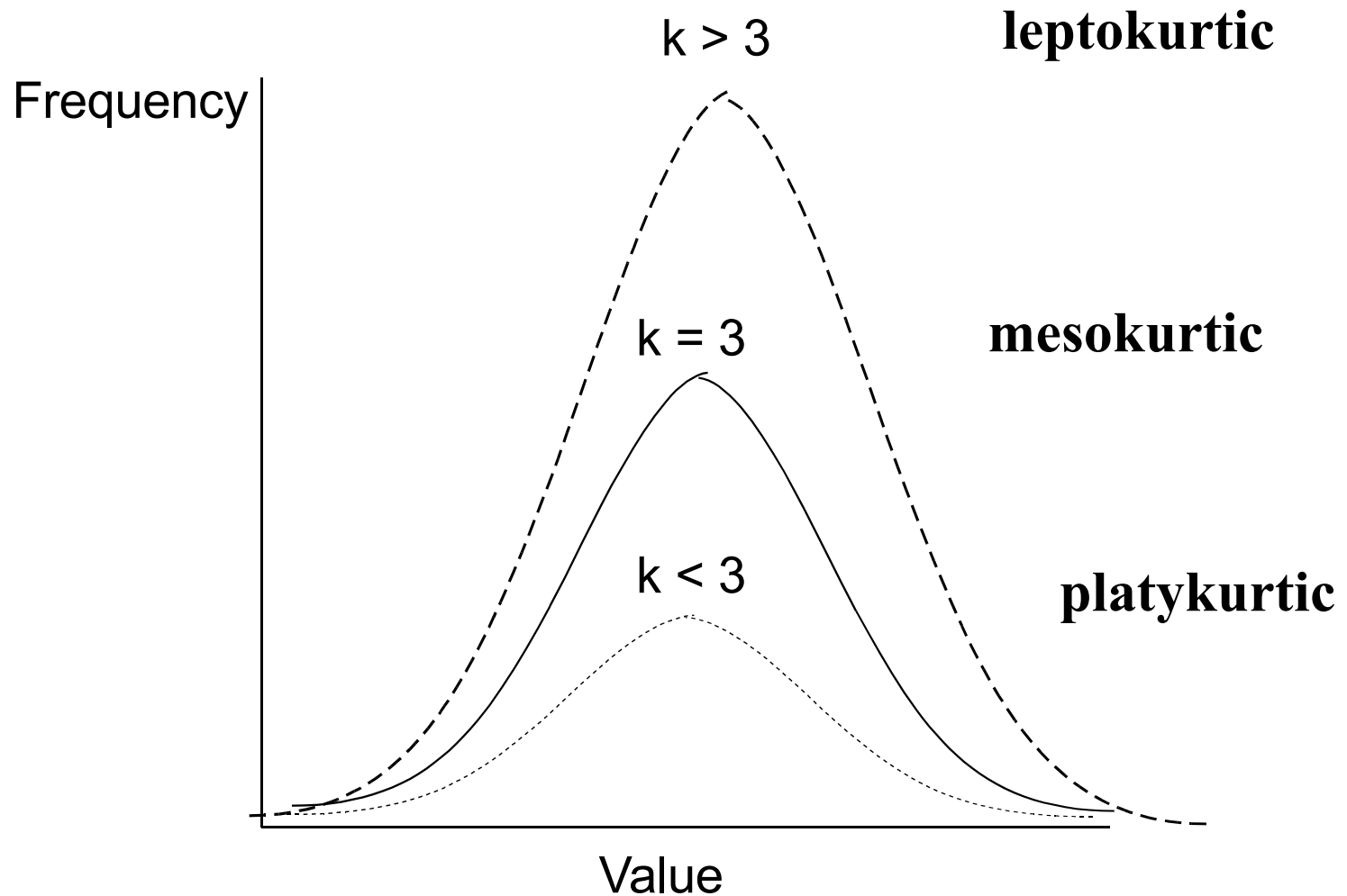
- Income
- Contribution to candidates
- Populations of countries
- “Residual vote” rates
- “Positive skew”
- “Right skew”

Skewness

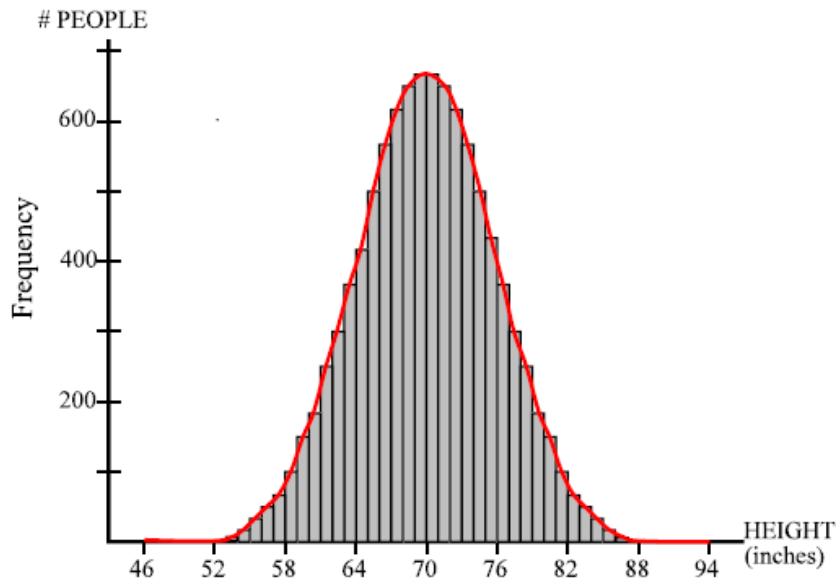
Frequency



Kurtosis



Normal distribution



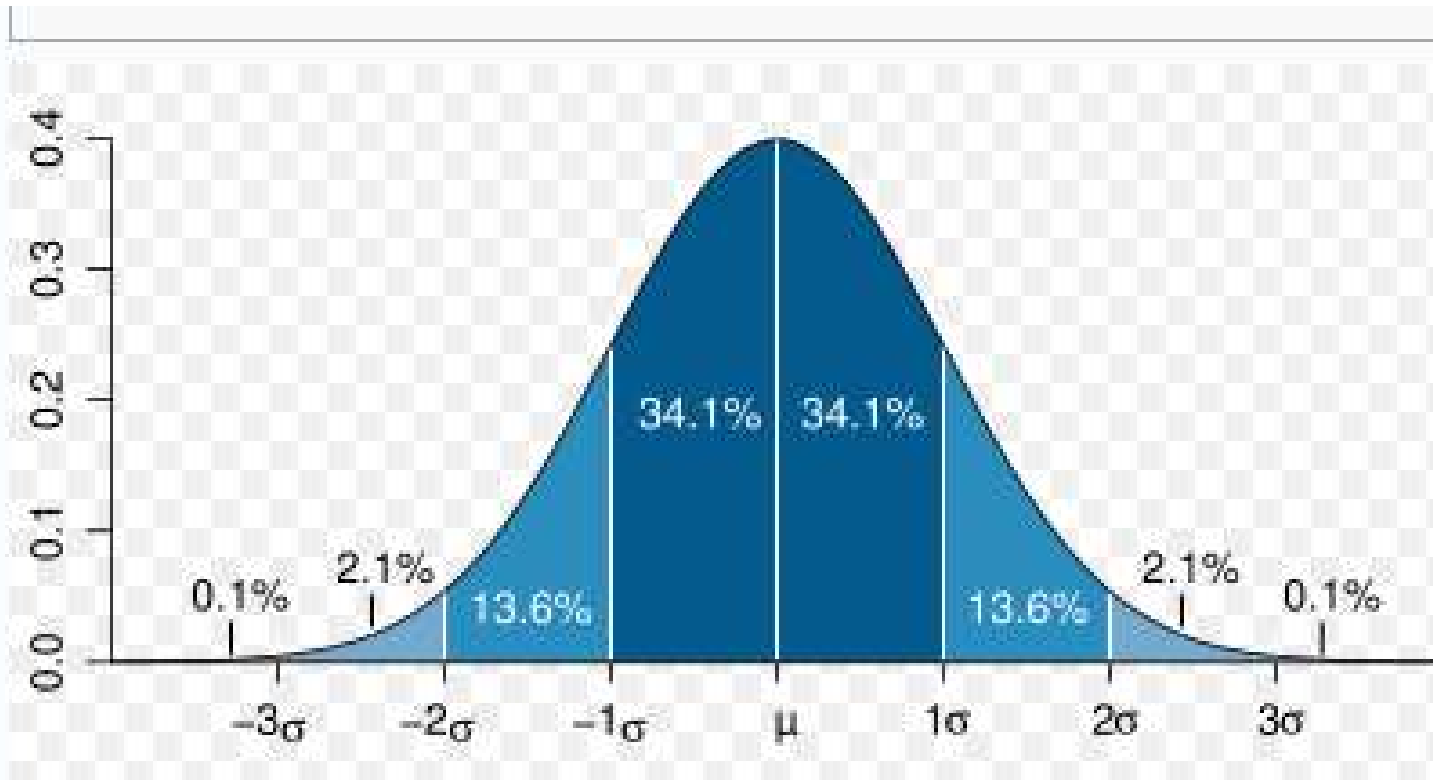
- Skewness = 0
- Kurtosis = 3

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

The z-score
or the
“standardized score”

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

More words about the normal curve



Commands in STATA for getting univariate statistics

- summarize *varname*
- summarize *varname*, detail
- histogram *varname*, *bin()* *start()* *width()*
density/fraction/frequency normal
- graph box *varnames*
- tabulate [NB: compare to table]

Example of Sophomore Test Scores

- High School and Beyond, 1980: A Longitudinal Survey of Students in the United States (ICPSR Study 7896)
- *totalscore* = % of questions answered correctly minus penalty for guessing
- *recodedtype* = (1=public school, 2=religious private, 3 = non-sectarian private)

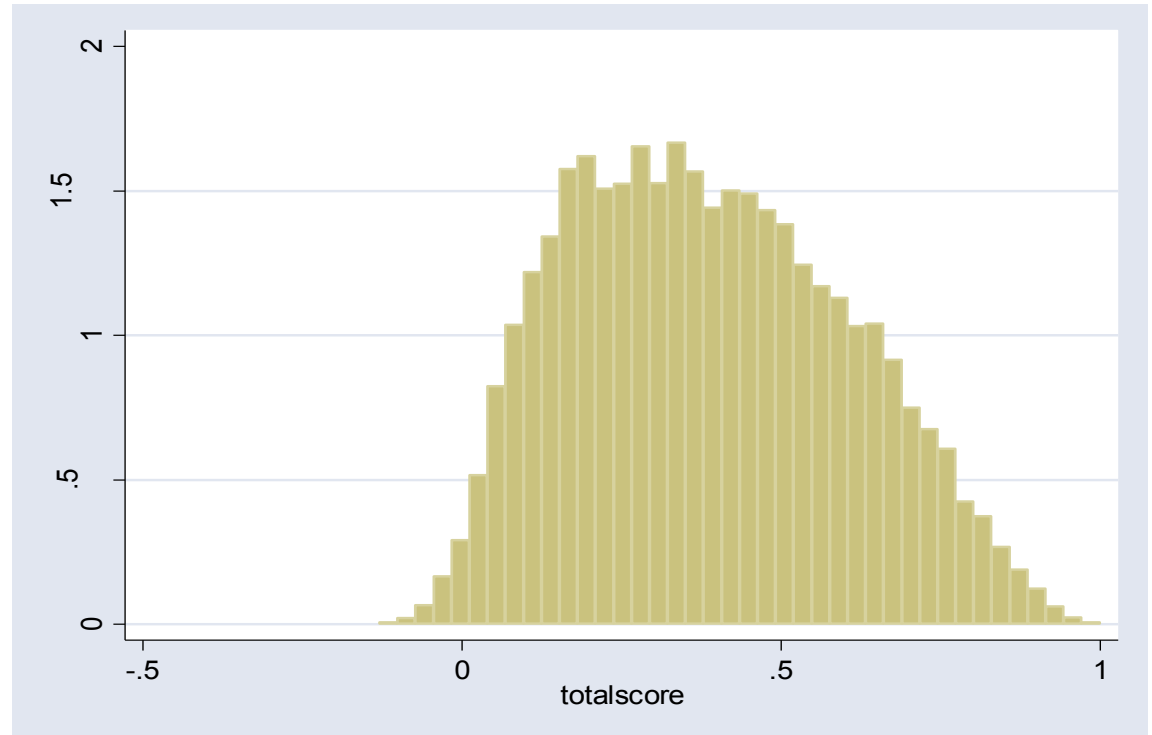
Explore totalscore some more

```
. table recodedtype,c(mean totalscore)
```

```
-----  
recodedty |  
pe         | mean(totals~e)  
-----+-----  
          1 |          .3729735  
          2 |          .4475548  
          3 |          .5898883  
-----
```

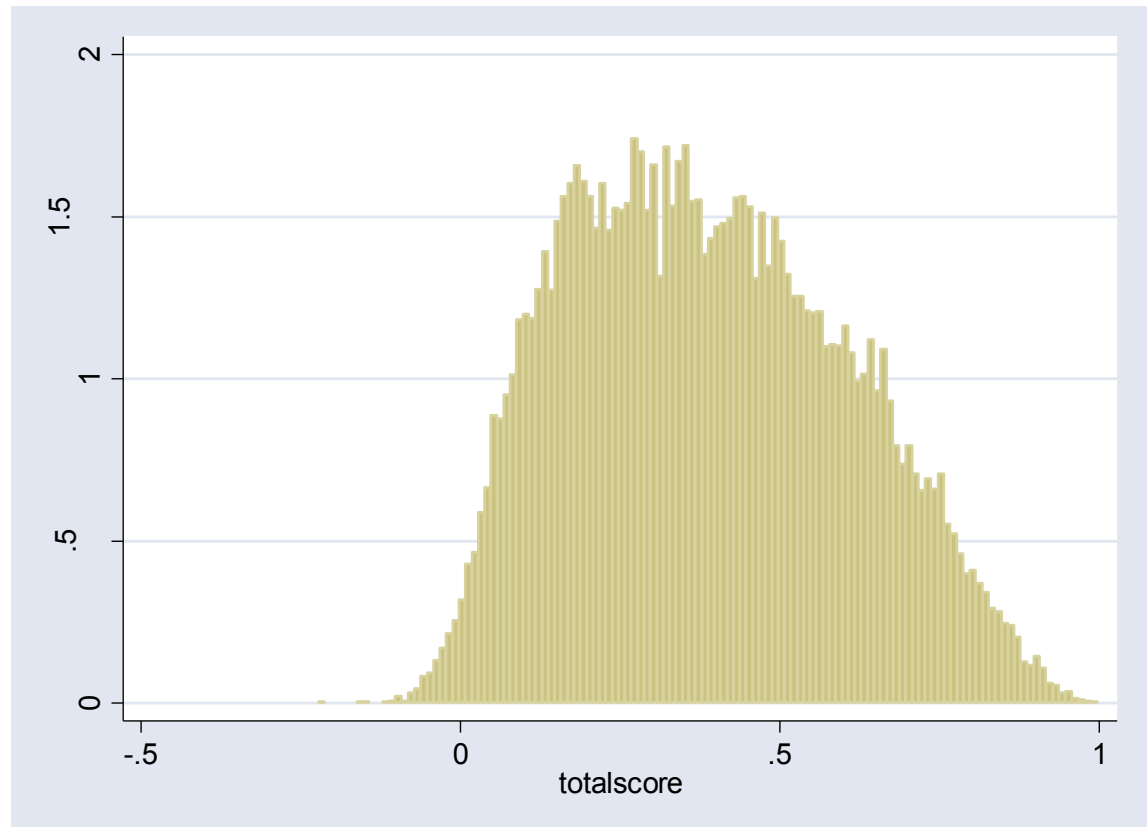
Graph totalscore

```
. hist totalscore
```



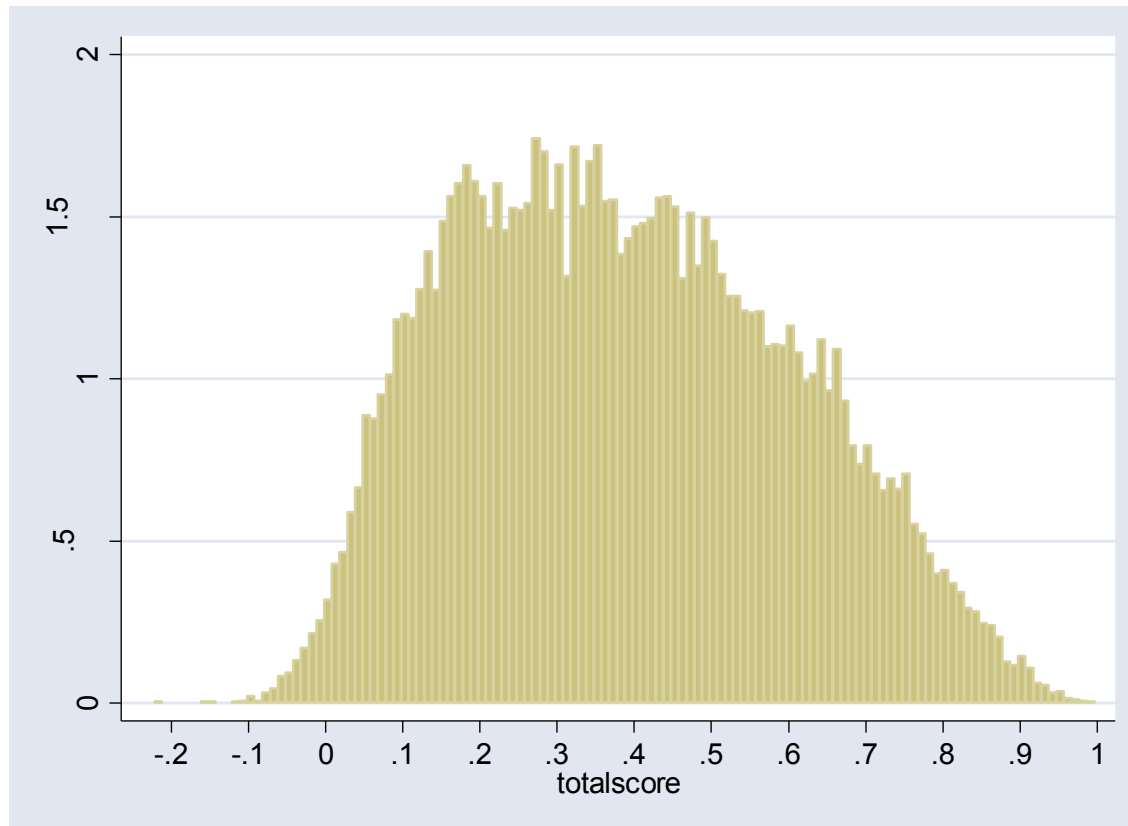
Divide into “bins” so that each bar represents 1% correct

- `hist totalscore,width(.01)`
- **`(bin=124, start=-.24209334, width=.01)`**



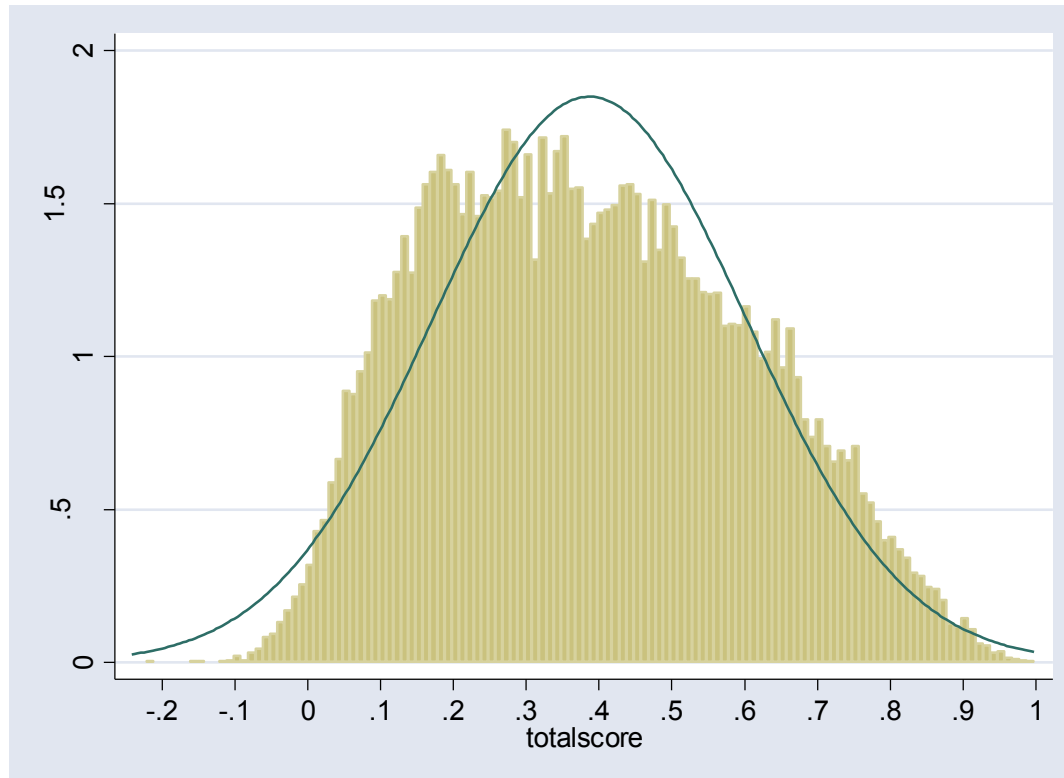
Add ticks at each 10% mark

```
histogram totalscore, width(.01) xlabel(-.2 (.1) 1)  
(bin=124, start=-.24209334, width=.01)
```



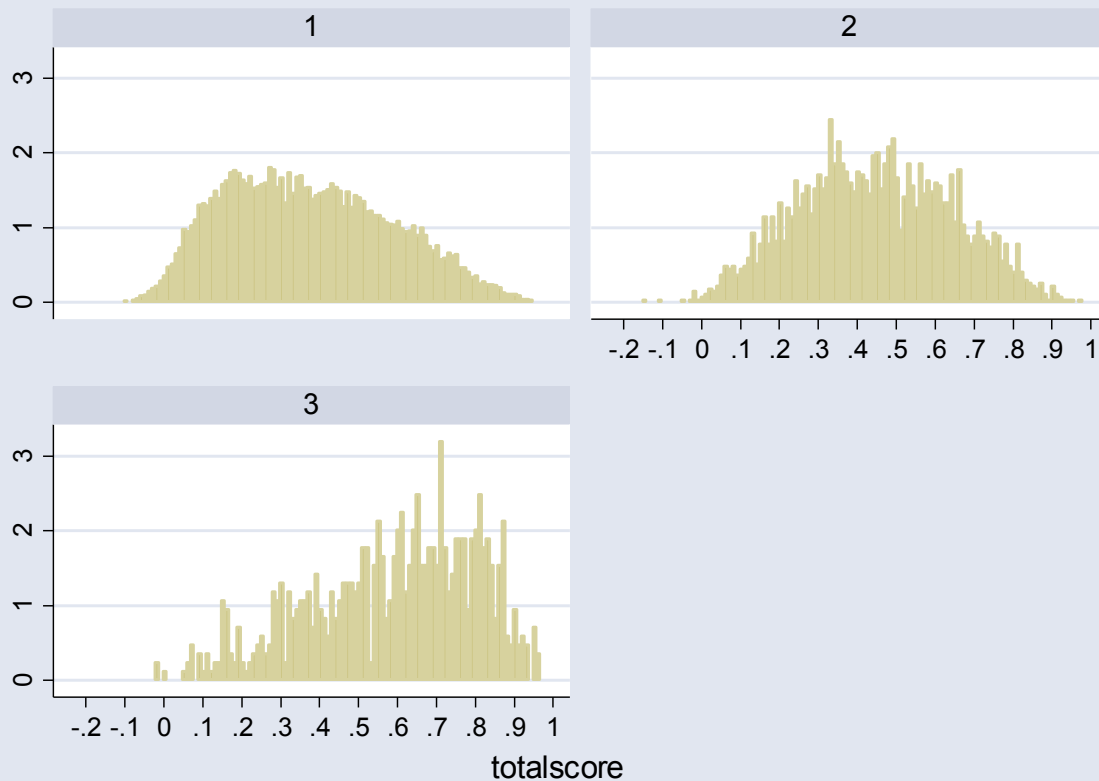
Superimpose the normal curve (with the same mean and s.d. as the empirical distribution)

```
. histogram totalscore, width(.01) xlabel(-.2 (.1) 1)  
normal  
(bin=124, start=-.24209334, width=.01)
```



Histograms by category

```
.histogram totalscore, width(.01) xlabel(-.2 (.1)1)  
  by(recodedtype)  
(bin=124, start=-.24209334, width=.01)
```



Graphs by recodedtype

Brief exercise: red versus blue states?

- Open CCES.dta from Examples folder in course locker
- Is America polarized? Create a histogram of partisan identification (pid7) by state
 - Necessary commands: tab (with option nolabel), recode, collapse, histogram
- Do most states fall within two standard deviations?
 - Create z-scores and use tabulate



Main issues with histograms

- Proper level of aggregation
- Non-regular data categories



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

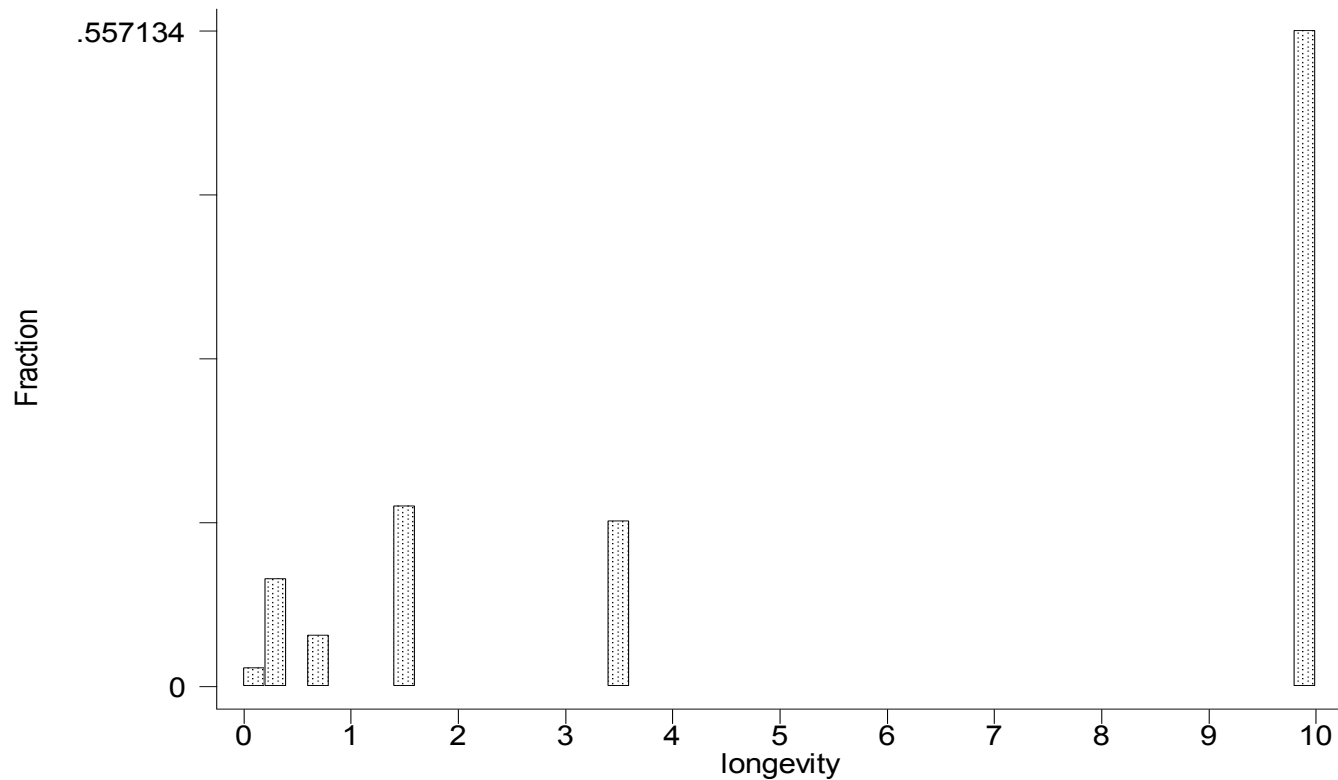
Solution, Step 1

Map artificial category onto “natural” midpoint

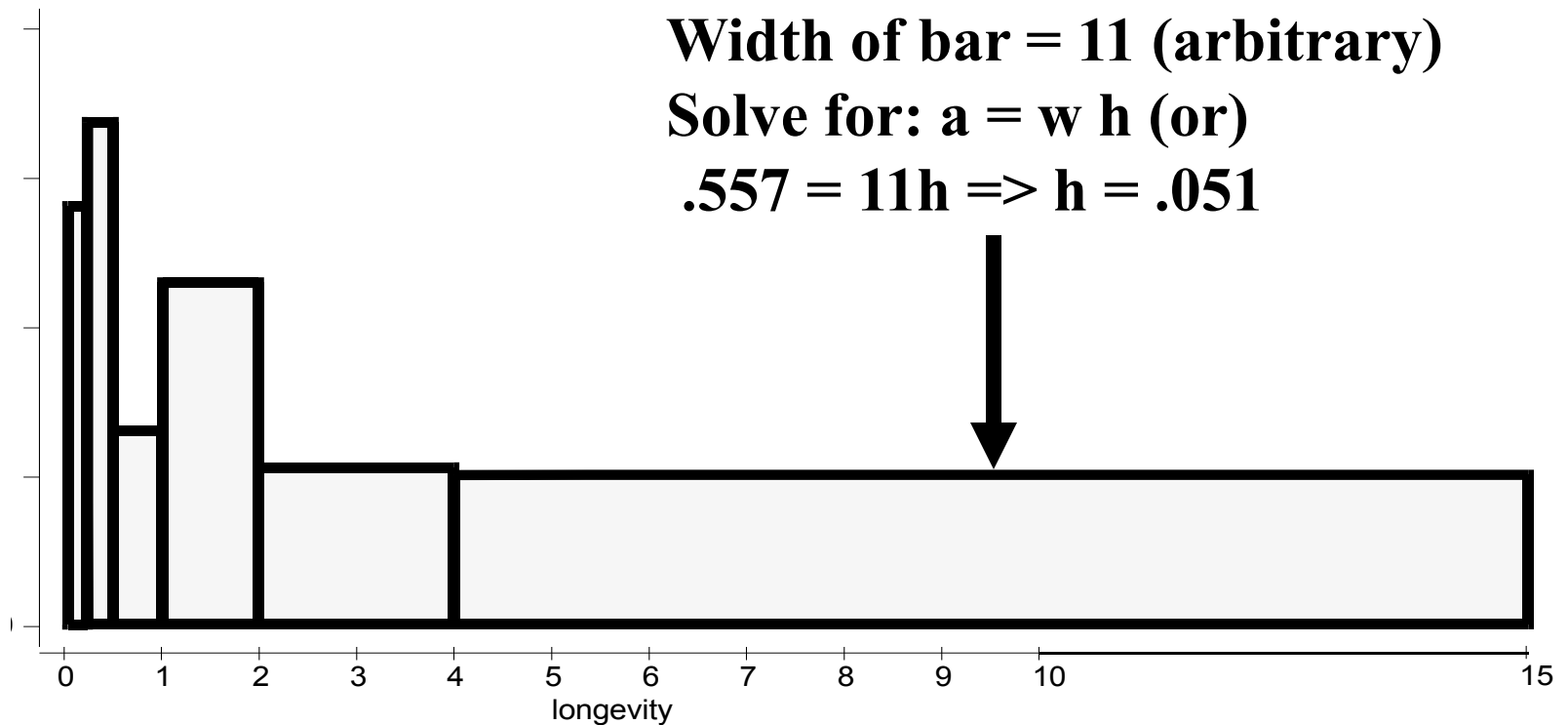
- 9 No Response → missing
- 3 Refused → missing
- 2 Don't know → missing
- 1 Not in universe → missing
- 1 Less than 1 month → $1/24 = 0.042$
- 2 1-6 months → $3.5/12 = 0.29$
- 3 7-11 months → $9/12 = 0.75$
- 4 1-2 years → 1.5
- 5 3-4 years → 3.5
- 6 5 years or longer → 10 (arbitrary)

Graph of recoded data

histogram longevity, fraction



Density plot of data



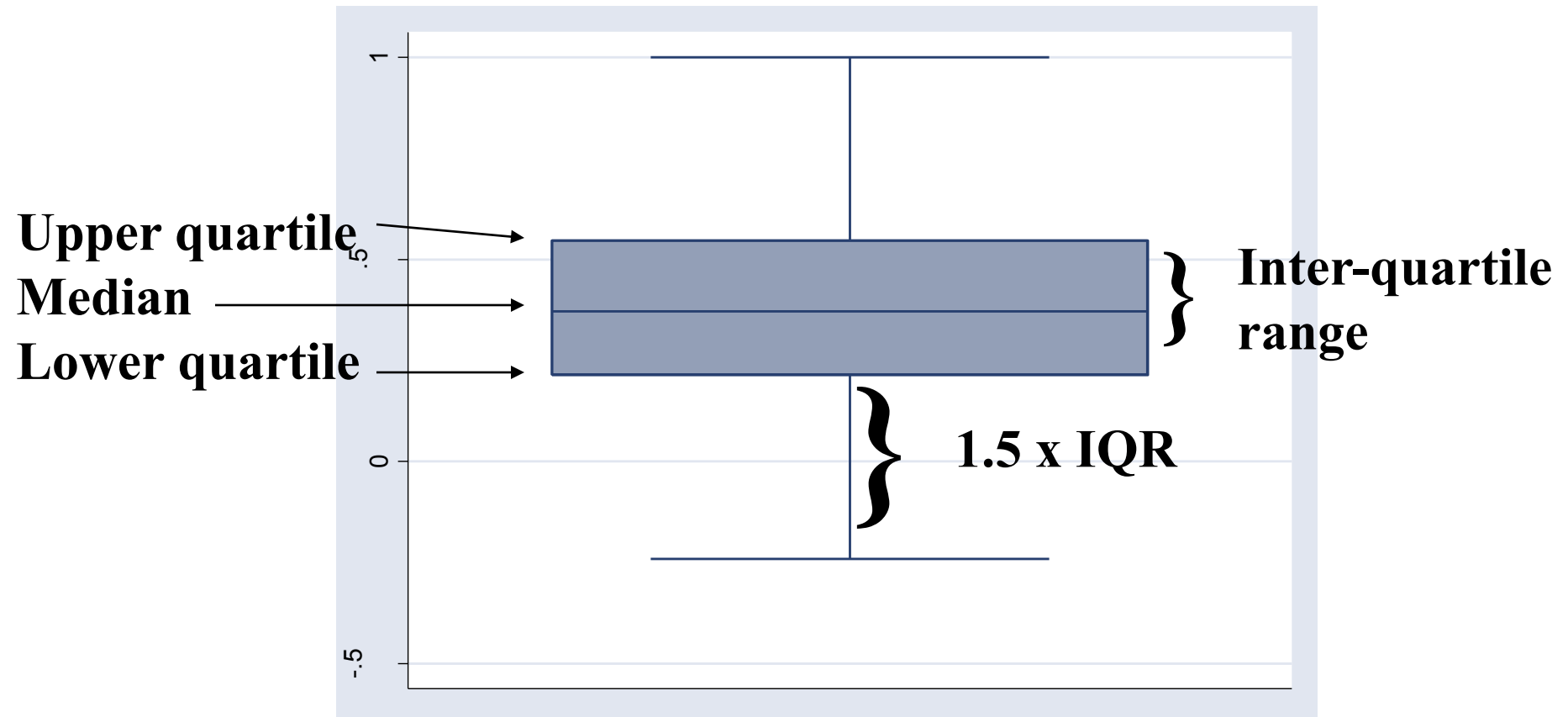
Density plot template

Category	Fraction	X-min	X-max	X-length	Height (density)
< 1 mo.	.0156	0	1/12	.082	.19*
1-6 mo.	.0909	1/12	1/2	.417	.22
7-11 mo.	.0430	1/2	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**

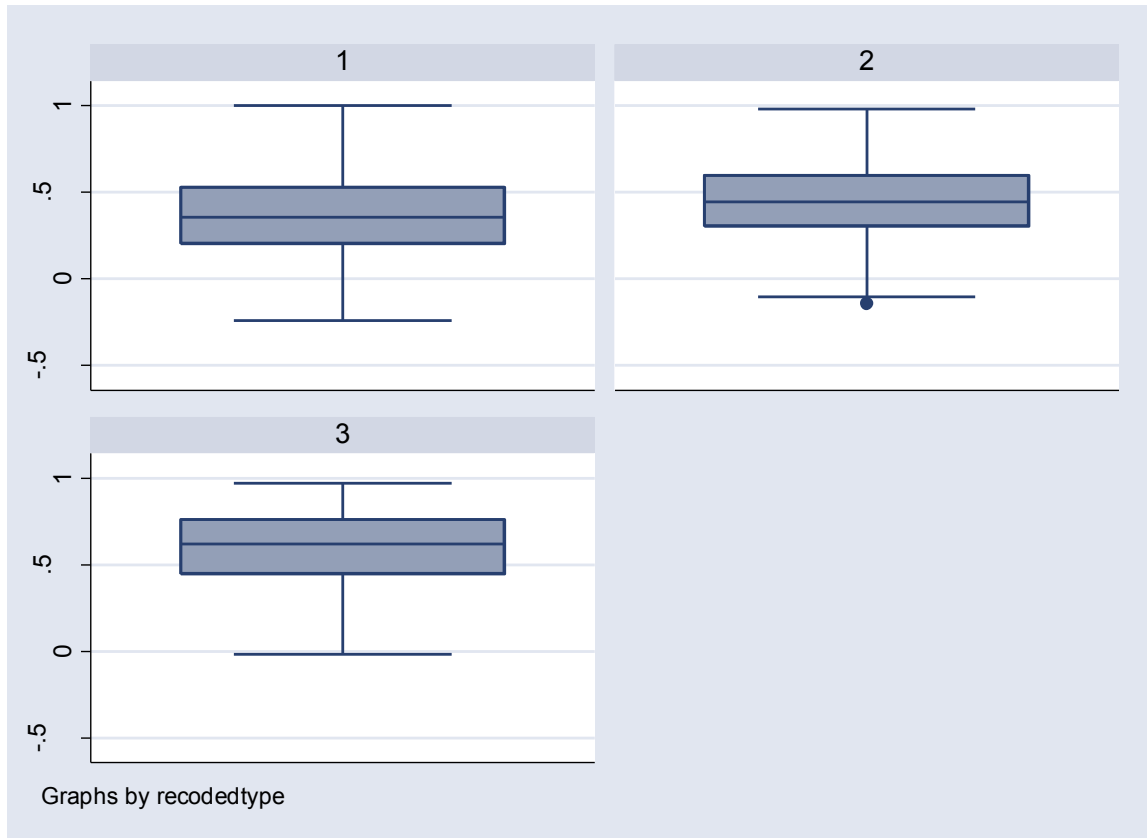
Draw the previous graph with a box plot

. graph box totalscore



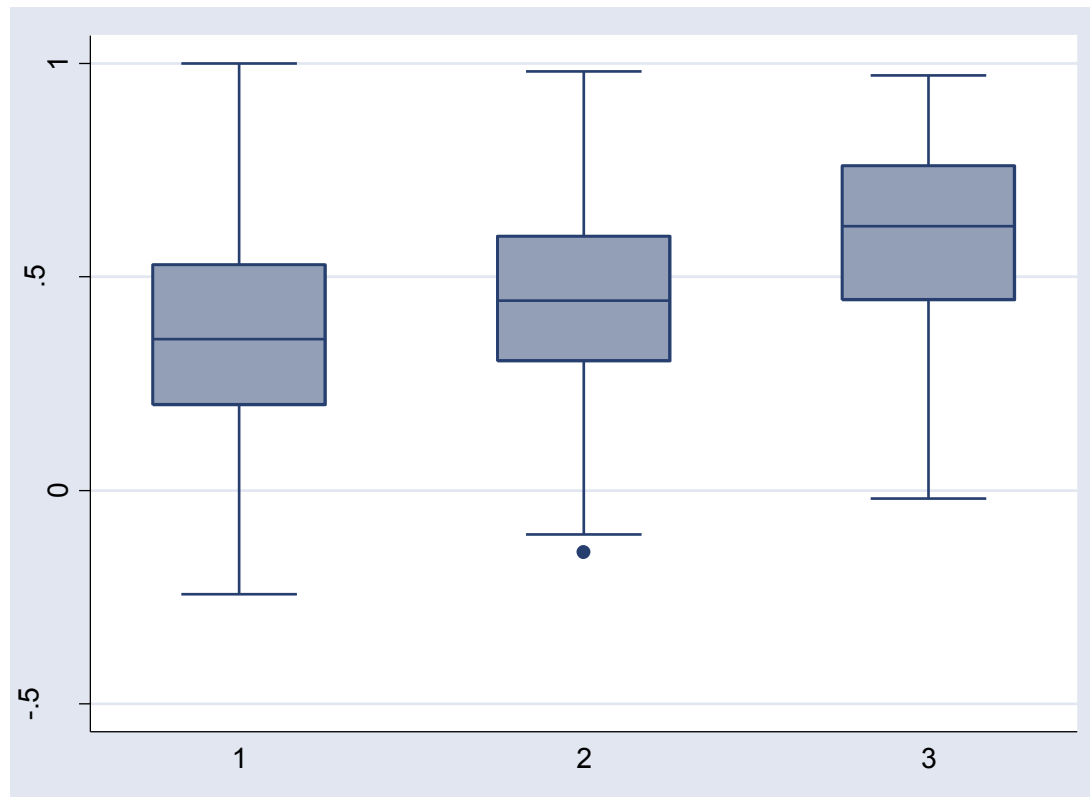
Draw the box plots for the different types of schools

```
. graph box totalscore, by(recodedtype)
```

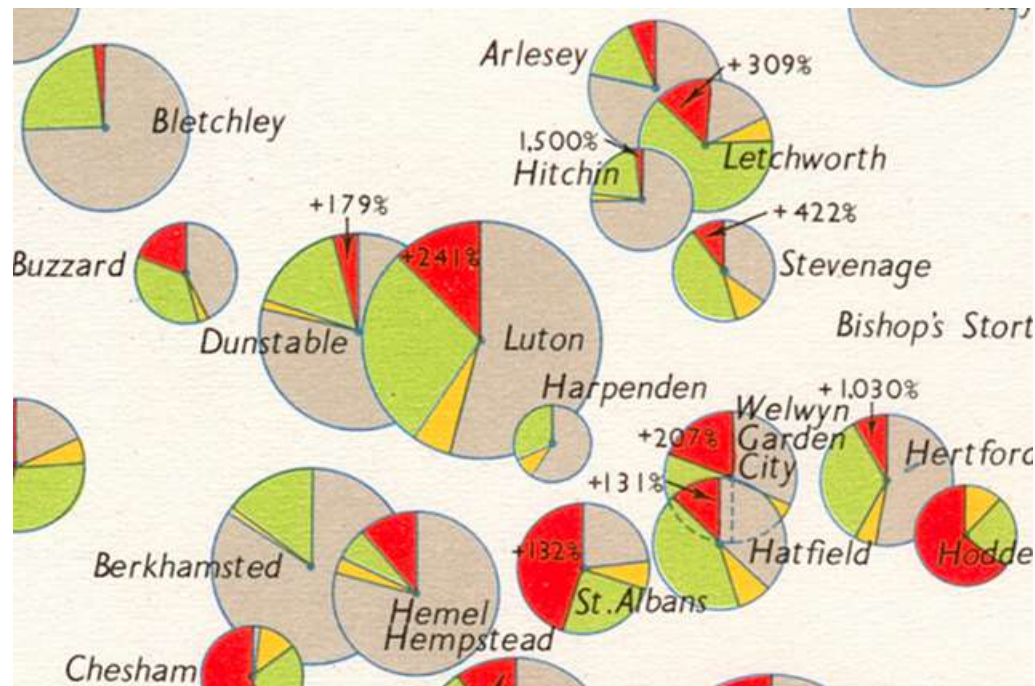


Draw the box plots for the different types of schools using “over” option

```
graph box totalscore, over (recodedtype)
```



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*)
 - Represent number in direct proportion to numerical quantities presented
 - Write clear labels on the graph
 - Show data variation, not design variation
 - Deflate and standardize money in time series