

Describing Distributions with Graphs



What to Use to Describe a Single Variable?

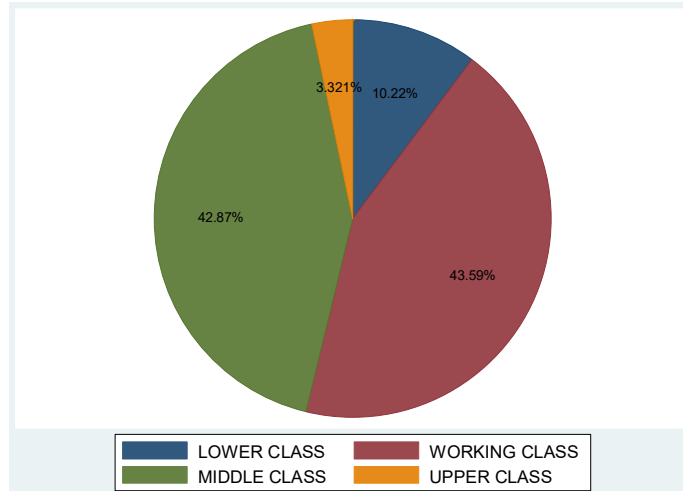
- For all: frequency distribution, percentages

Variable Type	Central Tendency	Dispersion	Graphs
Nominal	mode	variation ratio	bar, pie
Ordinal	median, mode, (sometimes: mean)	range, interquartile range	bar, pie
Interval/ratio	mean, median (preferred if outliers), mode (rarely useful)	range, interquartile range, standard deviation, variance	histogram, frequency polygon, ogive, boxplot, stem-and-leaf



Pie Chart in Stata: Our Variable is class

```
graph pie, over(class) plabel(_all percent)
```



Exporting Stata Graphs

- Export into png format

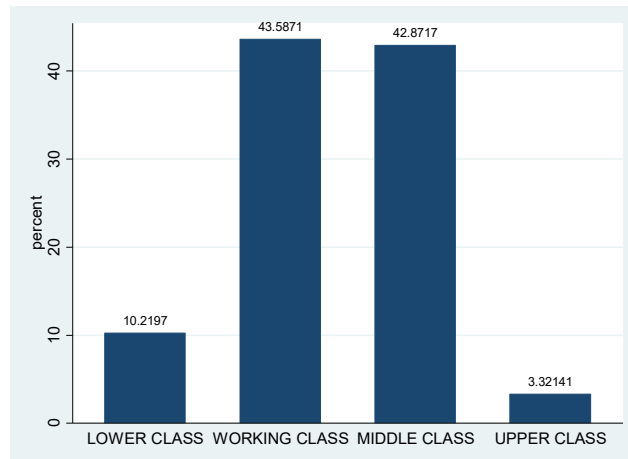
```
graph export class_pie.png
```

- When in Word, use Insert -> Picture to add graphs
- If file already exists and want to overwrite: use replace option

```
graph export class_pie.png, replace
```

Bar Graph in Stata: Our Variable is class

```
graph bar, over(class) blabel(bar)
```

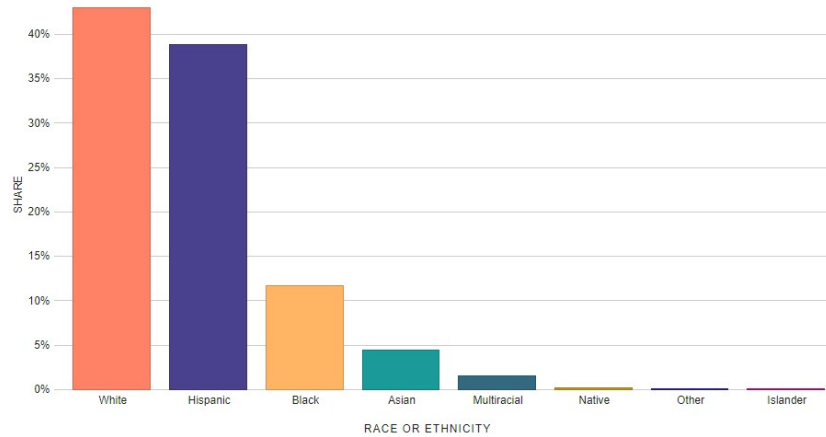


Univariate vs Bivariate Bar Graph

- Univariate Bar Graph = percentages for categories of one nominal/ordinal variable → add up to 100%
- Bivariate Bar Graph = represents the mean or percentage of something calculated separately for each group (2 variables – one main variable and one group variable) → does not add up to 100%

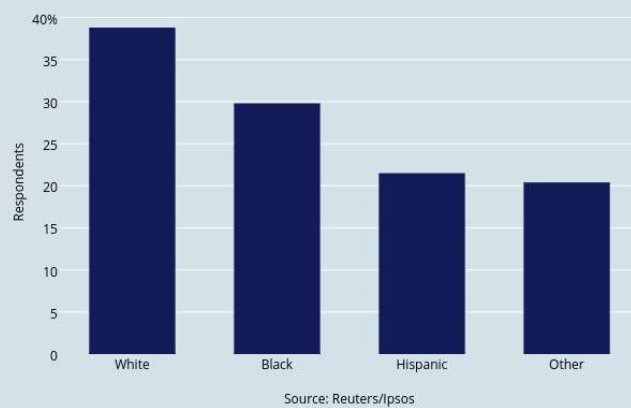
Univariate Bar Graph

Race and Ethnicity in Texas (2015)

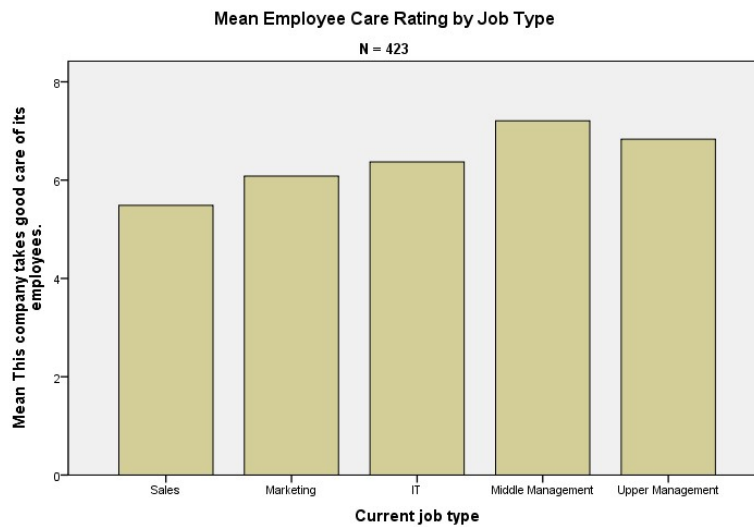


Bivariate Bar Graph

Americans with No Friends of Another Race



Bivariate Bar Graph: Means



Univariate Bar vs. Pie: Largely Interchangeable

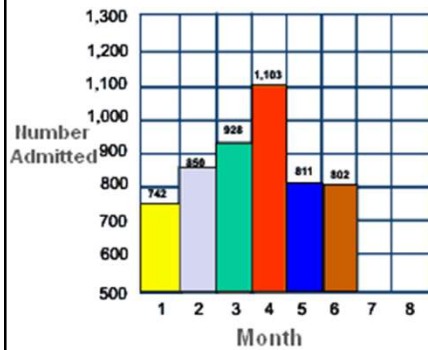


"...and this chart shows people clearly prefer bar graphs over pie charts."

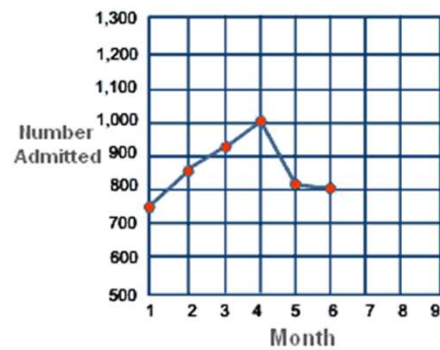
Line Instead of Bars = Line Plot (Always Bivariate)

Emergency Department Admissions per Month

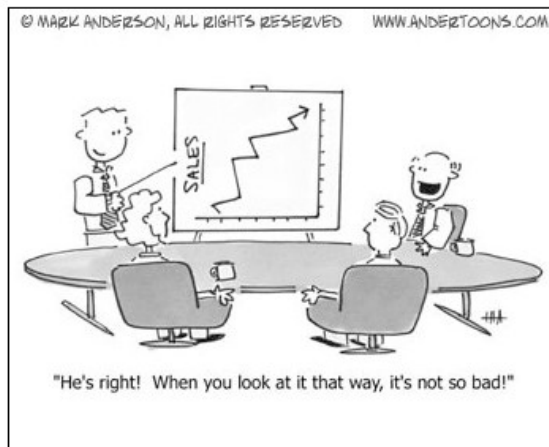
Bar Graph



Line Graph



Misleading Graphs

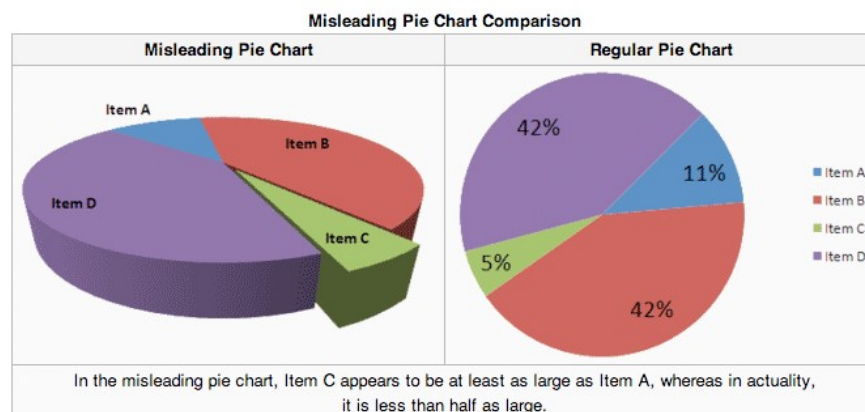


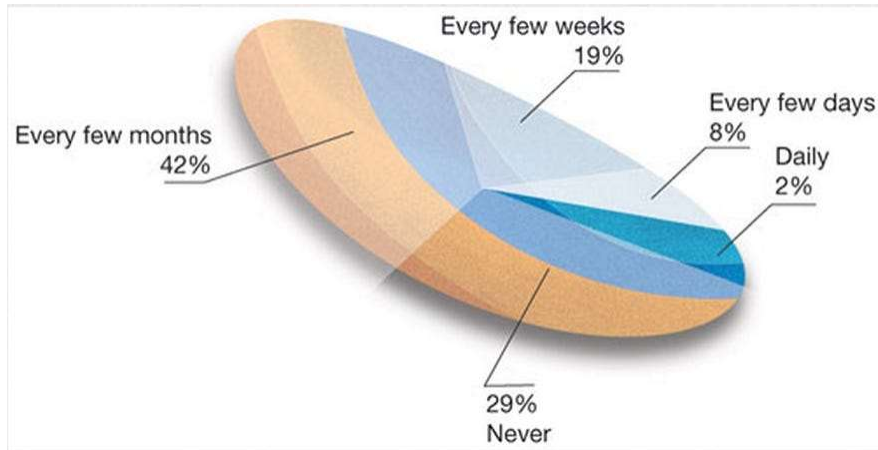
Typical Problems with Graphs

- The proportional ink principle is violated
- The graph isn't labeled properly
- The vertical scale is too big or too small, or skips numbers, or doesn't start at zero
- The horizontal scale is unevenly spaced
- Data are left out

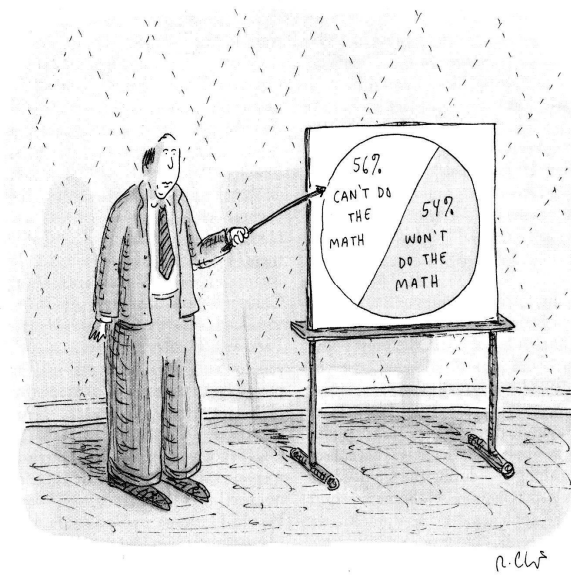


3D Pie Charts: Violating Proportional Ink Principle

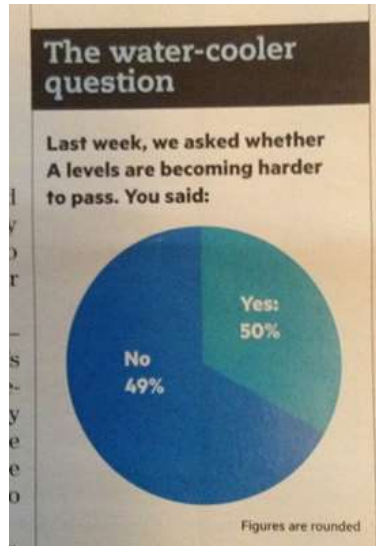




Pie Charts Must Add Up to 100%

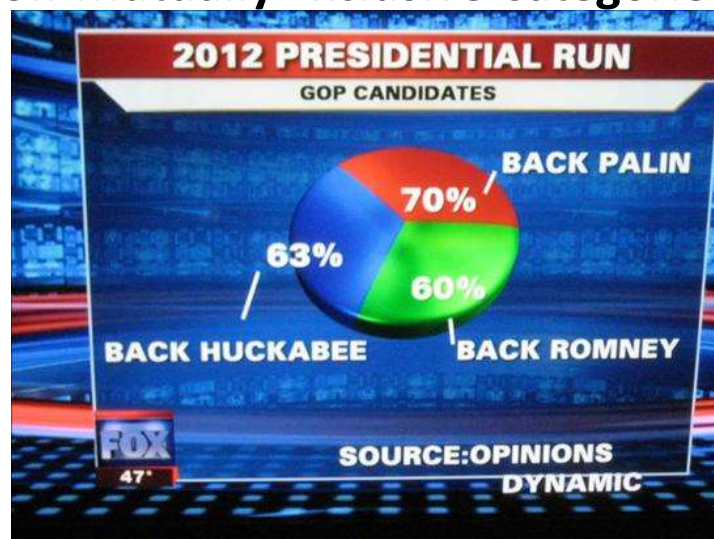


Rounding Issues and Visual Deception

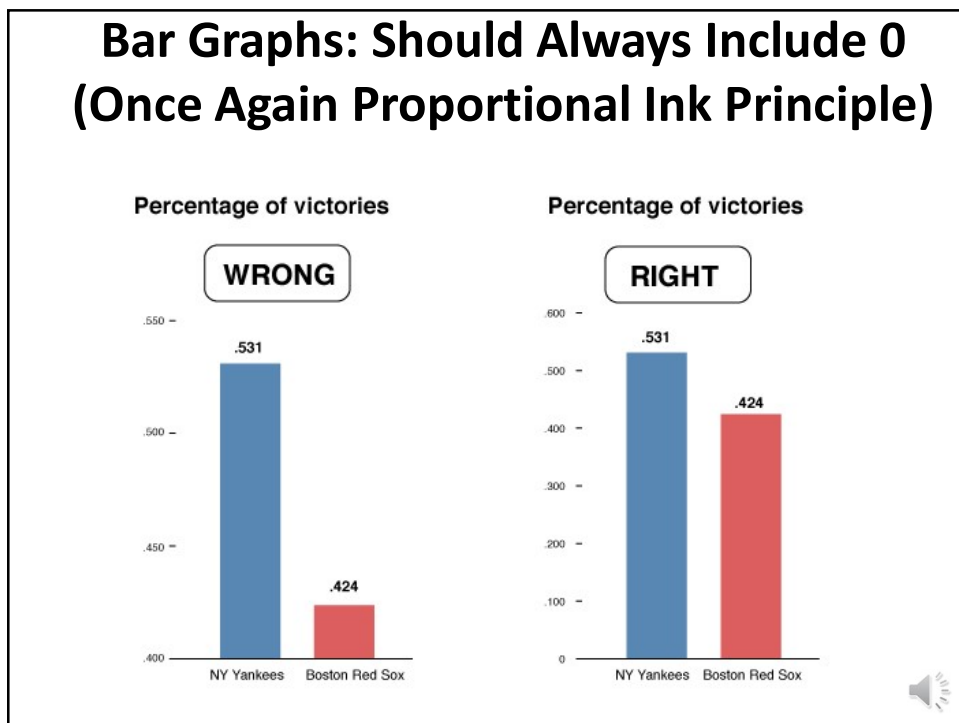
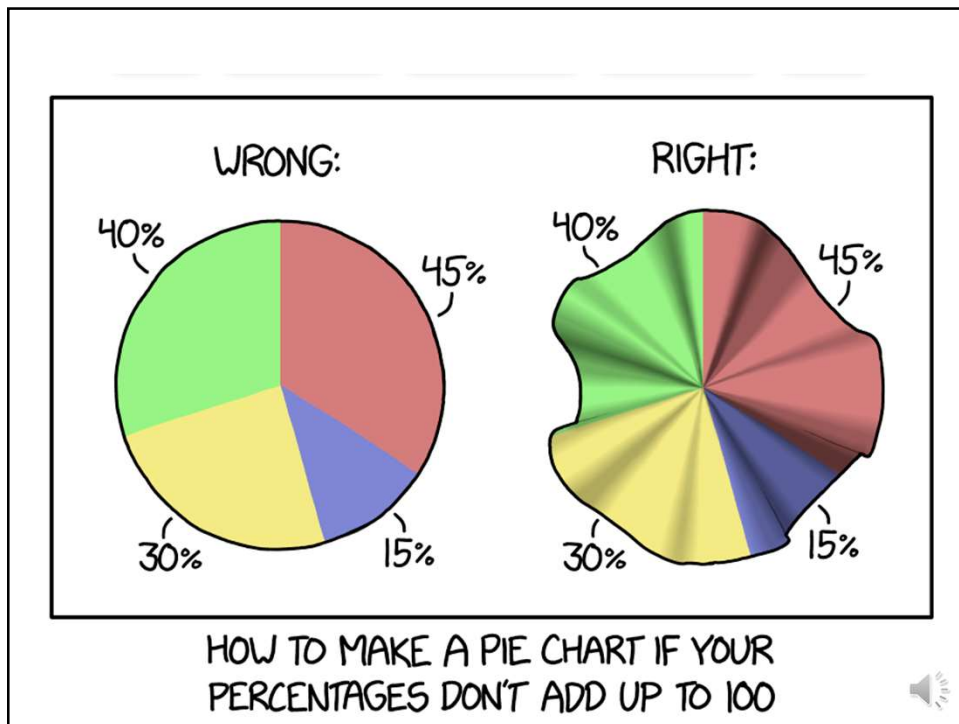


Source: <https://www.shinobicontrols.com/blog/6-common-mistakes-with-data-visualization>

Don't Use Pie Charts With Non-Mutually Exclusive Categories



Source: <https://www.shinobicontrols.com/blog/6-common-mistakes-with-data-visualization>



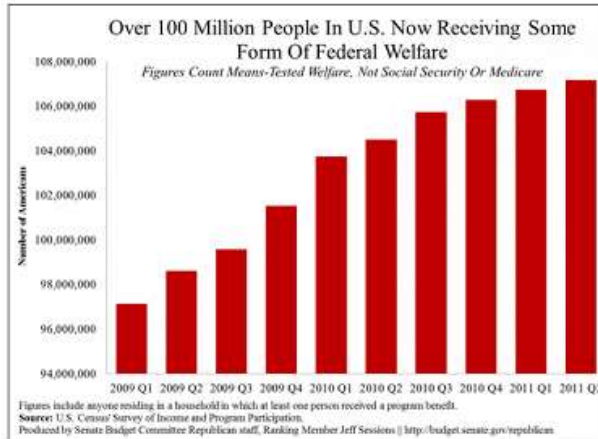
THE BLOG

Over 100 Million Now Receiving Federal Welfare

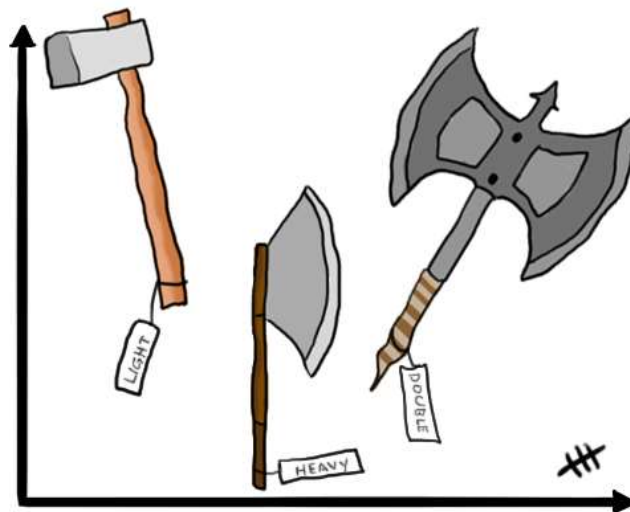
2:40 PM, AUG 8, 2012 - BY DANIEL HALPER

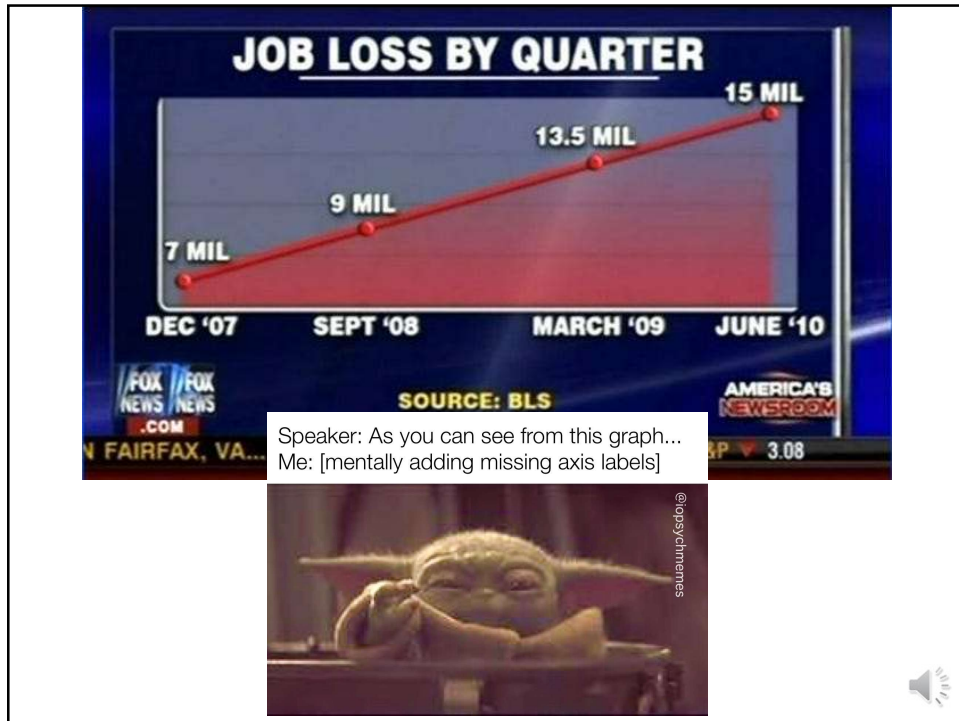
SHARE PAGE PRINT LARGER TEXT SMALLER TEXT RICKETS

A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."

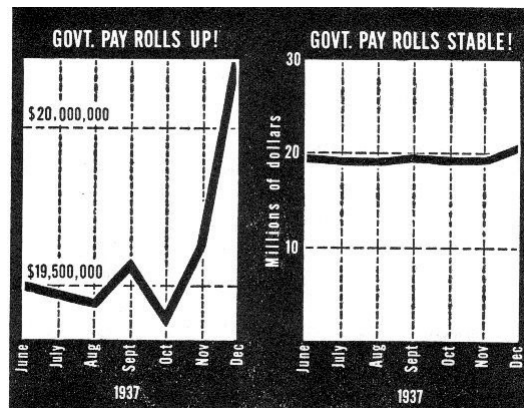


Always label your axes

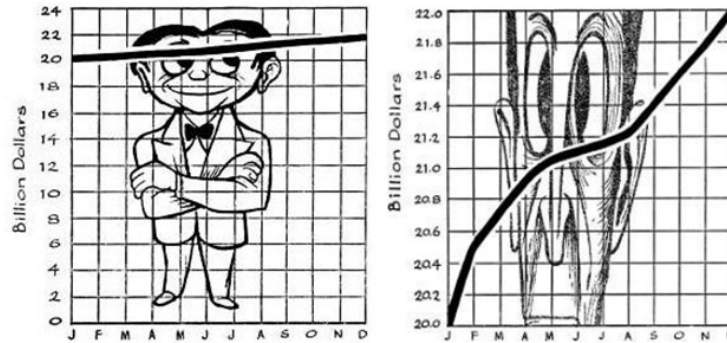




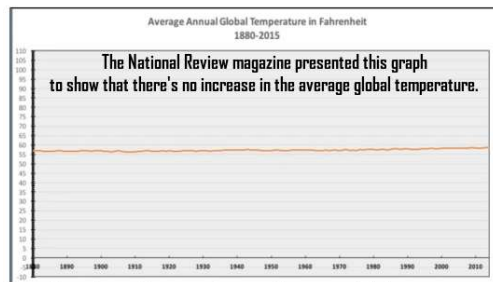
Scale Distortion in Line Graphs



Pick a Reasonable Range of Values



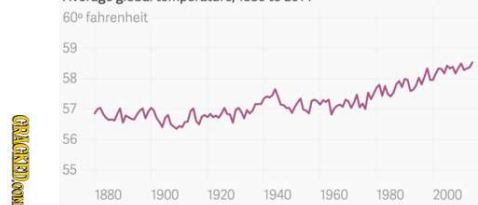
Line Graphs Don't Have to Include 0



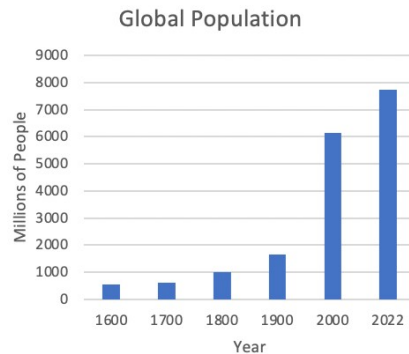
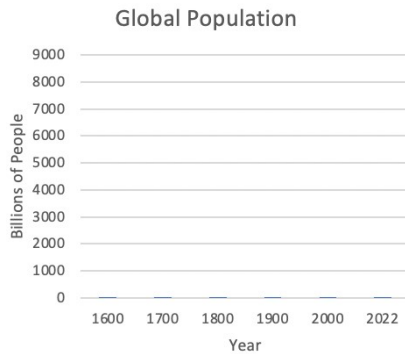
But a change of even one degree in the average global temperature is hugely significant.

Although the data are correct, if we zoom the graph we see the actual rise of global temperatures across the years.

Average global temperature, 1880 to 2014

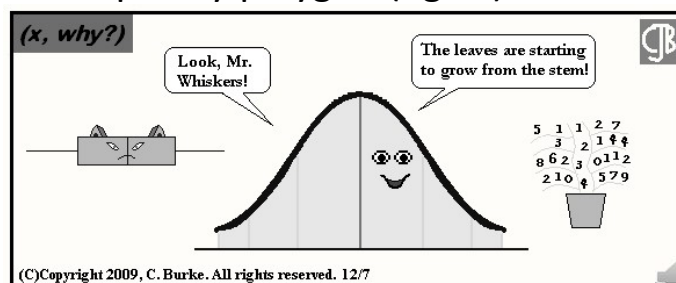


Axis labels should be about 10-20% higher than the highest value



Now For the More Technical Plots that Rarely Make News

- Histogram
- Stem-and-leaf plot
- Frequency polygon
- Cumulative frequency polygon (ogive)
- Boxplot



Frequency Distribution Is Too Long for a Bar Graph?

```
tab tvhours
```

HOURS PER DAY WATCHING TV	Freq.	Percent	Cum.
0	90	6.93	6.93
1	255	19.65	26.58
2	325	25.04	51.62
3	238	18.34	69.95
4	171	13.17	83.13
5	61	4.70	87.83
6	58	4.47	92.30
7	19	1.46	93.76
8	31	2.39	96.15
9	3	0.23	96.38
10	17	1.31	97.69
12	11	0.85	98.54
13	2	0.15	98.69
14	4	0.31	99.00
15	3	0.23	99.23
16	1	0.08	99.31
18	1	0.08	99.38
20	2	0.15	99.54
22	1	0.08	99.61
24	5	0.39	100.00
Total	1,298	100.00	



Combine into Equal Length Intervals (Binning)

```
tab tvhours
```

HOURS PER DAY WATCHING TV	Freq.	Percent	Cum.
0	90	6.93	6.93
1	255	19.65	26.58
2	325	25.04	51.62
3	238	18.34	69.95
4	171	13.17	83.13
5	61	4.70	87.83
6	58	4.47	92.30
7	19	1.46	93.76
8	31	2.39	96.15
9	3	0.23	96.38
10	17	1.31	97.69
12	11	0.85	98.54
13	2	0.15	98.69
14	4	0.31	99.00
15	3	0.23	99.23
16	1	0.08	99.31
18	1	0.08	99.38
20	2	0.15	99.54
22	1	0.08	99.61
24	5	0.39	100.00
Total	1,298	100.00	



Resulting Distribution

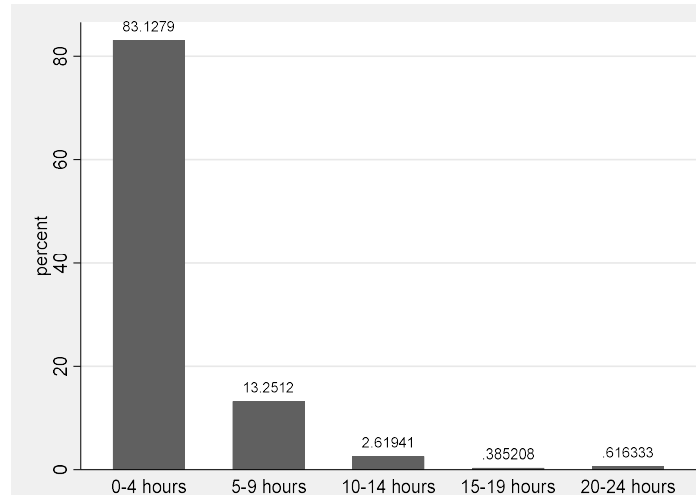
```
tab tvhours5
```

TV hours in intervals of 5 hours	Freq.	Percent	Cum.
0-4 hours	1,079	83.13	83.13
5-9 hours	172	13.25	96.38
10-14 hours	34	2.62	99.00
15-19 hours	5	0.39	99.38
20-24 hours	8	0.62	100.00
Total	1,298	100.00	



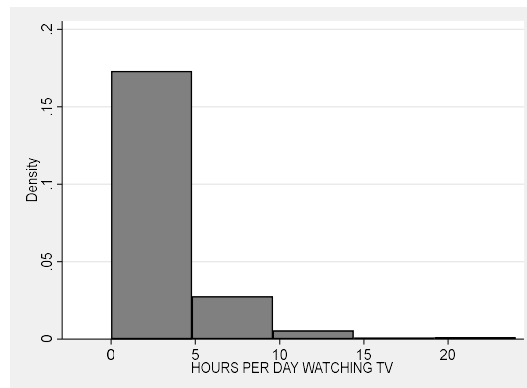
Bar Chart for Collapsed Distribution

```
graph bar, over(tvhours5) blabel(bar)
```



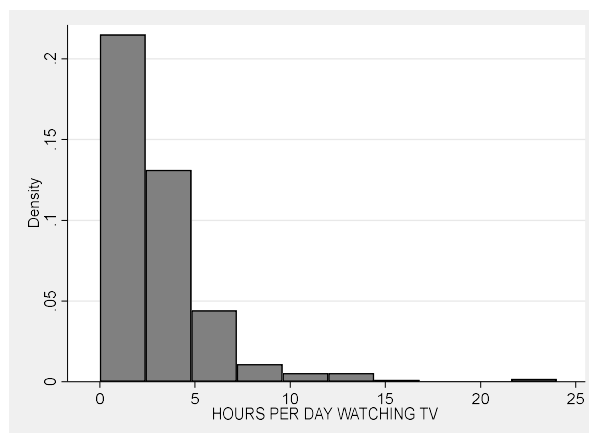
Bar Chart for Collapsed Distribution = Histogram (with 5 Intervals)

```
histogram tvhours, bin(5)
```



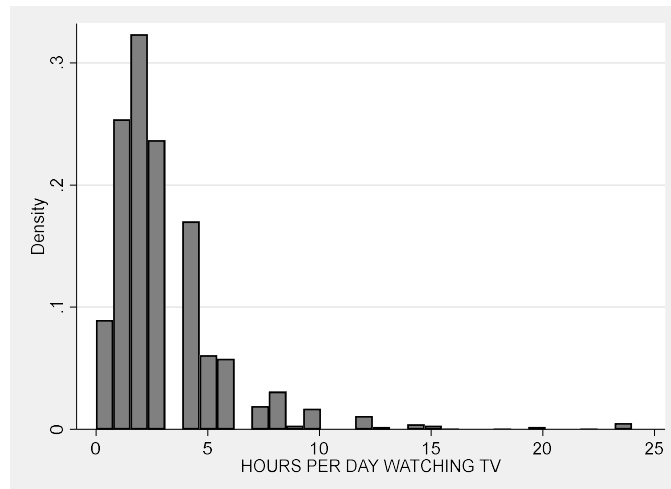
Easy to Change: Histogram with 10 Intervals

```
histogram tvhours, bin(10)
```



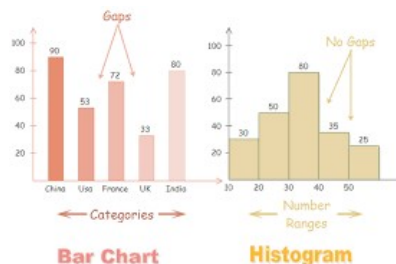
Stata Default Version

histogram tvhours

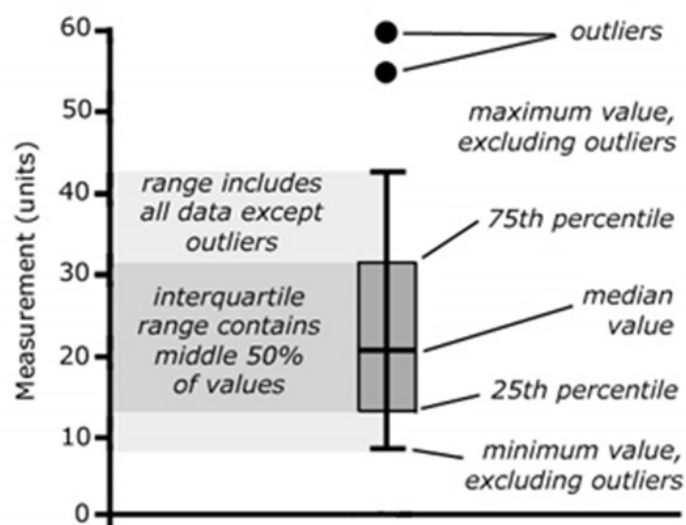


Histogram or Univariate Bar Graph?

- Bar graph (nominal/ordinal variable or 2 vars):
 - Spaces between bars
 - Bars separately labeled
- Histogram (one interval/ratio variable):
 - No spaces between bars
 - No individual labels for bars

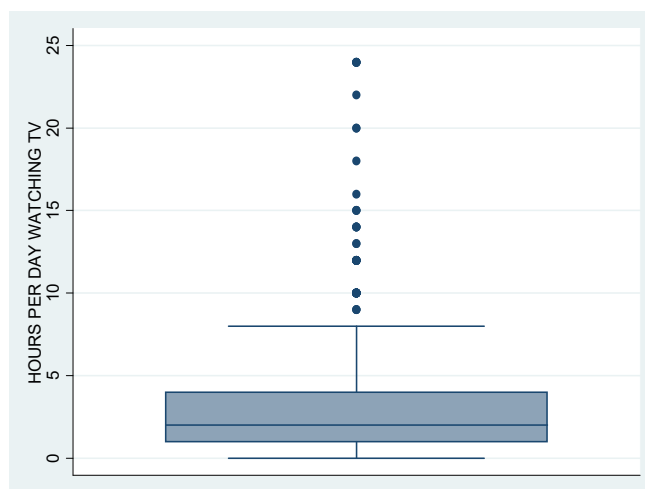


Box-and-Whisker Plot (Boxplot)



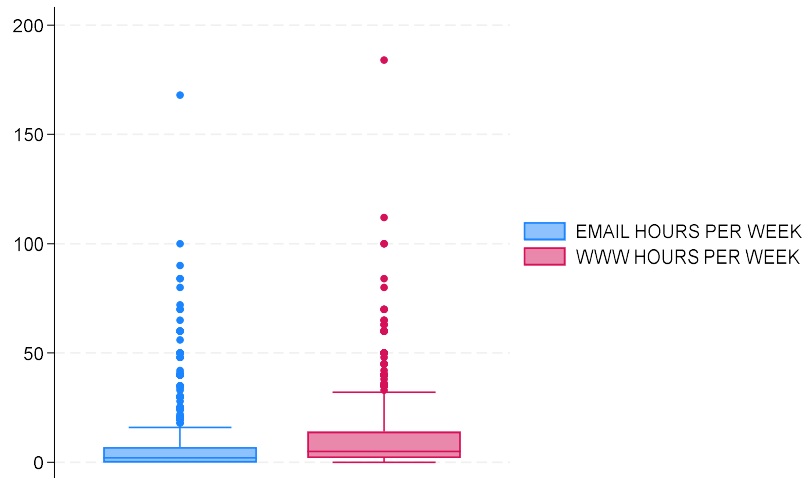
Boxplot in Stata

```
. graph box tvhours
```



Comparing Distributions

```
. graph box emailhr wwwhr
```



Outliers in Boxplot

- Below $Q1 - 1.5 \times IQR$
- Above $Q3 + 1.5 \times IQR$



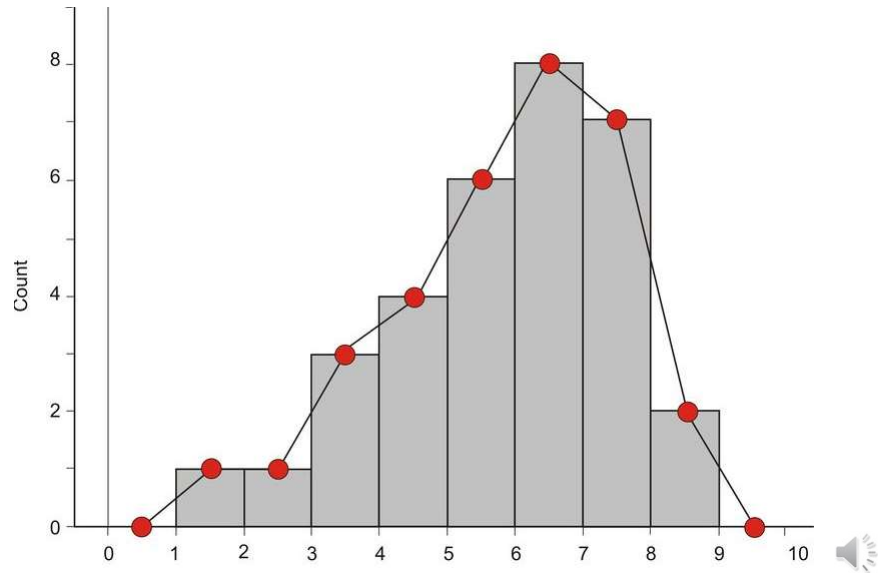
Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

how to place "32"

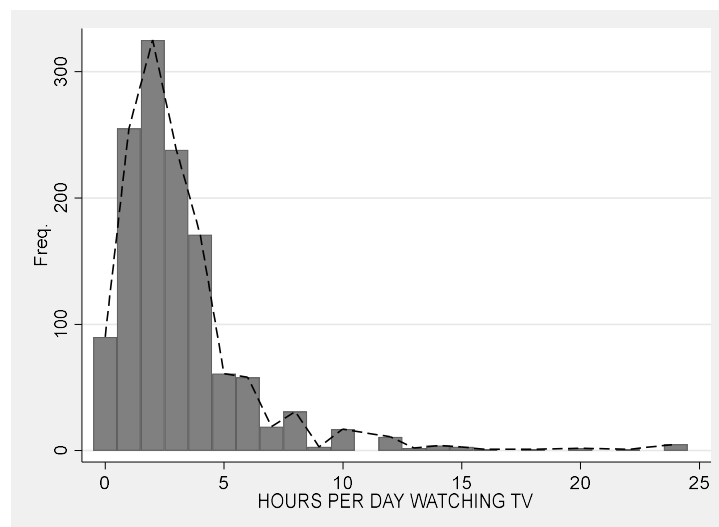


```
. stem tvhours
```


Frequency Polygon: Connect the Dots

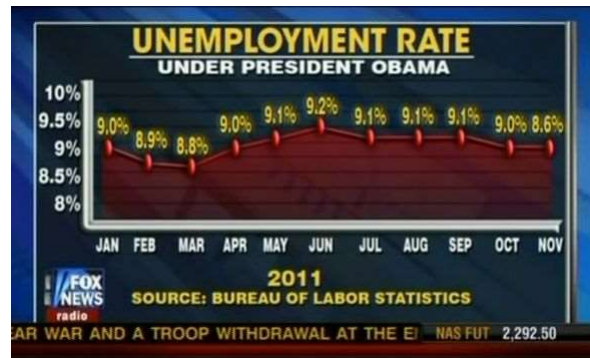


Frequency Polygon for TV hours

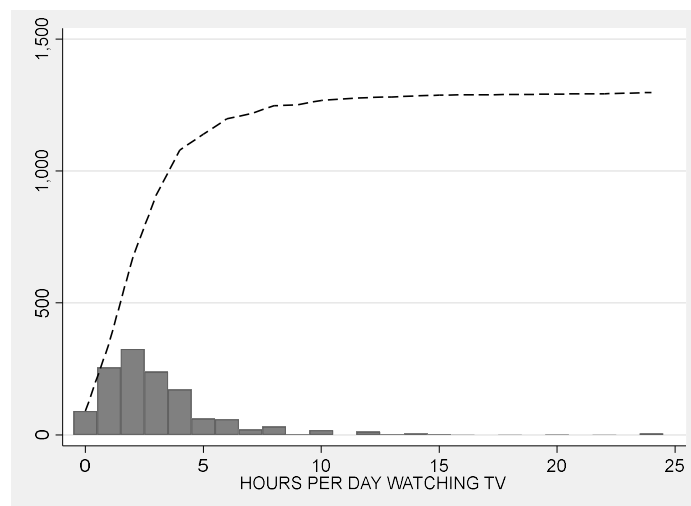


Line Plot vs Frequency Polygon

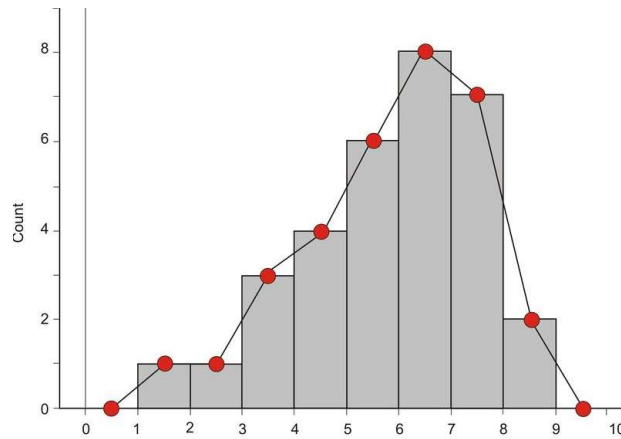
- Line plot is bivariate (two variables)
- Frequency polygon is univariate (one variable)



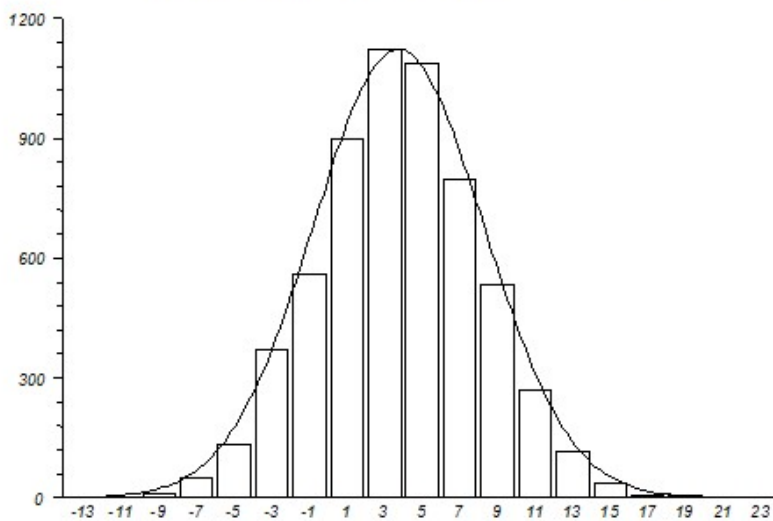
Cumulative Frequency Polygon (Ogive) for TV hours



Shape of a Distribution: Connecting the Dots

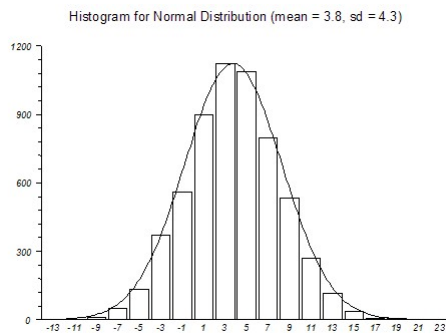


Histogram for Normal Distribution (mean = 3.8, sd = 4.3)

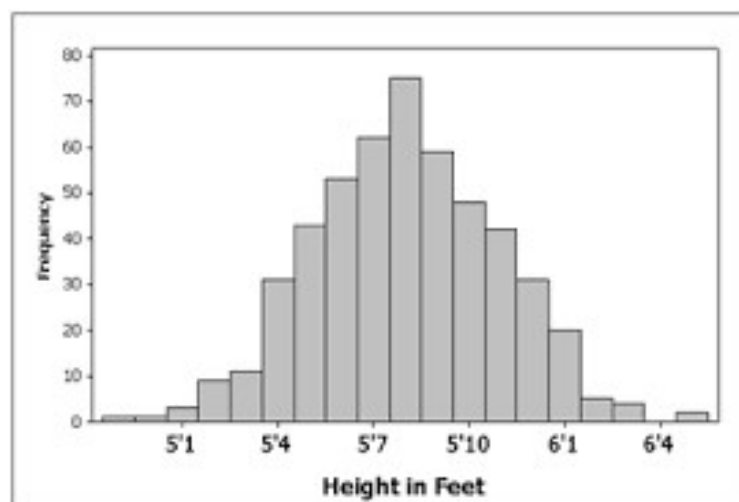


Normal Curve (Bell Curve)

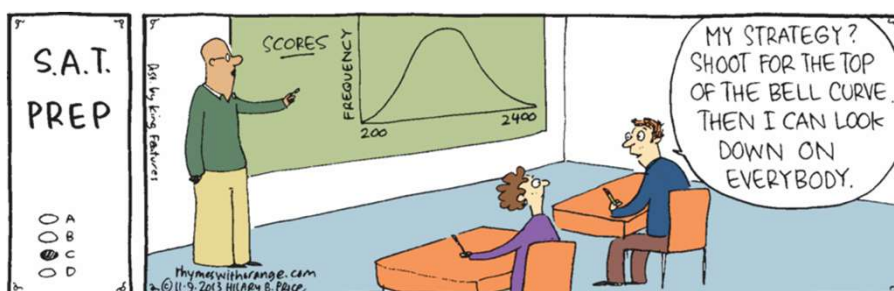
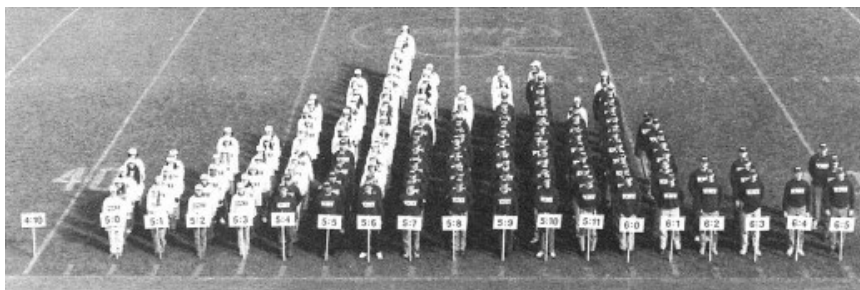
- The most well known “shape” is the normal curve
- Random processes (chance) often result in a normal curve



Height Distribution



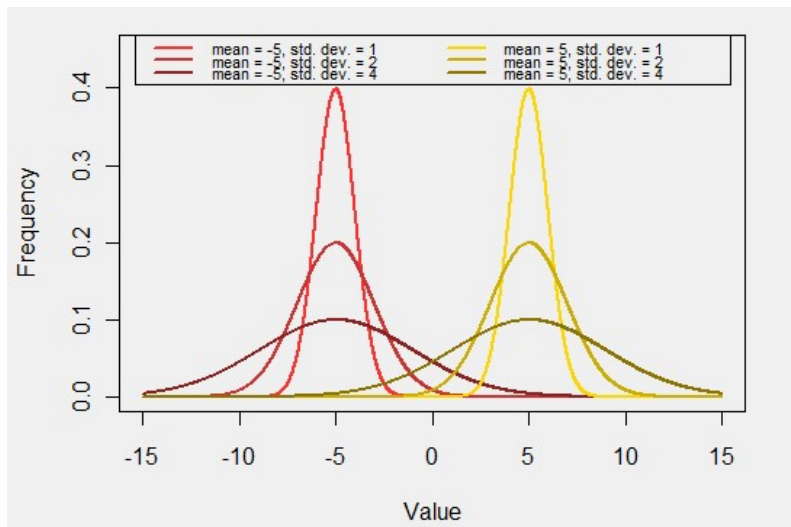
Height Distribution



- Normal curve: Mean = median = mode

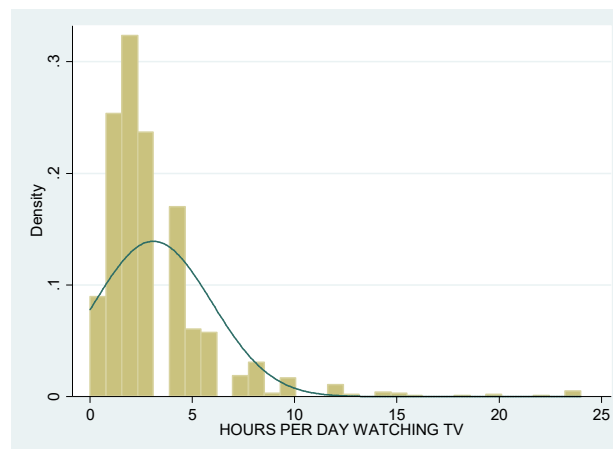


Same Shape, Different Versions

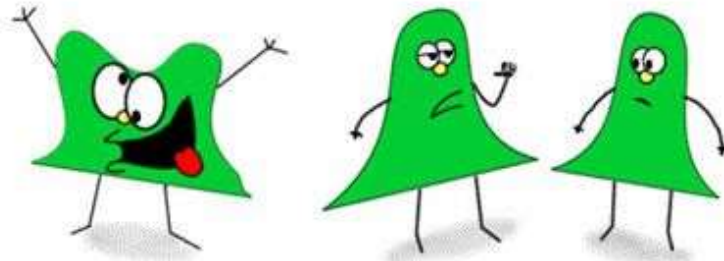


Histogram with a Normal Curve In Stata

histogram tvhours, normal



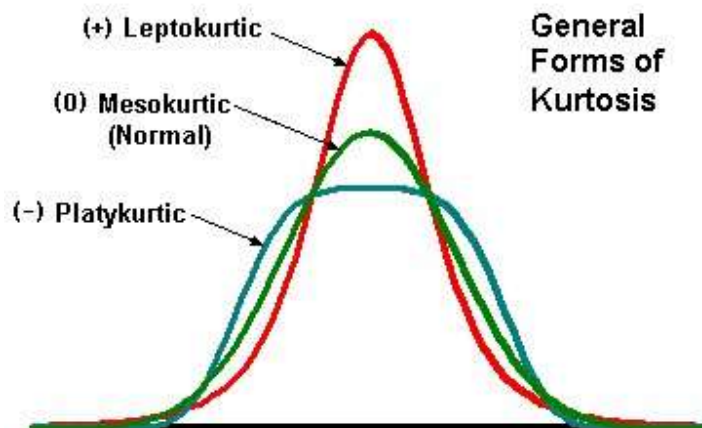
What's Not Normal?



"KEEP YOUR EYE ON THAT GUY, TOM. HE'S NOT, YOU KNOW...NORMAL!"

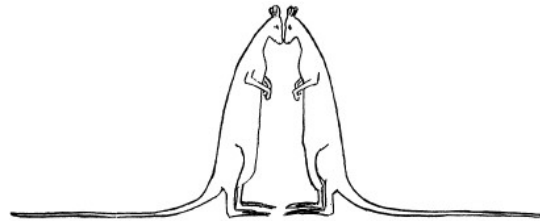


Measures of Distribution's Shape: Kurtosis

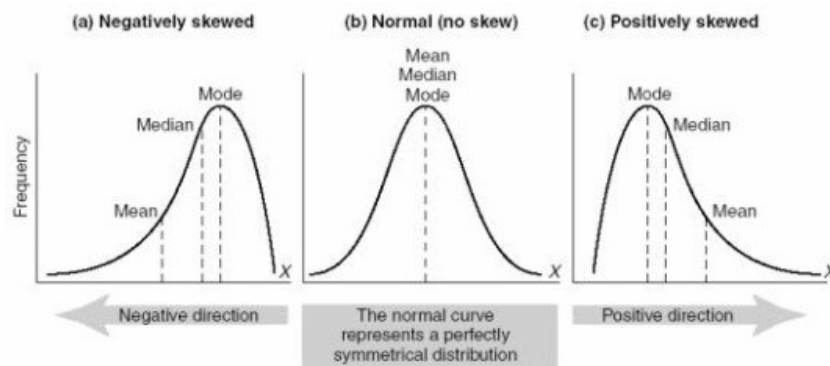


Platypus = platykurtic

Leaping kangaroos =
leptokurtic



Measures of Distribution's Shape: Skewness



Interpretation of Skewness and Kurtosis Statistics

- Skewness:
 - Near zero = symmetric (close to normal)
 - Positive = right, positive skew (especially >2)
 - Negative = left, negative skew (especially <-2)
- Kurtosis:
 - In Stata: subtract 3 before interpreting!
 - Near zero = close to normal
 - Positive = leptokurtic (especially >2)
 - Negative = platykurtic (especially <-2)



Skewness and Kurtosis in Stata

```
tabstat tvhours, stats(mean skewness  
kurtosis)
```

variable	mean	skewness	kurtosis
tvhours	3.088598	3.123997	18.48296

