# Math 1710 Class 13

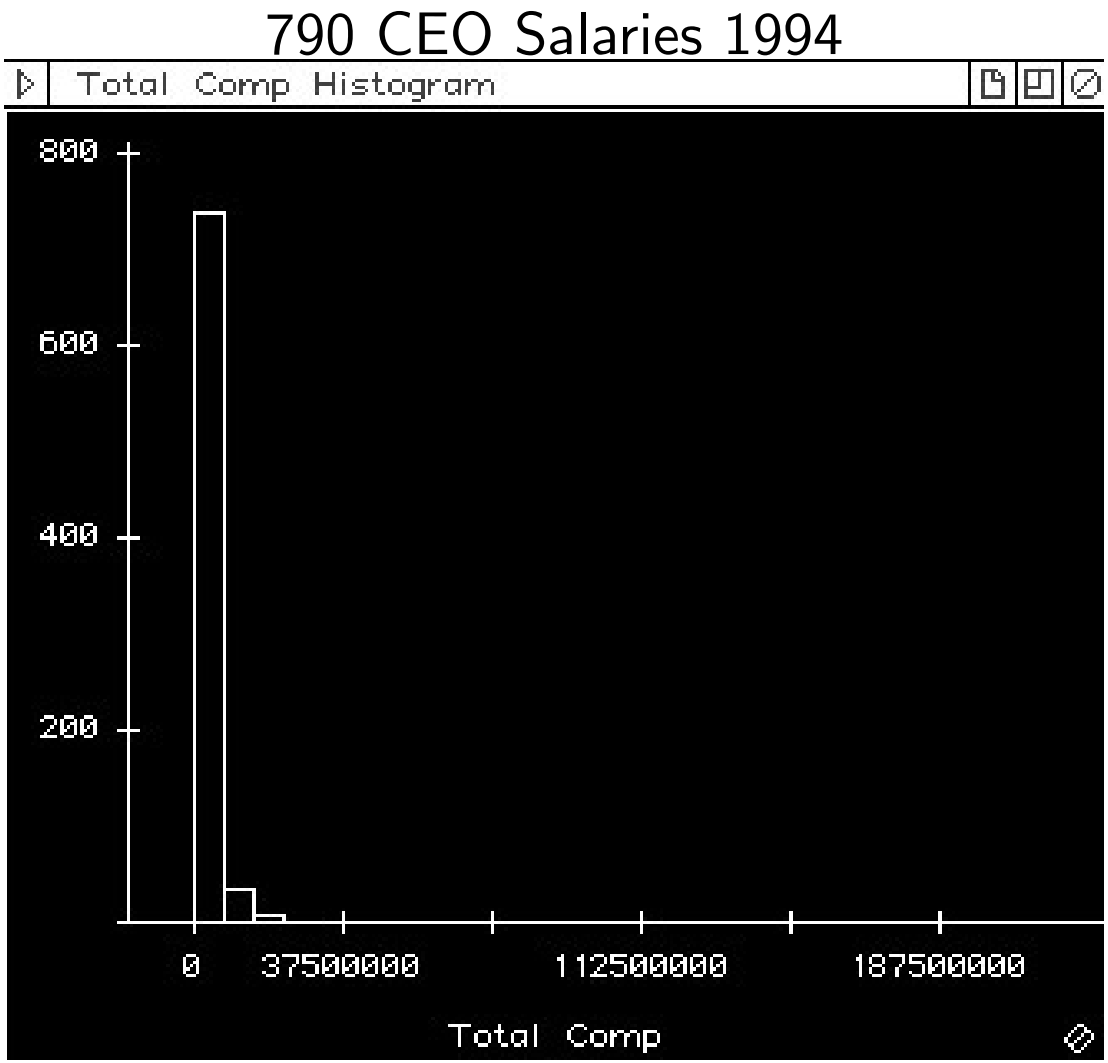
Normal Distributions, Outliers, and Summary Statistics

Dr. Back

Sep. 25, 2009

# Forbes




790 CEO Salaries 1994
Total Comp Histogram
800
600
400
200
0
37500000
112500000
187500000
Total Comp

# Forbes



## 790 CEO Salaries 1994

Total Comp

Summary of **Total Comp**
No Selector
800 total cases of which 10 are missing

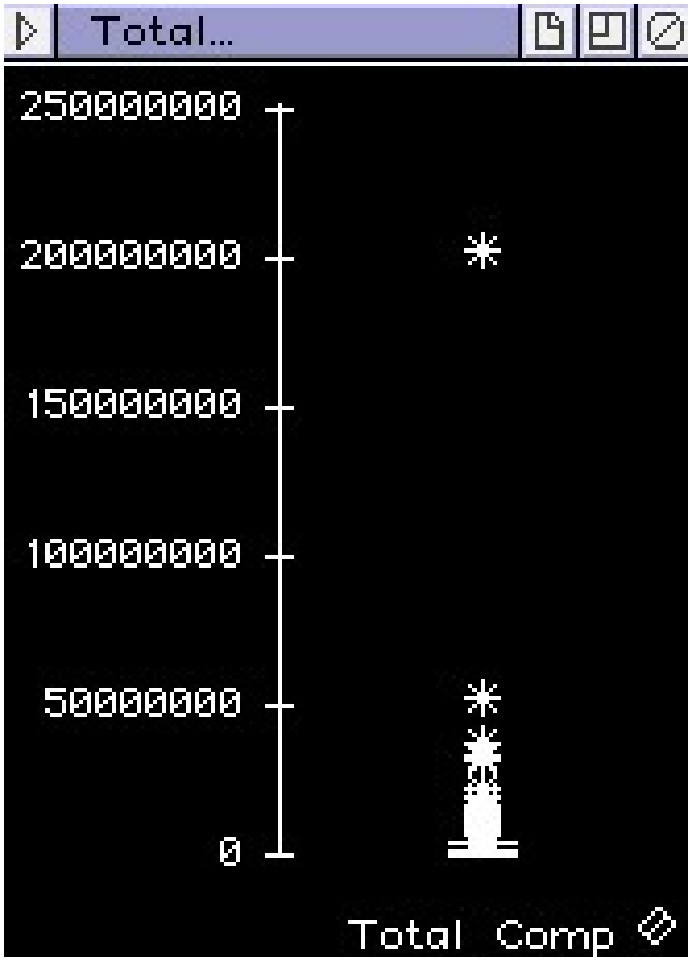
Percentile 25

| | |
|---|---|
| **Count** | 790 |
| **Mean** | 2.81874e6 |
| **Median** | 1.30447e6 |
| **StdDev** | 8.32005e6 |
| **Range** | 202.991e6 |
| **IntQRange** | 1.7274e6 |
| **Lower ith %tile** | 787841 |
| **Upper ith %tile** | 2.51524e6 |

# Forbes



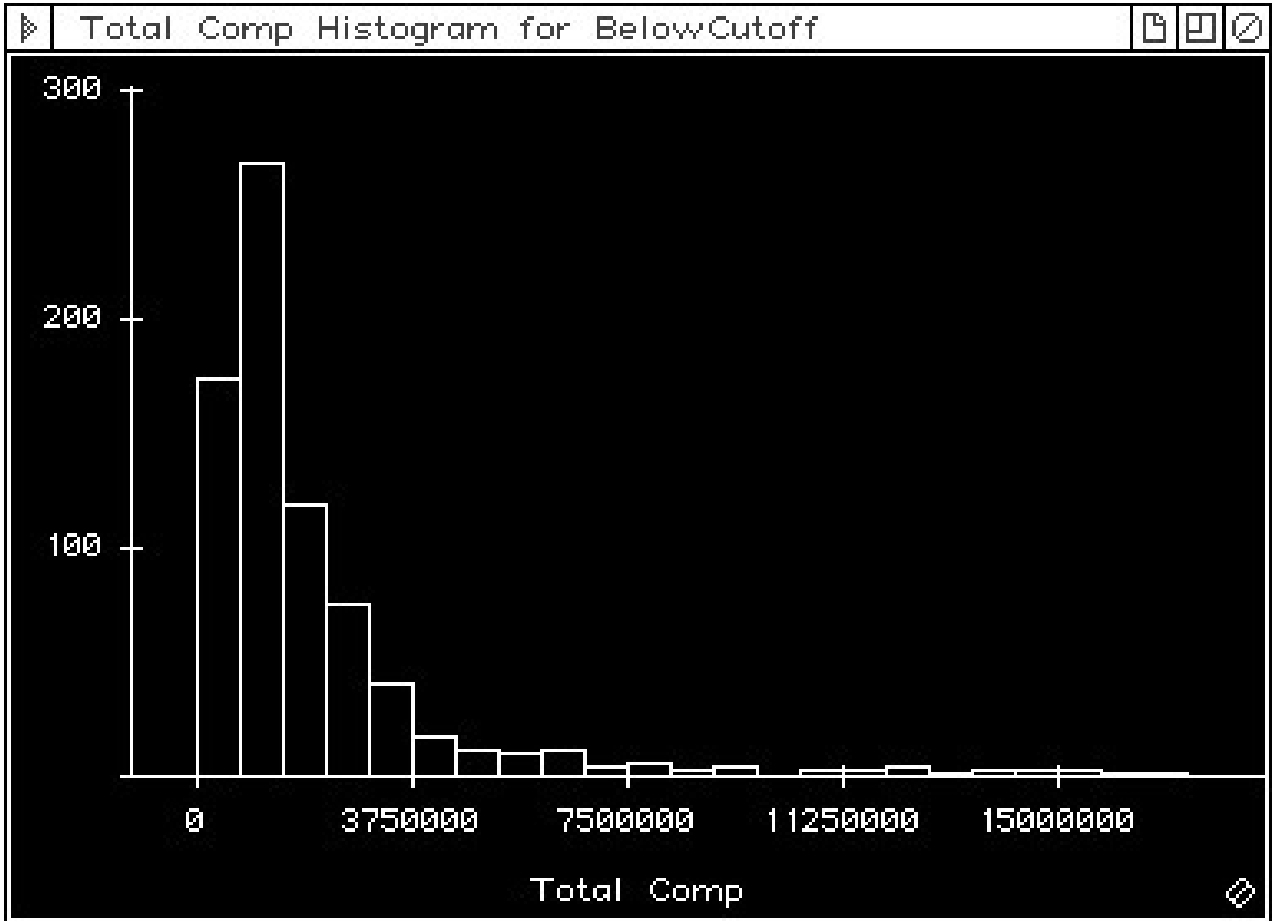
CEO Salaries 1994 Boxplot



# Forbes



All But Top 9 CEO Salaries 1994



# Forbes



## All But Top 9 CEO Salaries 1994

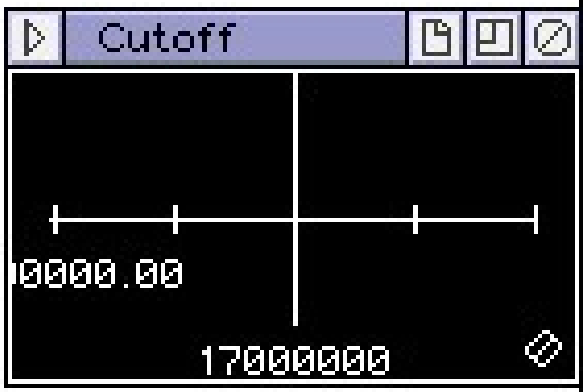
Total Comp

Summary of cases selected according to **Total Comp Below Cutoff**

800 total cases of which 19 are missing

Percentile 25

| | |
|---|---|
| **Count** | 781 |
| **Mean** | 2.24289e6 |
| **Median** | 1.29613e6 |
| **StdDev** | 2.72437e6 |
| **Range** | 16.7457e6 |
| **IntQRange** | 1.65856e6 |
| **Lower ith %tile** | 784797 |
| **Upper ith %tile** | 2.44336e6 |



# Forbes



Last Time

Quartiles in the CEO Data Set

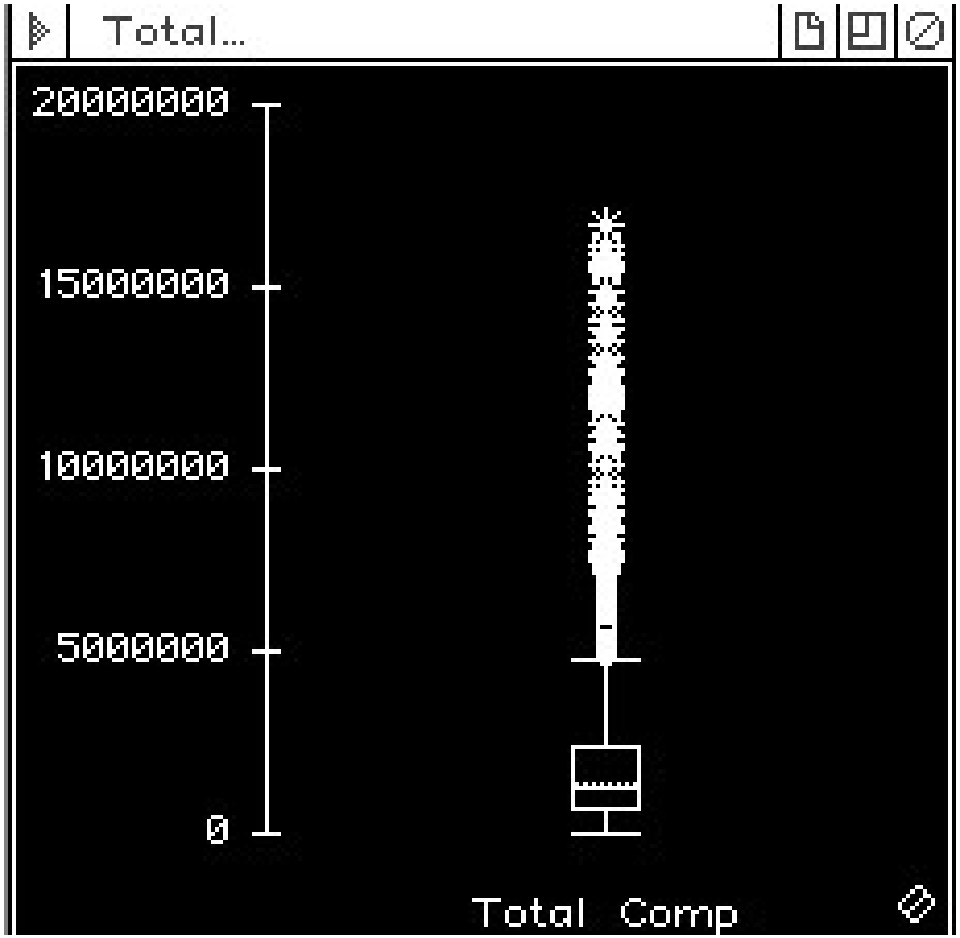
Transforming

Preferred Measures

Outliers

Stemplots

CEO Salaries 1994 Boxplot



# Comparing Measures of Center



| | n | **Mean $\bar{x}$** | **Median** |
|---|---|---|---|
| **All** | 790 | 2.82M | 1.304M |
| **Without top 9** | 781 | 2.24M | 1.296M |

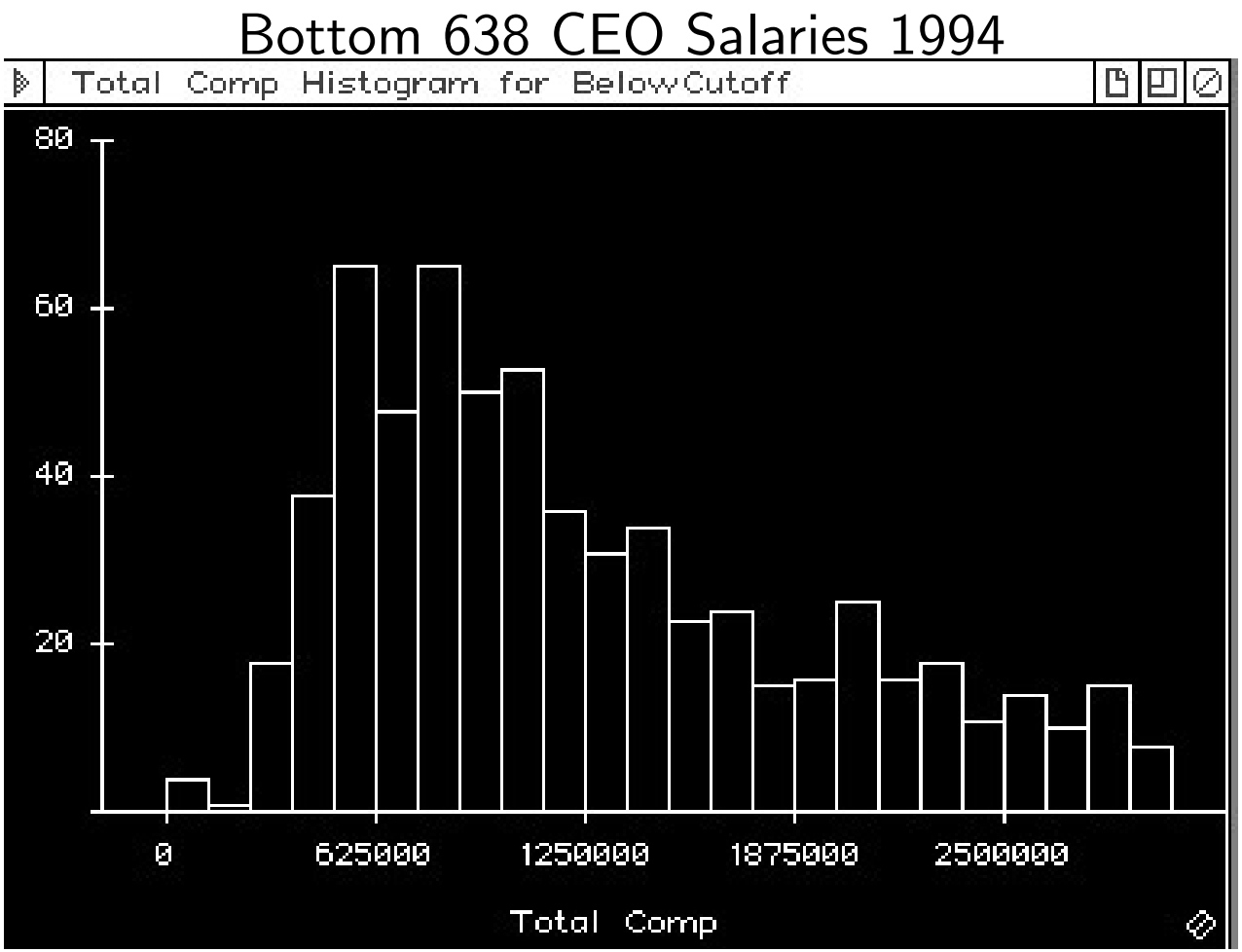
# Comparing Measures of Spread



| | n | **Std. Dev.s** | **IQR** |
|---|---|---|---|
| **All** | 790 | 8.32M | 1.731M |
| **Without top 9** | 781 | 2.724M | 1.662M |

# Forbes




Bottom 638 CEO Salaries 1994
Total Comp Histogram for BelowCutoff
80
60
40
20
0
625000
1250000
1875000
2500000
Total Comp

# Forbes



## Bottom 638 CEO Salaries 1994

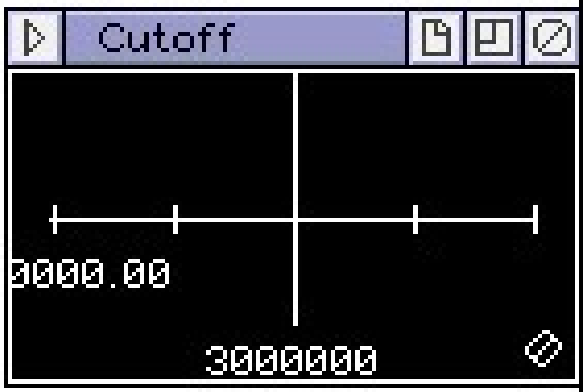
Total Comp

Summary of
cases selected according to
800 total cases of which 162 are missing

**Total Comp**
**BelowCutoff**

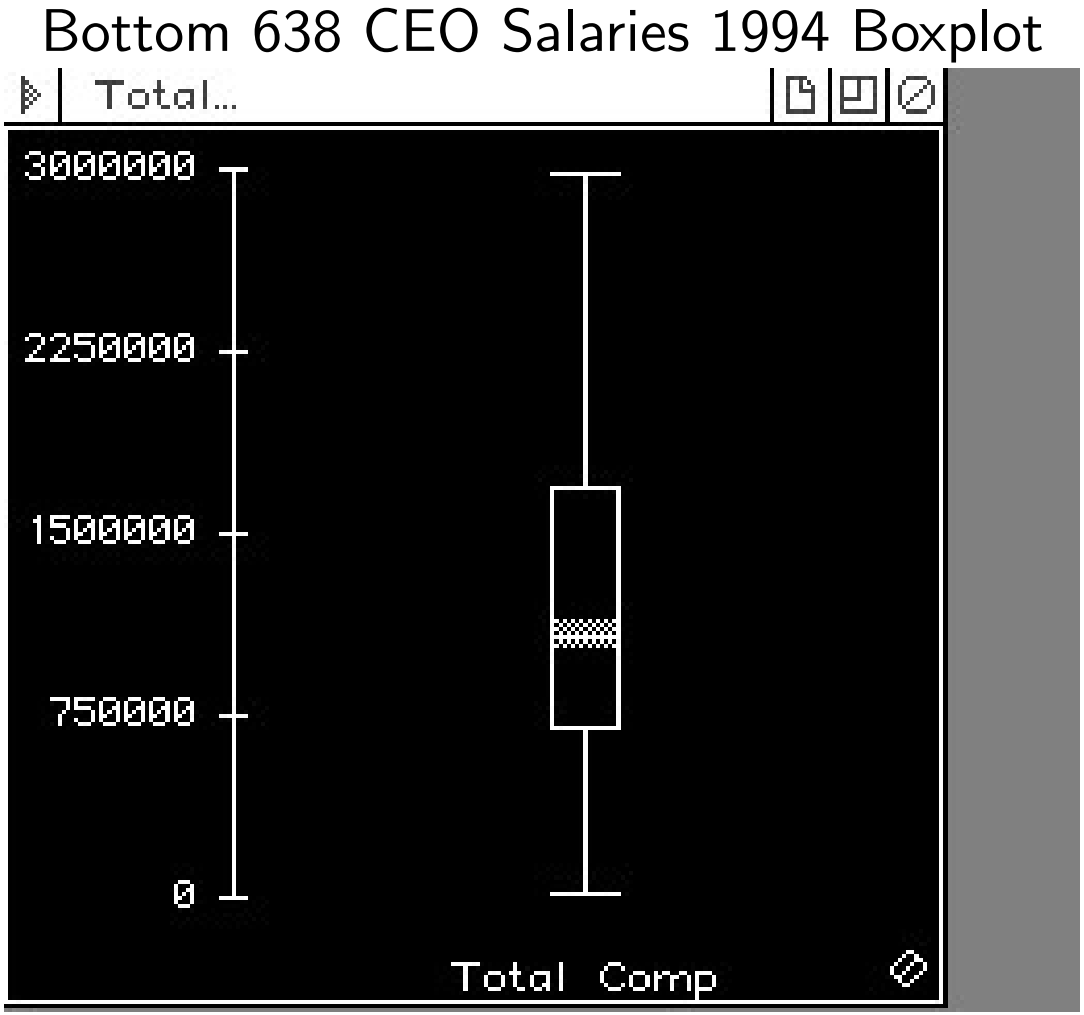
Percentile 25

| | |
|---|---|
| **Count** | 638 |
| **Mean** | 1.2529e6 |
| **Median** | 1.08931e6 |
| **StdDev** | 689151 |
| **Range** | 2.96418e6 |
| **IntQRange** | 977697 |
| **Lower ith %tile** | 715623 |
| **Upper ith %tile** | 1.69332e6 |



# Forbes




Bottom 638 CEO Salaries 1994 Boxplot
Total...
3000000
2250000
1500000
750000
0
Total Comp

# Comparing 84th Percentiles



| | n | s | $\bar{x}$ | **84%ile** | $\bar{x} + s$ |
|---|---|---|---|---|---|
| **All** | 790 | 8.32M | 2.82M | 3.392M | 11.14M |
| **Without top 9** | 781 | 2.724M | 2.24M | 3.347M | 4.97M |
| **Bottom 638** | 638 | .689M | 1.25M | 2.076M | 1.94M |

# Comparing 97.7%iles



| | n | s | $\bar{x}$ | **97.7%ile** | $\bar{x} + 2s$ |
|---|---|---|---|---|---|
| **All** | 790 | 8.32M | 2.82M | 14.85M | 19.46M |
| **Without top 9** | 781 | 2.724M | 2.24 | 12.33M | 7.68M |
| **Bottom 638** | 638 | .689M | 1.25M | 2.79M | 2.63M |

# Comparing IQR's



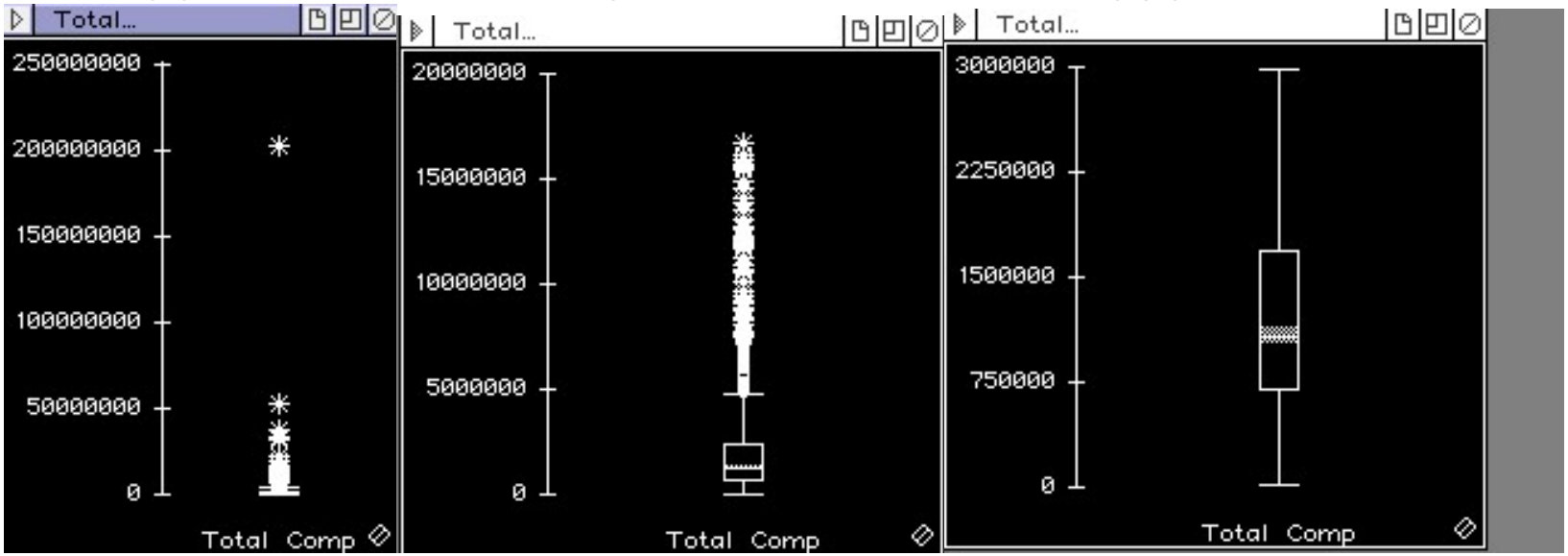
For a normal distribution $N(\mu, \sigma)$, the quartiles are at $\mu \pm .675^{-}\sigma$, so the IQR would be at $1.35\sigma$.

# Comparing IQR's



For a normal distribution $N(\mu, \sigma)$,
the quartiles are at $\mu \pm .675^{-}\sigma$,
so the IQR would be at $1.35\sigma$.

$n = 790$ $n = 781$ $n = 638$

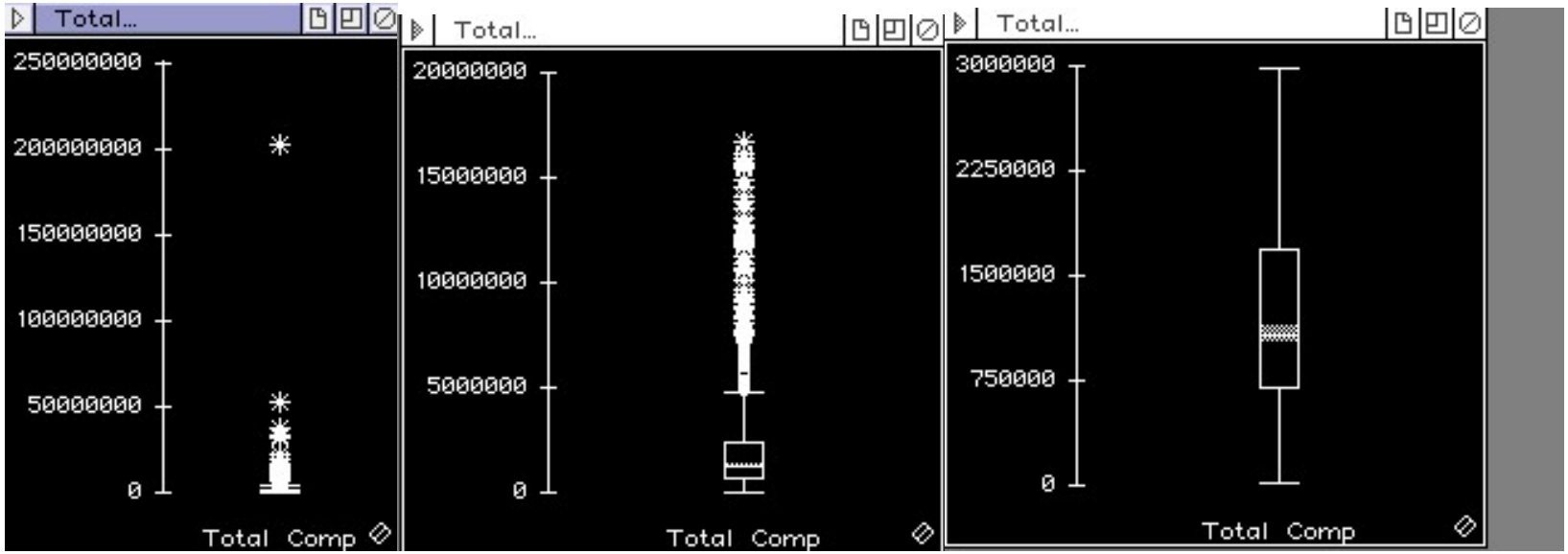


# Comparing IQR's



For a normal distribution $N(\mu, \sigma)$,
the quartiles are at $\mu \pm .675^{-}\sigma$,
so the IQR would be at $1.35\sigma$.

$n = 790$ $n = 781$ $n = 638$




Both $n = 790$ and $n = 781$ are strongly skewed to the right with lots of outliers.

# Comparing IQR's



For a normal distribution $N(\mu, \sigma)$,
the quartiles are at $\mu \pm .675^{-}\sigma$,
so the IQR would be at $1.35\sigma$.

| | n | **Std. Dev. s** | $\bar{x}$ | **IQR** | $1.35s$ |
|---|---|---|---|---|---|
| **All** | 790 | 8.32M | 2.82M | 1.73M | 11.23M |
| **Without top 9** | 781 | 2.724M | 2.24M | 1.66M | 3.68M |
| **Bottom 638** | 638 | .689M | 1.25M | .979M | .930M |

# Comparing IQR's



For a normal distribution $N(\mu, \sigma)$,
the quartiles are at $\mu \pm .675^{-}\sigma$,
so the IQR would be at $1.35\sigma$.

| | n | **Std. Dev. s** | $\bar{x}$ | **IQR** | $1.35s$ |
|---|---|---|---|---|---|
| **All** | 790 | 8.32M | 2.82M | 1.73M | 11.23M |
| **Without top 9** | 781 | 2.724M | 2.24M | 1.66M | 3.68M |
| **Bottom 638** | 638 | .689M | 1.25M | .979M | .930M |

Agreement of $1.35s$ with IQR is poor for the first two.

# Comparing IQR's



For a normal distribution $N(\mu, \sigma)$,
the quartiles are at $\mu \pm .675^{-}\sigma$,
so the IQR would be at $1.35\sigma$.

| | n | **Std. Dev. s** | $\bar{x}$ | **IQR** | $1.35s$ |
|---|---|---|---|---|---|
| **All** | 790 | 8.32M | 2.82M | 1.73M | 11.23M |
| **Without top 9** | 781 | 2.724M | 2.24M | 1.66M | 3.68M |
| **Bottom 638** | 638 | .689M | 1.25M | .979M | .930M |

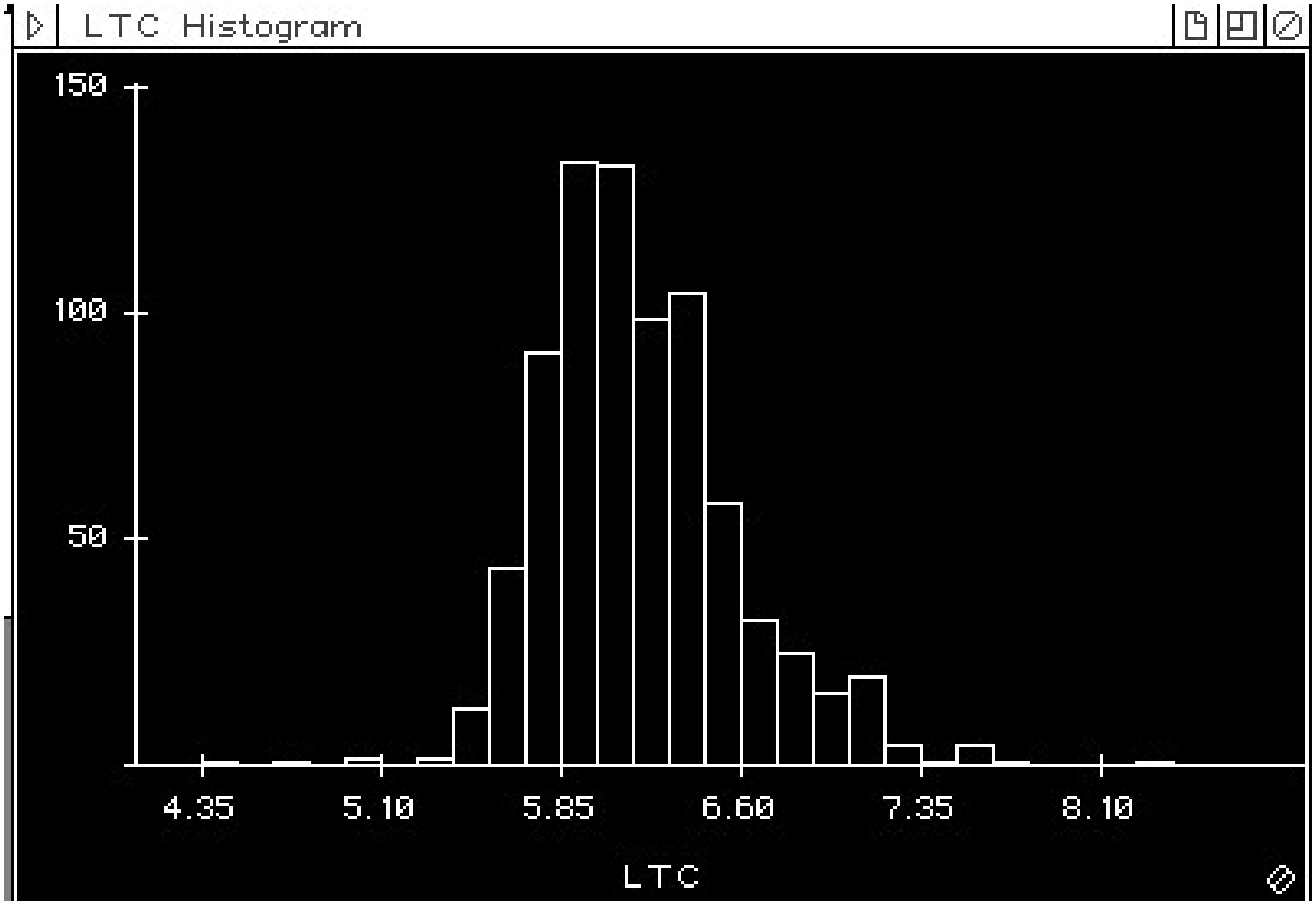
Agreement of $1.35s$ with IQR is poor for the first two.

Agreement of $1.35s$ with IQR is good when $n = 638$.

# Forbes



Logarithms of all 790 Salaries



# Forbes



## Logarithms of all 790 Salaries

LTC

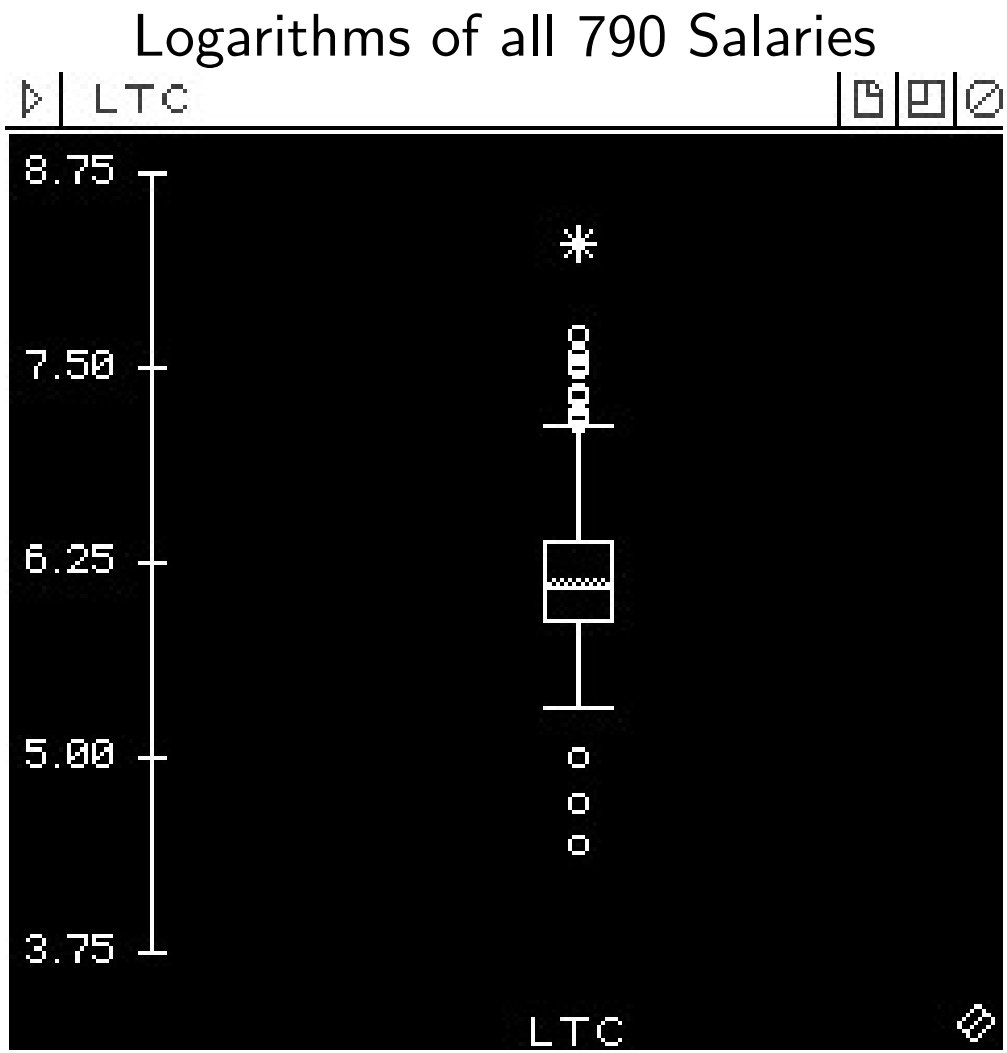
Summary of LTC
No Selector
800 total cases of which 10 are missing

Percentile 25

| | |
|---|---|
| **Count** | 790 |
| **Mean** | 6.17778 |
| **Median** | 6.11543 |
| **StdDev** | 0.412134 |
| **Range** | 3.84791 |
| **IntQRange** | 0.504141 |
| **Lower ith %tile** | 5.89644 |
| **Upper ith %tile** | 6.40058 |

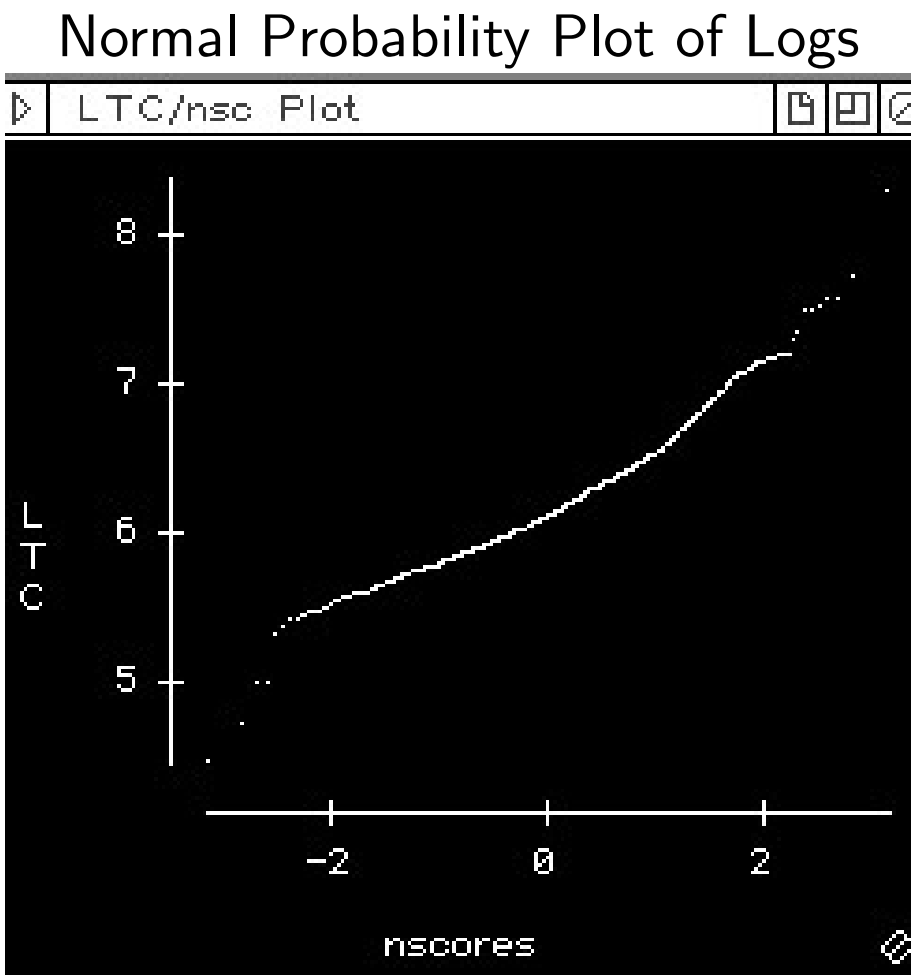
# Forbes





# Forbes



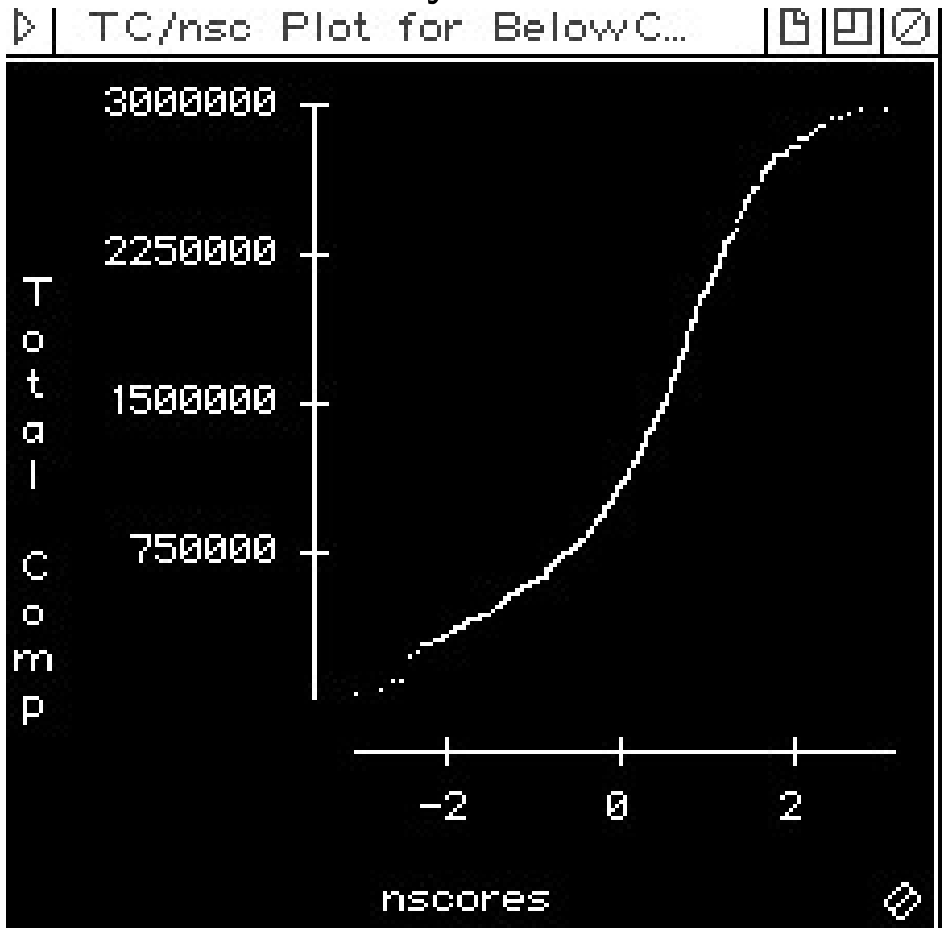



Straighter means closer to normal.

# Forbes



Normal Probability Plot of Bottom 638



# Preferred Measures of Center and Spread



Your textbook suggests

- Symmetric Distributions: mean and std. dev.
- Skewed Distributions: median and IQR.

# Preferred Measures of Center and Spread



Your textbook suggests

- Symmetric Distributions: mean and std. dev.
- Skewed Distributions: median and IQR.

This recommendation works well in the CEO data set case.

# Preferred Measures of Center and Spread



Your textbook suggests

- Symmetric Distributions: mean and std. dev.
- Skewed Distributions: median and IQR.

This recommendation works well in the CEO data set case.

Here are some examples where it doesn't:

# How Many Times Married?



People Aged 20-24.

# How Many Times Married?



People Aged 20-24.

| # Times | Women | Men |
|---|---|---|
| 0 | 5861 | 7024 |
| 1 | 2773 | 1541 |
| 2 | 105 | 43 |

# How Many Times Married?



People Aged 20-24.

| # Times | Women | Men |
|---|---|---|
| 0 | 5861 | 7024 |
| 1 | 2773 | 1541 |
| 2 | 105 | 43 |

Skewed to the right.
(Might be viewed as categorical rather than quantitative.)

# How Many Times Married?



| # Times | Women | Men |
|---|---|---|
| 0 | 5861 | 7024 |
| 1 | 2773 | 1541 |
| 2 | 105 | 43 |

| | Women | Men |
|---|---|---|
| **median** | 0 | 0 |
| $\bar{x}$ | .34 | .19 |
| $Q_1$ | 0 | 0 |
| $Q_3$ | 1 | 0 |
| IQR | 1 | 0 |
| $s$ | .499 | .403 |

A skewed distribution. Still the mean has some advantages.

# # People You've Known who were Victims of Homicide



4390 subjects

# # People You've Known who were Victims of Homicide



4390 subjects

| Number | Frequency |
|---|---|
| 0 | 3944 |
| 1 | 279 |
| 2 | 97 |
| 3 | 40 |
| 4 | 30 |

*(Actually the 4 is 4 or more.)*

# # People You've Known who were Victims of Homicide



4390 subjects

| Number | Frequency |
|---|---|
| 0 | 3944 |
| 1 | 279 |
| 2 | 97 |
| 3 | 40 |
| 4 | 30 |

*(Actually the 4 is 4 or more.)*

A shift of 1744 from 0 to 4 would not change the median.

# # People You've Known who were Victims of Homicide



4390 subjects

| Number | Frequency |
|---|---|
| 0 | 3944 |
| 1 | 279 |
| 2 | 97 |
| 3 | 40 |
| 4 | 30 |

*(Actually the 4 is 4 or more.)*

A shift of 1744 from 0 to 4 would not change the median.

Strongly skewed.

# # People You've Known who were Victims of Homicide



4390 subjects

| Number | Frequency |
|---|---|
| 0 | 3944 |
| 1 | 279 |
| 2 | 97 |
| 3 | 40 |
| 4 | 30 |

*(Actually the 4 is 4 or more.)*

A shift of 1744 from 0 to 4 would not change the median.

Strongly skewed.

Yet the mean has some advantages over the median here . . . .

# # People You've Known who were Victims of Homicide



4390 subjects

| Number | Frequency |
|---|---|
| 0 | 3944 |
| 1 | 279 |
| 2 | 97 |
| 3 | 40 |
| 4 | 30 |

*(Actually the 4 is 4 or more.)*

| | |
|---|---|
| **median** | 0 |
| $\bar{x}$ | .162 |
| $Q_1$ | 0 |
| $Q_3$ | 0 |
| IQR | 0 |
| $s$ | .563 |

# Just barely an ordinary outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,46

# Just barely an ordinary outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17, 18,19,20,21,22,23,24,25,26,27,28,29,46

| **median** | 15.5 |
|---|---|
| $Q_1$ | 8 |
| $Q_3$ | 23 |
| IQR | 15 |

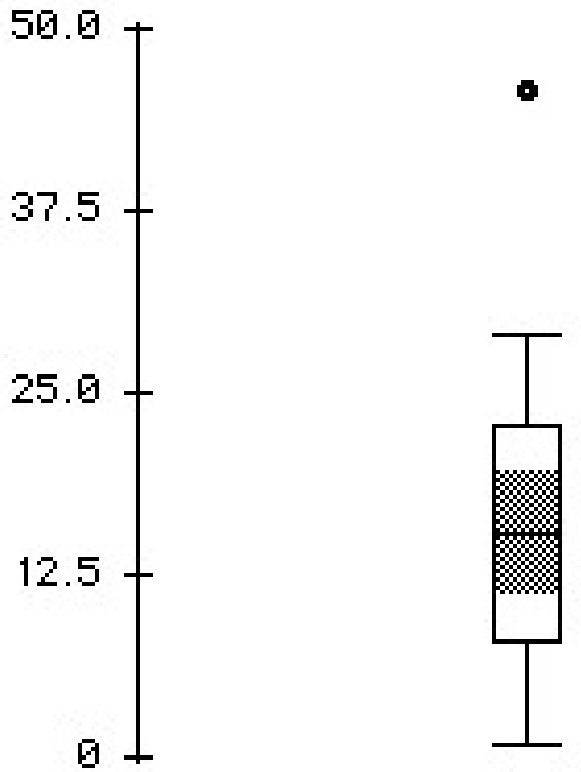
# Just barely an ordinary outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,46

$$Q_3 + 1.5 \cdot \text{IQR} = 23 + 22.5 = 45.5$$



# Just barely an ordinary outlier:



The data set with 45 instead:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,45

# Just barely an ordinary outlier:



The data set with 45 instead:
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17, 18,19,20,21,22,23,24,25,26,27,28,29,45

| **median** | 15.5 |
|---|---|
| $Q_1$ | 8 |
| $Q_3$ | 23 |
| IQR | 15 |

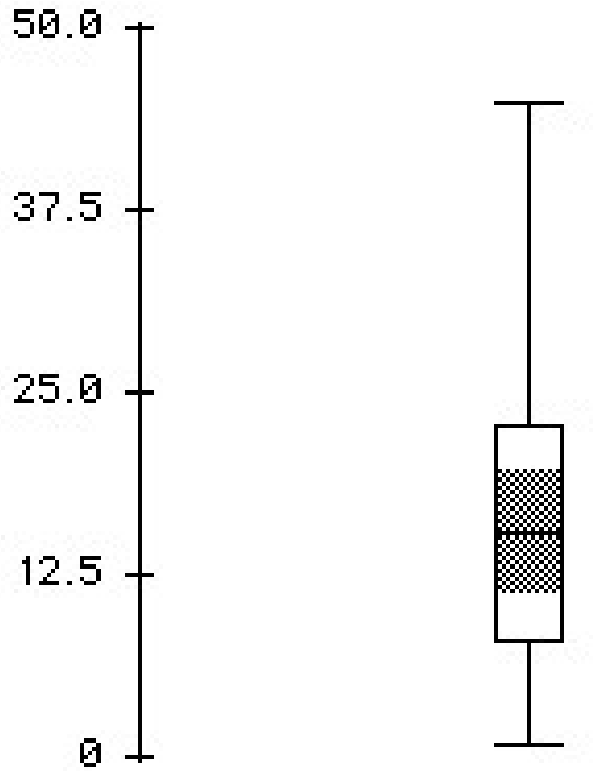
# Just barely an ordinary outlier:



The data set with 45 instead:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,45

$$Q_3 + 1.5 \cdot \text{IQR} = 23 + 22.5 = 45.5$$



# Just barely an extreme outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,69

# Just barely an extreme outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17, 18,19,20,21,22,23,24,25,26,27,28,29,69

| **median** | 15.5 |
|---|---|
| $Q_1$ | 8 |
| $Q_3$ | 23 |
| IQR | 15 |

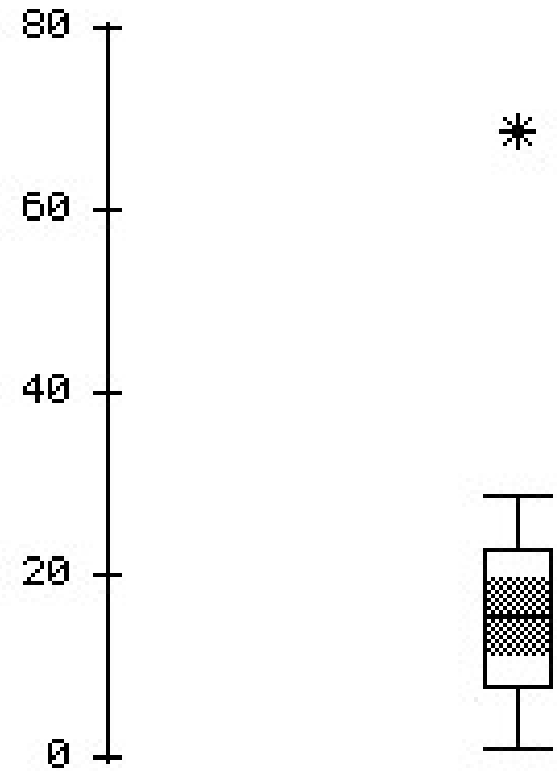
# Just barely an extreme outlier:



The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,69

$$Q_3 + 3 \cdot \text{IQR} = 23 + 45 = 68$$



# Just barely an extreme outlier:



The data set with 68 instead:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,68

# Just barely an extreme outlier:



The data set with 68 instead:
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17, 18,19,20,21,22,23,24,25,26,27,28,29,68

| **median** | 15.5 |
|---|---|
| $Q_1$ | 8 |
| $Q_3$ | 23 |
| IQR | 15 |

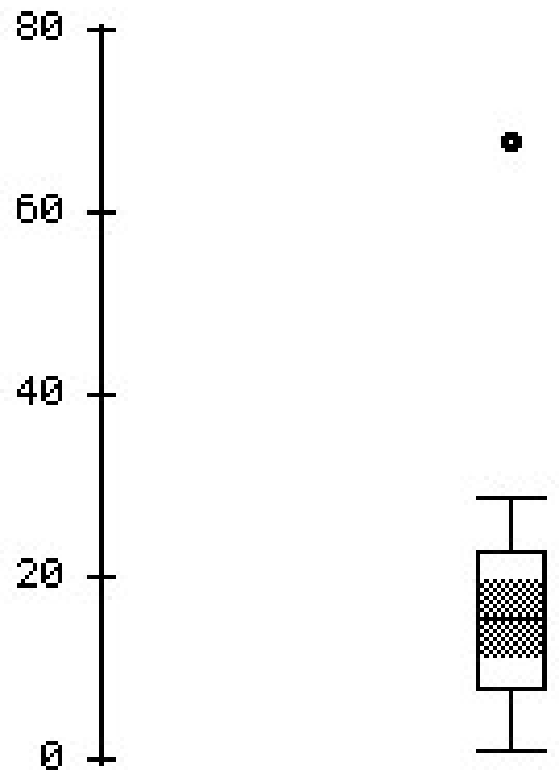
# Just barely an extreme outlier:



The data set with 68 instead:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,68

$$Q_3 + 3 \cdot \text{IQR} = 23 + 45 = 68$$



# Stemplots



## Roger Maris Yearly Home Run Totals

```
> maris <-c(14,9,19,28,16,39,61,33,23,26,8,13,9,5)
> stem(maris, scale=2)

  The decimal point is 1 digit(s) to the right of the |

  0 | 5899
  1 | 3469
  2 | 368
  3 | 39
  4 |
  5 |
  6 | 1
```

(A stemplot using R.)

# Stemplots



Stemplot Idea: First break each entry into a stem followed by a single digit leaf.

# Stemplots



Stemplot Idea: First break each entry into a stem followed by a single digit leaf.

You may have to round

or adjust the units of your numbers first.

# Stemplots



Stemplot Idea: First break each entry into a stem followed by a single digit leaf.
You may have to round
or adjust the units of your numbers first.
Then one row for each possible stem, with the leaves in increasing order.

# Stemplots



Willie Mays: 20, 4, 41, 51, 36, 35, 29, 34, 29, 40, 49, 38, 47, 52, 37, 22, 23, 13, 28, 18, 8, 6

Make a stemplot:

# Stemplots



Willie Mays: 20, 4, 41, 51, 36, 35, 29, 34, 29, 40, 49, 38, 47, 52, 37, 22, 23, 13, 28, 18, 8, 6

Make a stemplot:

First make a "pre-stemplot" where you don't worry about ordering the leaves.

# Stemplots



Willie Mays: 20, 4, 41, 51, 36, 35, 29, 34, 29, 40, 49, 38, 47, 52, 37, 22, 23, 13, 28, 18, 8, 6

Make a stemplot:

First make a "pre-stemplot" where you don't worry about ordering the leaves.

Pre-Stemplot

| stem | leaf |
|---|---|
| 0 | 4 8 6 |
| 1 | 3 8 |
| 2 | 0 9 9 2 3 8 |
| 3 | 6 5 4 8 7 |
| 4 | 1 0 9 7 |
| 5 | 1 2 |

# Stemplots



Willie Mays: 20, 4, 41, 51, 36, 35, 29, 34, 29, 40, 49, 38, 47, 52, 37, 22, 23, 13, 28, 18, 8, 6

Make a stemplot:

First make a "pre-stemplot" where you don't worry about ordering the leaves.

Pre-Stemplot

| stem | leaf |
|---|---|
| 0 | 4 8 6 |
| 1 | 3 8 |
| 2 | 0 9 9 2 3 8 |
| 3 | 6 5 4 8 7 |
| 4 | 1 0 9 7 |
| 5 | 1 2 |

Now rearrange the leaves.

# Stemplots



Pre-Stemplot

| stem | leaf |
|---|---|
| 0 | 4 8 6 |
| 1 | 3 8 |
| 2 | 0 9 9 2 3 8 |
| 3 | 6 5 4 8 7 |
| 4 | 1 0 9 7 |
| 5 | 1 2 |

Stemplot

| stem | leaf |
|---|---|
| 0 | 4 6 8 |
| 1 | 3 8 |
| 2 | 0 2 3 8 9 9 |
| 3 | 4 5 6 7 8 |
| 4 | 0 1 7 9 |
| 5 | 1 2 |

# Stemplots



Stemplot

| stem | leaf |
|---|---|
| 0 | 4 6 8 |
| 1 | 3 8 |
| 2 | 0 2 3 8 9 9 |
| 3 | 4 5 6 7 8 |
| 4 | 0 1 7 9 |
| 5 | 1 2 |

Willie Mays' season home run totals are unimodal with a center of about 30 home runs per year. (The median is 31.5.) Most seasons fall within approximately 20 home runs of the median. The distribution is somewhat skewed to the left, mostly due to a number of low season totals near the end of his career.

# Stemplots



Stemplot

| stem | leaf |
|---|---|
| 0 | 4 6 8 |
| 1 | 3 8 |
| 2 | 0 2 3 8 9 9 |
| 3 | 4 5 6 7 8 |
| 4 | 0 1 7 9 |
| 5 | 1 2 |

Hank Aaron: 13,27,26,44,30,39,40,34,45,44,
24,32,44,39,29,44,38,47,34,40,20,12,10,25

One can do "back to back stemplots:"

# Stemplots



Pre-Stemplot

| Aaron leaf | stem | Mays leaf |
|---|---|---|
| | 0 | 4 6 8 |
| 0 2 3 | 1 | 3 8 |
| 5 0 9 4 6 7 | 2 | 0 2 3 8 9 9 |
| 4 8 9 2 4 9 0 | 3 | 4 5 6 7 8 |
| 0 7 4 4 4 5 0 4 | 4 | 0 1 7 9 |
| | 5 | 1 2 |

# Stemplots



Pre-Stemplot

| Aaron leaf | stem | Mays leaf |
|---|---|---|
| | 0 | 4 6 8 |
| 0 2 3 | 1 | 3 8 |
| 5 0 9 4 6 7 | 2 | 0 2 3 8 9 9 |
| 4 8 9 2 4 9 0 | 3 | 4 5 6 7 8 |
| 0 7 4 4 4 5 0 4 | 4 | 0 1 7 9 |
| | 5 | 1 2 |

Stemplot

| Aaron leaf | stem | Mays leaf |
|---|---|---|
| | 0 | 4 6 8 |
| 3 2 0 | 1 | 3 8 |
| 9 7 6 5 4 0 | 2 | 0 2 3 8 9 9 |
| 9 9 8 4 4 2 0 | 3 | 4 5 6 7 8 |
| 7 5 4 4 4 4 0 0 | 4 | 0 1 7 9 |
| | 5 | 1 2 |

# Stemplots



Stemplot

| Aaron leaf | stem | Mays leaf |
|---|---|---|
| | 0 | 4 6 8 |
| 3 2 0 | 1 | 3 8 |
| 9 7 6 5 4 0 | 2 | 0 2 3 8 9 9 |
| 9 9 8 4 4 2 0 | 3 | 4 5 6 7 8 |
| 7 5 4 4 4 4 0 0 | 4 | 0 1 7 9 |
| | 5 | 1 2 |

Aaron's season home run totals have less spread than those of Mays. This is evident graphically or can be seen by comparing Aaron's IQR of 16.5 to Mays' 20. Aaron's median of 34 is somewhat higher than Mays' median of 31.5. Both distributions are skewed to the left.