Math 1710 Class 12

Normal Distributions, Outliers, and Summary Statistics
Dr. Back

Sep. 23, 2009
Per Capita CO$_2$ Emissions

Units are metric tons per person per year.
8 Most Populous Countries in the World:

<table>
<thead>
<tr>
<th>Country</th>
<th>tons/yr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>2.3</td>
</tr>
<tr>
<td>India</td>
<td>1.1</td>
</tr>
<tr>
<td>US</td>
<td>19.7</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1.2</td>
</tr>
<tr>
<td>Brazil</td>
<td>1.8</td>
</tr>
<tr>
<td>Russia</td>
<td>9.8</td>
</tr>
<tr>
<td>Pakistan</td>
<td>.7</td>
</tr>
<tr>
<td>?</td>
<td>.2</td>
</tr>
</tbody>
</table>
### Per Capita CO₂ Emissions

8 Most Populous Countries in the World:

<table>
<thead>
<tr>
<th>Country</th>
<th>tons/yr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>2.3</td>
</tr>
<tr>
<td>India</td>
<td>1.1</td>
</tr>
<tr>
<td>US</td>
<td>19.7</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1.2</td>
</tr>
<tr>
<td>Brazil</td>
<td>1.8</td>
</tr>
<tr>
<td>Russia</td>
<td>9.8</td>
</tr>
<tr>
<td>Pakistan</td>
<td>0.7</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Per Capita CO₂ Emissions

In order:

\textbf{tons/yr} : \begin{align*}
.2 & \quad .7 & \quad 1.1 & \quad 1.2 & \quad 1.8 & \quad 2.3 & \quad 9.8 & \quad 19.7
\end{align*}
In order with positions: \((n = 8)\)

\[
\begin{array}{cccccccc}
\text{tons/yr} & : & .2 & .7 & 1.1 & 1.2 & 1.8 & 2.3 & 9.8 & 19.7 \\
\text{Posn.} & : & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8
\end{array}
\]
Per Capita CO₂ Emissions

In order with positions: \((n = 8)\)

<table>
<thead>
<tr>
<th>tons/yr</th>
<th>0.2</th>
<th>0.7</th>
<th>1.1</th>
<th>1.2</th>
<th>1.8</th>
<th>2.3</th>
<th>9.8</th>
<th>19.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posn.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

The *median* is the middle value. A basic measure of center. 
*When the sample size is even, we average the two middle values.*
Per Capita CO\textsubscript{2} Emissions

The *median* is the middle value. A basic measure of center. **When the sample size is even, we average the two middle values.**

Sample size \((n = 8)\)

\[
\text{tons/yr} : \quad .2 \quad .7 \quad 1.1 \quad 1.2 \quad 1.8 \quad 2.3 \quad 9.8 \quad 19.7
\]

\[
\text{Posn.} : \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8
\]

\[
\text{median} = \frac{1.8 + 1.2}{2} = 1.5.
\]
Per Capita CO$_2$ Emissions

tons/yr: 0.2 0.7 1.1 1.2 1.8 2.3 9.8 19.7

Histogram of all 8
Per Capita CO$_2$ Emissions

**tons/yr**: \(0.2, 0.7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7\)

The US value of 19.7 is not in keeping with the rest if the data. Such a value is called an outlier.
Per Capita CO$_2$ Emissions

The US value of 19.7 is not in keeping with the rest of the data. Such a value is called an outlier.

Histogram of all but US. (n=7)
The US value of 19.7 is not in keeping with the rest of the data. Such a value is called an outlier.

Removal of the outlier gives a much more revealing histogram.
Per Capita CO\textsubscript{2} Emissions

Without the outlier: \((n = 7)\)

\[
\begin{array}{cccccccc}
\text{tons/yr} & 0.2 & 0.7 & 1.1 & 1.2 & 1.8 & 2.3 & 9.8 \\
\text{Posn.} & 1 & 2 & 3 & 4 & 5 & 6 & 7
\end{array}
\]

With \(n\) odd, the median is just the middle value of 1.2.
### Per Capita CO$_2$ Emissions

**Without the outlier:** \((n = 7)\)

<table>
<thead>
<tr>
<th>tons/yr</th>
<th>0.2</th>
<th>0.7</th>
<th>1.1</th>
<th>1.2</th>
<th>1.8</th>
<th>2.3</th>
<th>9.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posn.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

With \(n\) odd, the median is just the middle value of 1.2.
The first quartile $Q_1$ or 25th percentile is defined to be the median of the bottom half of our data.
The first quartile $Q_1$ or 25th percentile is defined to be the median of the bottom half of our data. For a data set of odd sample size, we do not include the median in the bottom half:
The first quartile $Q_1$ or 25th percentile is defined to be the median of the bottom half of our data. For a data set of odd sample size, we do not include the median in the bottom half:
Without the outlier: ($n = 7$)

<table>
<thead>
<tr>
<th>tons/yr</th>
<th>.2</th>
<th>.7</th>
<th>1.1</th>
<th>1.2</th>
<th>1.8</th>
<th>2.3</th>
<th>9.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posn.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

$Q_1 = .7$
The first quartile $Q_1$ or 25th percentile is defined to be the median of the bottom half of our data. For a data set of odd sample size, we do not include the median in the bottom half:
Without the outlier: ($n = 7$)

\[
\begin{array}{cccccccc}
\text{tons/yr} & .2 & .7 & 1.1 & 1.2 & 1.8 & 2.3 & 9.8 \\
\text{Posn.} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\end{array}
\]

$Q_1 = .7$
The first quartile $Q_1$ or 25th percentile is defined to be the median of the bottom half of our data. For a data set of odd sample size, we do not include the median in the bottom half:
Without the outlier: ($n = 7$)

$$\begin{align*}
\text{tons/yr} & : \ 0.2 \ 0.7 \ 1.1 \ 1.2 \ 1.8 \ 2.3 \ 9.8 \\
\text{Posn.} & : \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7
\end{align*}$$

$$Q_1 = 0.7$$

(And $Q_3 = 2.3$.)
The convention about not including the middle changed in the 3rd edition of our text.
Per Capita CO₂ Emissions

For the original data set: \((n = 8)\)

\[
\text{tons/yr} : \quad 0.2 \quad 0.7 \quad 1.1 \quad 1.2 \quad 1.8 \quad 2.3 \quad 9.8 \quad 19.7
\]

\[
\text{Posn.} : \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8
\]

\[
Q_1 = \frac{0.7 + 1.1}{2} = 0.9 \quad \text{and} \quad Q_3 = \frac{2.3 + 9.8}{2} = 6.05
\]
Per Capita CO$_2$ Emissions

The 5-number summary:

<table>
<thead>
<tr>
<th></th>
<th>min,</th>
<th>$Q_1$,</th>
<th>median,</th>
<th>$Q_3$,</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 8</td>
<td>.2</td>
<td>.9</td>
<td>1.5</td>
<td>6.05</td>
<td>19.7</td>
</tr>
<tr>
<td>w/o US</td>
<td>.2</td>
<td>.7</td>
<td>1.2</td>
<td>2.3</td>
<td>9.8</td>
</tr>
</tbody>
</table>
Per Capita CO$_2$ Emissions

The 5-number summary:

- **all 8**: min 0.2, Q$_1$ 0.9, median 1.5, Q$_3$ 6.05, max 19.7
- **w/o US**: min 0.2, Q$_1$ 0.7, median 1.2, Q$_3$ 2.3, max 9.8

Boxplot - Graphical form of the 5 number summary:

- **all - n=8**
The 5-number summary:

- **min:** 0.2, 0.7, 1.2, 2.3, 9.8
- **Q1:** 0.2, 0.9, 1.5, 6.05, 19.7
- **median:** 0.2, 0.9, 1.5, 6.05, 19.7
- **Q3:** 0.2, 0.9, 1.5, 6.05, 19.7
- **max:** 0.2, 0.9, 1.5, 6.05, 19.7

Boxplot - Graphical form of the 5 number summary:

**Without the Outlier - n=7**

![Boxplot Image]
Per Capita CO$_2$ Emissions

Interquartile Range: $\text{IQR} = Q_3 - Q_1$.
A basic measure of spread.
Interquartile Range: $\text{IQR} = Q_3 - Q_1$.
A basic measure of spread.
The median and IQR are usually little affected by outliers.
“Resistant to Outliers”
Interquartile Range: \( \text{IQR} = Q_3 - Q_1 \).

A basic measure of spread.

The median and IQR are usually little affected by outliers.

“Resistant to Outliers”

<table>
<thead>
<tr>
<th>( n )</th>
<th>( Q_1 )</th>
<th>( \text{med} )</th>
<th>( Q_3 )</th>
<th>( \text{IQR} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>with outlier</td>
<td>8</td>
<td>.9</td>
<td>1.5</td>
<td>6.05</td>
</tr>
<tr>
<td>w/o outlier</td>
<td>7</td>
<td>.7</td>
<td>1.2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

It is mostly because of the small sample size that the median and IQR change as much as they do here.
Per Capita CO$_2$ Emissions

Spreadsheets do “wild” things when computing quartiles:

\[ \text{tons/yr: } .2 \quad .7 \quad 1.1 \quad 1.2 \quad 1.8 \quad 2.3 \quad 9.8 \quad 19.7 \]

Open Office, an Excel Clone

One way to get such numbers:
With 8 numbers there are 7 intervals in between.
.7 is the \( \frac{100}{7} \)%ile.
1.1 is the \( \frac{200}{7} \)%ile.

\[ 25 = \frac{1}{4} \cdot \frac{100}{7} + \frac{3}{4} \cdot \frac{200}{7} \]

So the 25th %ile is \( \frac{1}{4} \cdot .7 + \frac{3}{4} \cdot 1.1 = 1 \).
Forbes

790 CEO Salaries 1994

Total Comp Histogram

0 37500000 112500000 187500000

Total Comp

Proof of Normal Approximation

Empirical Rule and Outliers

A Simple Median and Quartile Example

Last Time
790 CEO Salaries 1994

Summary of
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
Int QRange 1.7274e6
Lower 1st %ile 787841
Upper 1st %ile 2.51524e6
A Simple Median and Quartile Example

Empirical Rule and Outliers

Proof of Normal Approximation

Forbes CEO Salaries 1994 Boxplot
All But Top 9 CEO Salaries 1994

Forbes

Math 1710
Class 12
V4

Last Time
A Simple
Median and
Quartile
Example

Empirical Rule
and Outliers

Proof of
Normal
Approximation
All But Top 9 CEO Salaries 1994

Summary of cases selected according to 800 total cases of which 19 are missing

Percentile 25

<table>
<thead>
<tr>
<th>Count</th>
<th>781</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.24289e6</td>
</tr>
<tr>
<td>Median</td>
<td>1.29613e6</td>
</tr>
<tr>
<td>StdDev</td>
<td>2.72437e6</td>
</tr>
<tr>
<td>Range</td>
<td>16.7457e6</td>
</tr>
<tr>
<td>IntQR</td>
<td>1.65856e6</td>
</tr>
<tr>
<td>Lower ith %tile</td>
<td>784797</td>
</tr>
<tr>
<td>Upper ith %tile</td>
<td>2.44336e6</td>
</tr>
</tbody>
</table>
CEO Salaries 1994 Boxplot
## Comparing Measures of Center

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>790</td>
<td>2.82M</td>
<td>1.304M</td>
</tr>
<tr>
<td><strong>Without top 9</strong></td>
<td>781</td>
<td>2.24M</td>
<td>1.296M</td>
</tr>
</tbody>
</table>

Means heavily affected by outliers. Medians resistant to outliers.
Comparing Measures of Center

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>2.82M</td>
<td>1.304M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.24M</td>
<td>1.296M</td>
</tr>
</tbody>
</table>

Means heavily affected by outliers.
## Comparing Measures of Center

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>790</td>
<td>2.82M</td>
<td>1.304M</td>
</tr>
<tr>
<td><strong>Without top 9</strong></td>
<td>781</td>
<td>2.24M</td>
<td>1.296M</td>
</tr>
</tbody>
</table>

Means heavily affected by outliers.

Medians resistant to outliers.
### Comparing Measures of Spread

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev.s</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>1.731M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>1.662M</td>
</tr>
</tbody>
</table>
Comparing Measures of Spread

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev.s</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>790</td>
<td>8.32M</td>
<td>1.731M</td>
</tr>
<tr>
<td><strong>Without top 9</strong></td>
<td>781</td>
<td>2.724M</td>
<td>1.662M</td>
</tr>
</tbody>
</table>

Std. Dev heavily affected by outliers.
Comparing Measures of Spread

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev.s</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>1.731M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>1.662M</td>
</tr>
</tbody>
</table>

Std. Dev heavily affected by outliers.

IQR resistant to outliers.
Forbes

Bottom 638 CEO Salaries 1994
Bottom 638 CEO Salaries 1994

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

Percentile 25

<table>
<thead>
<tr>
<th>Count</th>
<th>638</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.2529e6</td>
</tr>
<tr>
<td>Median</td>
<td>1.08931e6</td>
</tr>
<tr>
<td>StdDev</td>
<td>689151</td>
</tr>
<tr>
<td>Range</td>
<td>2.96418e6</td>
</tr>
<tr>
<td>IntQRRange</td>
<td>977697</td>
</tr>
<tr>
<td>Lower ith 25th</td>
<td>715623</td>
</tr>
<tr>
<td>Upper ith 25th</td>
<td>1.69332e6</td>
</tr>
</tbody>
</table>
Forbes

Bottom 638 CEO Salaries 1994 Boxplot
Comparing 84th Percentiles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + \sigma$. 
Comparing 84th Percentiles
Comparing 84th Percentiles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + \sigma$.

For $n = 790$, $n = 781$, and $n = 638$:

- Both $n = 790$ and $n = 781$ are strongly skewed to the right with lots of outliers.
Comparing 84th Percentiles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + \sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>84%ile</th>
<th>$\bar{x} + s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>3.392M</td>
<td>11.14M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>3.347M</td>
<td>4.97M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.076M</td>
<td>1.94M</td>
</tr>
</tbody>
</table>
Comparing 84th Percentiles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + \sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>84%ile</th>
<th>$\bar{x} + s$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>3.392M</td>
<td>11.14M</td>
</tr>
<tr>
<td><strong>Without top 9</strong></td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>3.347M</td>
<td>4.97M</td>
</tr>
<tr>
<td><strong>Bottom 638</strong></td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.076M</td>
<td>1.94M</td>
</tr>
</tbody>
</table>

Agreement of $\bar{x} + s$ with 84%ile is poor for the first two.
Comparing 84th Percentiles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + \sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>84%ile</th>
<th>$\bar{x} + s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>3.392M</td>
<td>11.14M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>3.347M</td>
<td>4.97M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.076M</td>
<td>1.94M</td>
</tr>
</tbody>
</table>

Agreement of $\bar{x} + s$ with 84%ile is poor for the first two.

Agreement of $\bar{x} + s$ with 84%ile is good when $n = 638$. 
Comparing IQR’s

For a normal distribution $N(\mu, \sigma)$, the quartiles are at $\mu \pm 0.675 \sigma$, so the IQR would be at $1.35\sigma$. 
For a normal distribution $N(\mu, \sigma)$, the quartiles are at $\mu \pm 0.675\sigma$, so the IQR would be at $1.35\sigma$.

$n = 790 \quad n = 781 \quad n = 638$
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 $1.35\sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev. s</th>
<th>$\bar{x}$</th>
<th>IQR</th>
<th>1.35s</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>1.73M</td>
<td>11.23M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>1.66M</td>
<td>3.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>.979M</td>
<td>.930M</td>
</tr>
</tbody>
</table>
Comparing IQR’s

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

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev. $s$</th>
<th>$\bar{x}$</th>
<th>IQR</th>
<th>$1.35s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>1.73M</td>
<td>11.23M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>1.66M</td>
<td>3.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>.979M</td>
<td>.930M</td>
</tr>
</tbody>
</table>

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 \( \mu \pm 0.675\sigma \), so the IQR would be at 1.35\( \sigma \).

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Std. Dev. s</th>
<th>( \bar{x} )</th>
<th>IQR</th>
<th>1.35s</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>1.73M</td>
<td>11.23M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24M</td>
<td>1.66M</td>
<td>3.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>.979M</td>
<td>.930M</td>
</tr>
</tbody>
</table>

Agreement of 1.35\( \sigma \) with IQR is poor for the first two.

Agreement of 1.35\( \sigma \) with IQR is good when \( n = 638 \).
Comparing 97.7%iles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$. 
Comparing 97.7%iles

\[ n = 790 \quad n = 781 \quad n = 638 \]
Comparing 97.7%iles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$.

$n = 790$  \hspace{1cm} $n = 781$  \hspace{1cm} $n = 638$

Both $n = 790$ and $n = 781$ are strongly skewed to the right with lots of outliers.
Comparing 97.7%iles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>97.7%ile</th>
<th>$\bar{x} + 2s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>14.85M</td>
<td>19.46M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24</td>
<td>12.33M</td>
<td>7.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.79M</td>
<td>2.63M</td>
</tr>
</tbody>
</table>
Comparing 97.7%iles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>97.7%ile</th>
<th>$\bar{x} + 2s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>14.85M</td>
<td>19.46M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24</td>
<td>12.33M</td>
<td>7.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.79M</td>
<td>2.63M</td>
</tr>
</tbody>
</table>

Agreement of $\bar{x} + 2s$ with 97.7%ile is poor for the first two.
Comparing 97.7%iles

For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>s</th>
<th>$\bar{x}$</th>
<th>97.7%ile</th>
<th>$\bar{x} + 2s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>790</td>
<td>8.32M</td>
<td>2.82M</td>
<td>14.85M</td>
<td>19.46M</td>
</tr>
<tr>
<td>Without top 9</td>
<td>781</td>
<td>2.724M</td>
<td>2.24</td>
<td>12.33M</td>
<td>7.68M</td>
</tr>
<tr>
<td>Bottom 638</td>
<td>638</td>
<td>.689M</td>
<td>1.25M</td>
<td>2.79M</td>
<td>2.63M</td>
</tr>
</tbody>
</table>

Agreement of $\bar{x} + 2s$ with 97.7%ile is poor for the first two.

Agreement of $\bar{x} + 2s$ with 97.7%ile is good when $n = 638$. 
Normal Distribution Formula

\[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \]
Normal Distribution Formula

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Normal Distribution Formula

\[ N(\mu, \sigma) \]

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

These formulas and the following argument are far above the basic level of our course.
Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$

$\mu = m$ and $\sigma = \sqrt{.5m}$. 
Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set
$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$
Suppose $X \sim \text{Binom}(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set
$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

The z-score of $m + k$ is $\frac{k\sqrt{2}}{\sqrt{m}}$. 
Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

The z-score of $m + k$ is $\frac{k\sqrt{2}}{\sqrt{m}}$.
So we want to show $a_k \sim ce^{-\frac{k^2}{m}}$. 
Why Normal Out of Binomial?

Suppose \( X \sim Binom(2m, .5) \)

We want to understand why \( X \sim N(m, \sqrt{.5m}) \) approximately.

Set

\[
 a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}
\]

So we want to show \( a_k \sim ce^{-\frac{k^2}{m}} \).

i.e. \( \ln a_k \sim \ln c - \frac{k^2}{m} \).
Why Normal Out of Binomial?

Suppose $X \sim Binom(2m, .5)$
We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.
Set
\[ a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m} \]

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

**Strategy:** Compare $a_k$ to $a_0$ using the approximation $\ln (1 + x) \sim x$ for $x$ small.
Suppose \( X \sim \text{Binom}(2m, .5) \)

We want to understand why \( X \sim N(m, \sqrt{.5m}) \) approximately.

Set

\[
a_k = P(X = m + k) = \binom{2m}{m + k} (.5)^{2m}
\]

i.e. \( \ln a_k \sim \ln c - \frac{k^2}{m} \).

\[
a_k = \frac{(2m)! (.5)^{2m}}{(m + k)!(m - k)!} = a_0 \frac{(m)(m - 1) \ldots (m - k + 1)}{(m + k)(m + k - 1) \ldots (m + 1)}
\]

\[
= a_0 \frac{(1)(1 - \frac{1}{m}) \ldots (1 - \frac{k-1}{m})}{(1 + \frac{k}{m})(1 + \frac{k-1}{m}) \ldots (1 + \frac{1}{m})}
\]
Why Normal Out of Binomial?

Suppose $X \sim \text{Binom}(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

$$a_k = \frac{(2m)!(.5)^{2m}}{(m+k)!(m-k)!} = a_0 \frac{(m)(m-1)\ldots(m-k+1)}{(m+k)(m+k-1)\ldots(m+1)}$$

$$= a_0 \frac{(1)(1-\frac{1}{m})\ldots(1-\frac{k-1}{m})}{(1+\frac{k}{m})(1+\frac{k-1}{m})\ldots(1+\frac{1}{m})}$$

So using $\ln (1 + x) \sim x$,

$$\ln a_k \sim \ln a_0 - 2 \left( \frac{1}{m} + \frac{2}{m} + \ldots + \frac{k-1}{m} \right) - \frac{k}{m}$$
Suppose $X \sim \text{Binom}(2m, .5)$

We want to understand why $X \sim N(m, \sqrt{.5m})$ approximately.

Set

$$a_k = P(X = m + k) = \binom{2m}{m+k} (.5)^{2m}$$

i.e. $\ln a_k \sim \ln c - \frac{k^2}{m}$.

So using $\ln (1 + x) \sim x$,

$$\ln a_k \sim \ln a_0 - 2 \left( \frac{1}{m} + \frac{2}{m} + \ldots + \frac{k-1}{m} \right) - \frac{k}{m}$$

But $1 + 2 + \ldots k - 1 = \frac{k(k-1)}{2}$, so
Why Normal Out of Binomial?

Suppose \( X \sim Binom(2m, .5) \)

We want to understand why \( X \sim N(m, \sqrt{.5m}) \) approximately.

Set

\[
a_k = P(X = m + k) = \binom{2m}{m+k}(.5)^{2m}
\]

i.e. \( \ln a_k \sim \ln c - \frac{k^2}{m} \).

So using \( \ln (1 + x) \sim x \),

\[
\ln a_k \sim \ln a_0 - 2 \left( \frac{1}{m} + \frac{2}{m} + \ldots + \frac{k-1}{m} \right) - \frac{k}{m}
\]

But \( 1 + 2 + \ldots + k - 1 = \frac{k(k-1)}{2} \), so

\[
\ln a_k \sim \ln a_0 - \frac{k^2}{m}
\]

as desired.
Last Time

A Simple Median and Quartile Example

Empirical Rule and Outliers

Proof of Normal Approximation