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.

Per Capita CO_2 Emissions

8 Most Populous Countries in the World:

Country	**tons/yr.**
China	2.3
India	1.1
US	19.7
Indonesia	1.2
Brazil	1.8
Russia	9.8
Pakistan	.7
?	.2

Per Capita CO_2 Emissions

8 Most Populous Countries in the World:

Country	**tons/yr.**
China	2.3
India	1.1
US	19.7
Indonesia	1.2
Brazil	1.8
Russia	9.8
Pakistan	.7
Bangladesh	.2

Per Capita CO_2 Emissions

In order:

$$\textbf{tons/yr}: \quad .2 \quad .7 \quad 1.1 \quad 1.2 \quad 1.8 \quad 2.3 \quad 9.8 \quad 19.7$$

Per Capita CO_2 Emissions

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

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8	19.7
Posn. :	1	2	3	4	5	6	7	8

Per Capita CO_2 Emissions

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

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8	19.7
Posn. :	1	2	3	4	5	6	7	8

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_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 $(n = 8)$

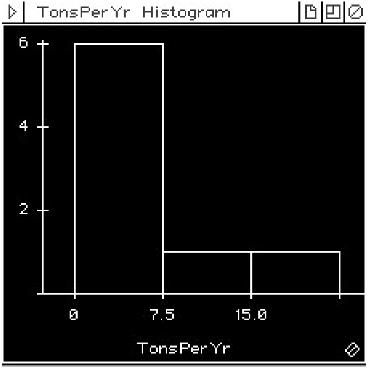
tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8	19.7
Posn. :	1	2	3	4	5	6	7	8

$$\text{median} = \frac{1.8 + 1.2}{2} = 1.5.$$

Per Capita CO_2 Emissions

tons/yr : .2 .7 1.1 1.2 1.8 2.3 9.8 19.7

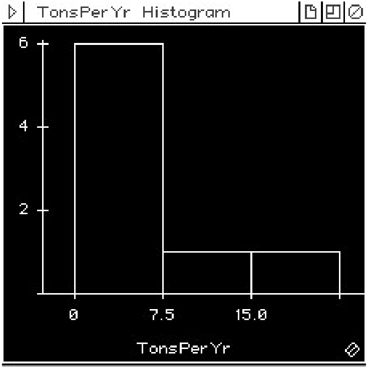
Histogram of all 8



Per Capita CO_2 Emissions

tons/yr : .2 .7 1.1 1.2 1.8 2.3 9.8 19.7

Histogram of all 8

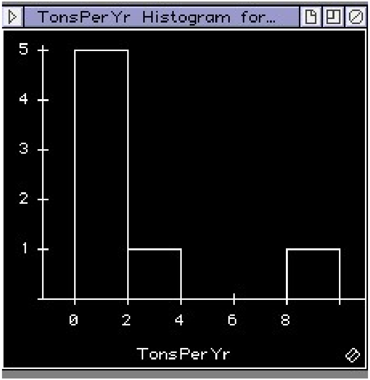


The US value of 19.7 is not in keeping with the rest if the data.
Such a value is called an outlier.

Per Capita CO_2 Emissions

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

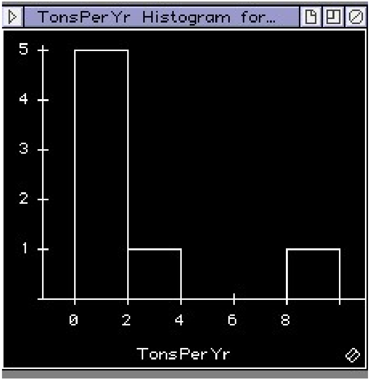
Histogram of all but US. (n=7)



Per Capita CO_2 Emissions

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

Histogram of all but US. (n=7)



Removal of the outlier gives a much more revealing histogram.

Per Capita CO_2 Emissions

Without the outlier: $(n = 7)$

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8
Posn. :	1	2	3	4	5	6	7

With n odd, the median is just the middle value of 1.2.

Per Capita CO_2 Emissions

Without the outlier: $(n = 7)$

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8
Posn. :	1	2	3	4	5	6	7

With n odd, the median is just the middle value of 1.2.

Per Capita CO_2 Emissions

The first quartile Q_1 or 25th percentile is defined to be the median of the bottom half of our data.

Per Capita CO_2 Emissions

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:

Per Capita CO_2 Emissions

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: $(n = 7)$

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8
Posn. :	1	2	3	4	5	6	7

$$Q_1 = .7$$

Per Capita CO_2 Emissions

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: $(n = 7)$

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8
Posn. :	1	2	3	4	5	6	7

$$Q_1 = .7$$

Per Capita CO_2 Emissions

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: $(n = 7)$

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8
Posn. :	1	2	3	4	5	6	7

$$Q_1 = .7$$

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

Per Capita CO_2 Emissions

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

tons/yr :	.2	.7	1.1	1.2	1.8	2.3	9.8	19.7
Posn. :	1	2	3	4	5	6	7	8

$$Q_1 = \frac{.7 + 1.1}{2} = .9 \text{ and } Q_3 = \frac{2.3 + 9.8}{2} = 6.05$$

Per Capita CO_2 Emissions

The 5-number summary:

	min,	Q_1,	median,	Q_3,	max
all 8	.2,	.9,	1.5,	6.05,	19.7
w/o US	.2,	.7,	1.2,	2.3,	9.8

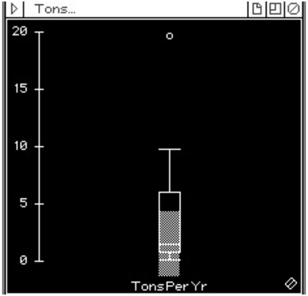
Per Capita CO_2 Emissions

The 5-number summary:

	min,	Q_1,	median,	Q_3,	max
all 8	.2,	.9,	1.5,	6.05,	19.7
w/o US	.2,	.7,	1.2,	2.3,	9.8

Boxplot - Graphical form of the 5 number summary:

all - n=8



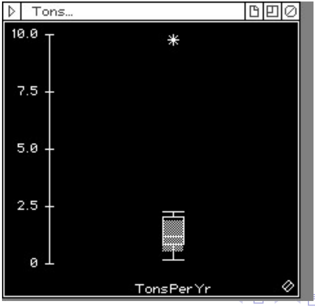
Per Capita CO_2 Emissions

The 5-number summary:

	min,	Q_1,	median,	Q_3,	max
all 8	.2,	.9,	1.5,	6.05,	19.7
w/o US	.2,	.7,	1.2,	2.3,	9.8

Boxplot - Graphical form of the 5 number summary:

Without the Outlier - n=7



Per Capita CO_2 Emissions

Interquartile Range: IQR$=Q_3 - Q_1$.
A basic measure of spread.

Per Capita CO_2 Emissions

Interquartile Range: IQR$=Q_3 - Q_1$.
A basic measure of spread.
The median and IQR are usually little affected by outliers.
"Resistant to Outliers"

Per Capita CO_2 Emissions

Interquartile Range: IQR$=Q_3 - Q_1$.
A basic measure of spread.
The median and IQR are usually little affected by outliers.
"Resistant to Outliers"

Here:

	n	Q_1	med	Q_3	IQR
with outlier	8	.9	1.5	6.05	5.15
w/o outlier	7	.7	1.2	2.3	1.6

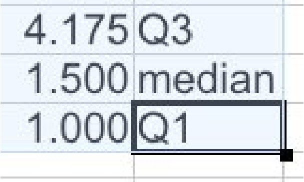
It is mostly because of the small sample size that the median and IQR change as much as they do here.

Per Capita CO_2 Emissions

Spreadsheets do "wild" things when computing quartiles:

tons/yr : .2 .7 1.1 1.2 1.8 2.3 9.8 19.7

Open Office, an Excel Clone



4.175	Q3
1.500	median
1.000	Q1

One way to get such numbers:

With 8 numbers there are 7 intervals in between.

.7 is the $\frac{100}{7}$ %ile.

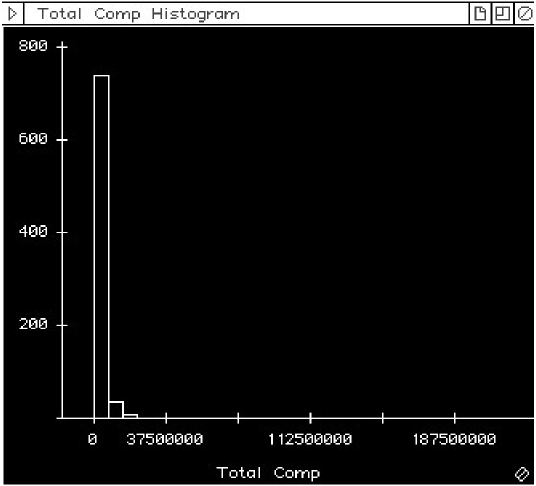
1.1 is the $\frac{200}{7}$ %ile.

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

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

Forbes

790 CEO Salaries 1994



Forbes

790 CEO Salaries 1994

Total Comp

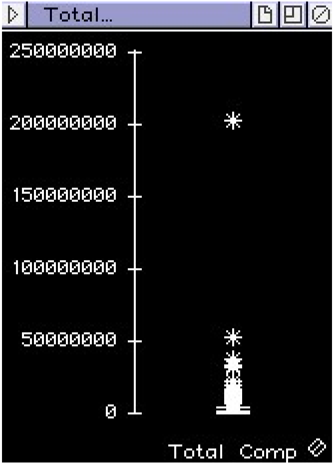
Summary of **Total Comp**
No Selector
800 total cases of which 10 are missing

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

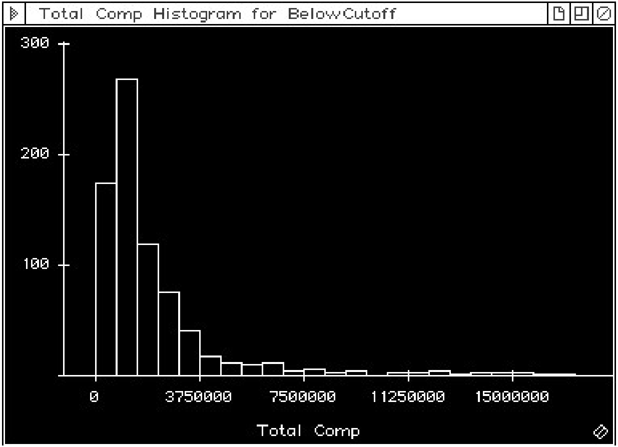
Forbes

CEO Salaries 1994 Boxplot



Forbes

All But Top 9 CEO Salaries 1994



Forbes

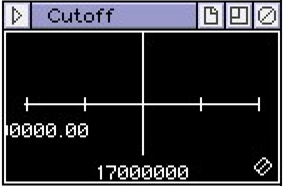
All But Top 9 CEO Salaries 1994

Total Comp

Summary of **Total Comp**
cases selected according to **BelowCutoff**
800 total cases of which 19 are missing

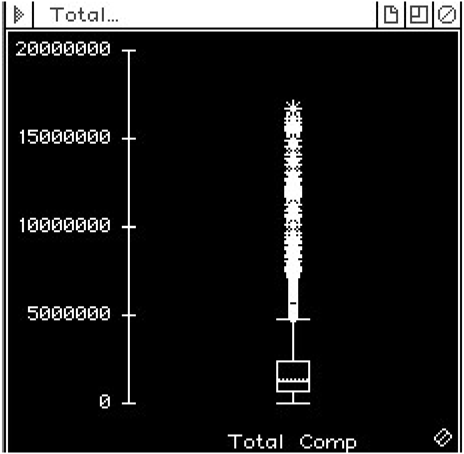
Percentile 25

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

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 Center

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

Means heavily affected by outliers.

Comparing Measures of Center

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

Means heavily affected by outliers.

Medians resistant to outliers.

Comparing Measures of Spread

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

Comparing Measures of Spread

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

Std. Dev heavily affected by outliers.

Comparing Measures of Spread

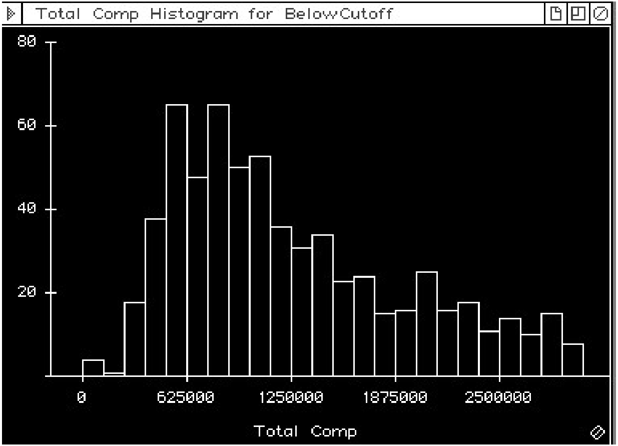
	n	**Std. Dev.s**	**IQR**
All	790	8.32M	1.731M
Without top 9	781	2.724M	1.662M

Std. Dev heavily affected by outliers.

IQR resistant to outliers.

Forbes

Bottom 638 CEO Salaries 1994



Forbes

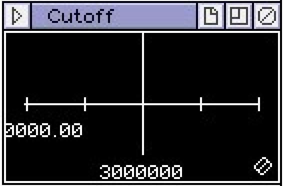
Bottom 638 CEO Salaries 1994

Total Comp

Summary of **Total Comp**
cases selected according to **BelowCutoff**
800 total cases of which 162 are missing

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

Math 1710
Class 12

V4

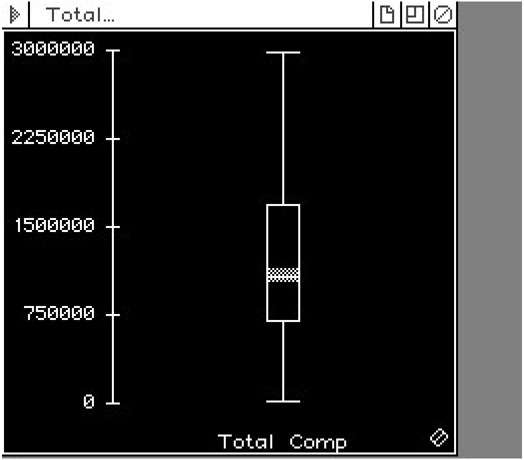
Last Time

A Simple Median and Quartile Example

Empirical Rule and Outliers

Proof of Normal Approximation

Bottom 638 CEO Salaries 1994 Boxplot

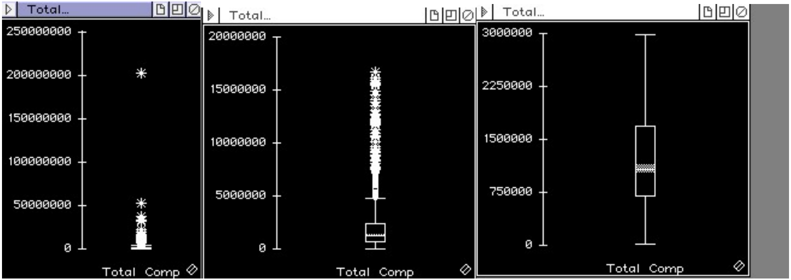


Comparing 84th Percentiles

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

Comparing 84th Percentiles

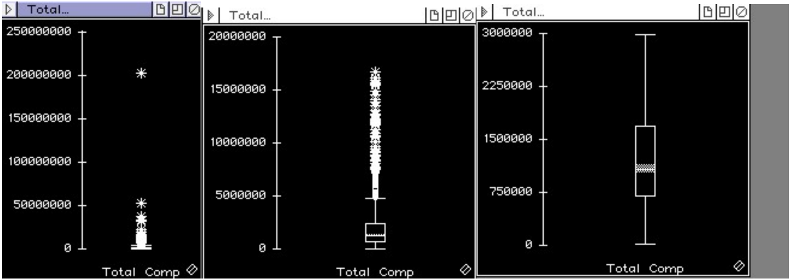
$n = 790$ $n = 781$ $n = 638$



Comparing 84th Percentiles

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

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



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 $\mu + \sigma$.

	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 84th Percentiles

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

	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

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 $\mu + \sigma$.

	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

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 $n = 638$.

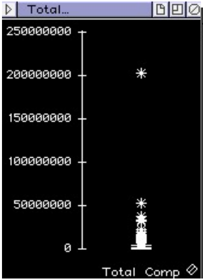
Comparing IQR's

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

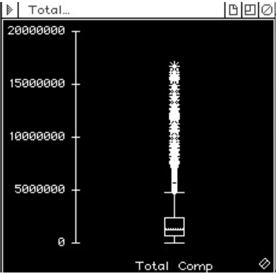
Comparing IQR's

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

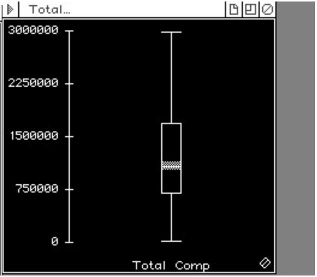
$n = 790$



$n = 781$



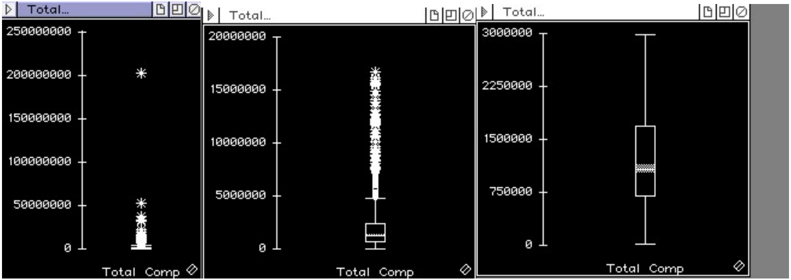
$n = 638$



Comparing IQR's

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

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



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 1.35σ.

	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 1.35σ.

	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 1.35σ.

	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.

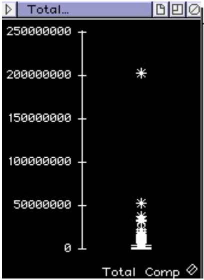
Agreement of $1.35s$ with IQR is good when $n = 638$.

Comparing 97.7%iles

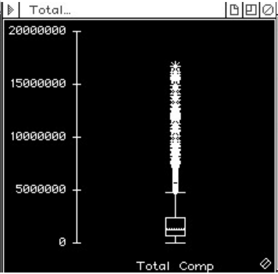
For a normal distribution $N(\mu, \sigma)$, this would be at $\mu + 2\sigma$.

Comparing 97.7%iles

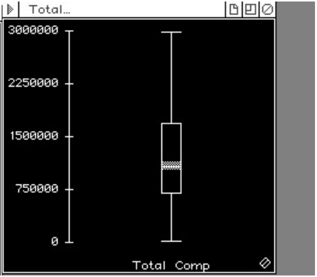
$n = 790$



$n = 781$



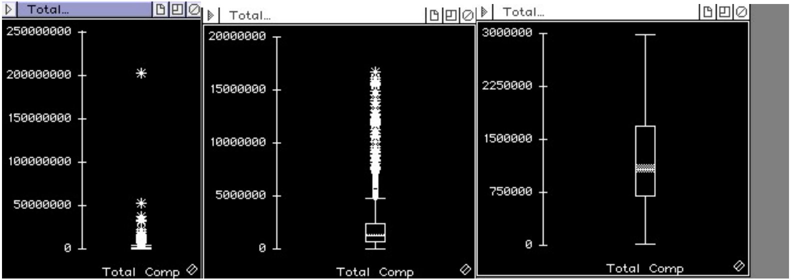
$n = 638$



Comparing 97.7%iles

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

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



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 $\mu + 2\sigma$.

	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 97.7%iles

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

	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

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 $\mu + 2\sigma$.

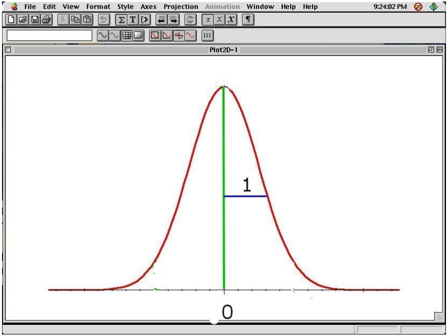
	n	s	$\bar{x}$	**97.7%ile**	$\bar{x} + 2s$
All	790	8.32M	2.82M	14.85M	19.46M
Without top 9	781	2.724M	2.24	12.33M	7.68M
Bottom 638	638	.689M	1.25M	2.79M	2.63M

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

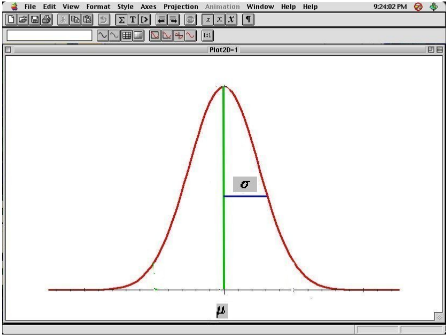
N(0,1)



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

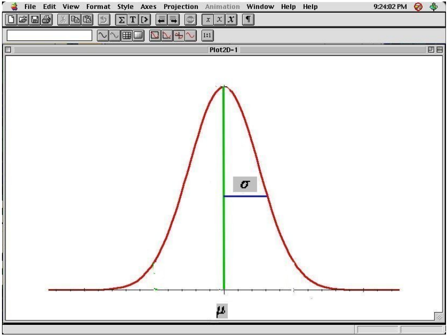
Normal Distribution Formula

$N(\mu, \sigma)$



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

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 $\sigma = \sqrt{.5m}$.

Why Normal Out of Binomial?

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

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

Why Normal Out of Binomial?

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

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

Set

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

Why Normal Out of Binomial?

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

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

Set

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

The z-score of $m + k$ is $\frac{k\sqrt{2}}{\sqrt{m}}$.

Why Normal Out of Binomial?

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

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

Set

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

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.

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

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

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

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

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

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

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

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

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

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.

C

C