

Statistics ...

Statistics is the study of how best to

1. collect **data**
2. summarize/describe **data**
3. draw conclusions/inferences from **data**

all in a framework that explicitly recognizes the reality and omnipresence of **variability**.

"Probability" is the mathematical description of "chance" and is a tool for the above that must be part of the course ... but it is *not* statistics.

Statistics is the backbone of intelligent business decisions.

Video: Against All Odds, Unit #1

Basic Descriptive Statistics

Real data are variable, so it is important to describe/measure that variability and the pattern of variation.

We'll consider 2 kinds of simple methods:

- graphical/pictorial methods
- numerical methods

Simple Statistical Graphics

Histograms/Bar Charts *Example* 1950 values from Frequency Table 1.4,
page 25 of MMD&S

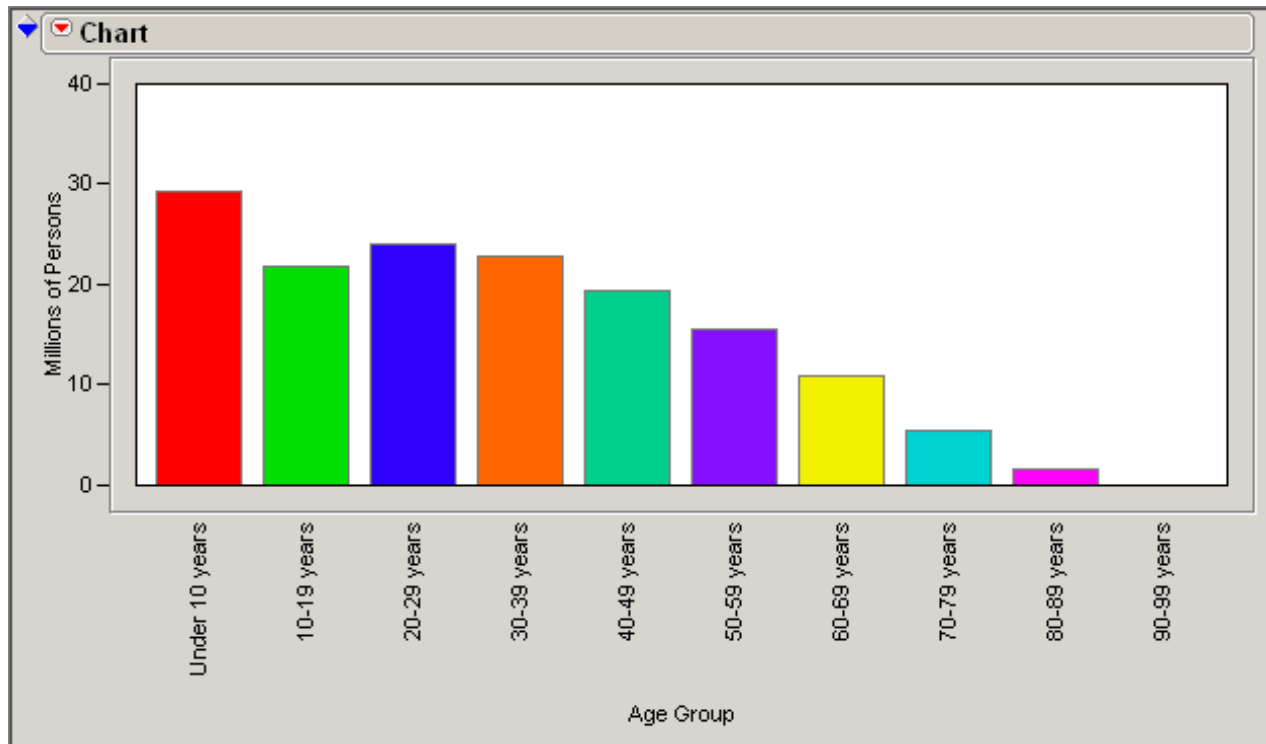


Figure 1: Histogram for MMD&S Table 1.4

On a well-made histogram

area of a bar \longleftrightarrow fraction of data values represented

"Shapes" of histograms often suggest mechanisms/causes/sources, trade-offs possible, interventions, etc. (and can therefore inform and guide good business decisions).

Example "Bimodal" histogram

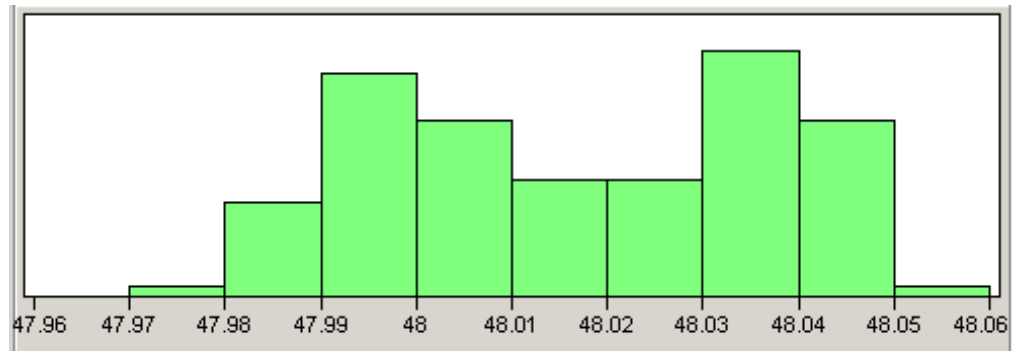


Figure 2: A Frequency Histogram of Cylinder Diameters of Parts Made on Two Different Machines

Example "Left-truncated" histogram

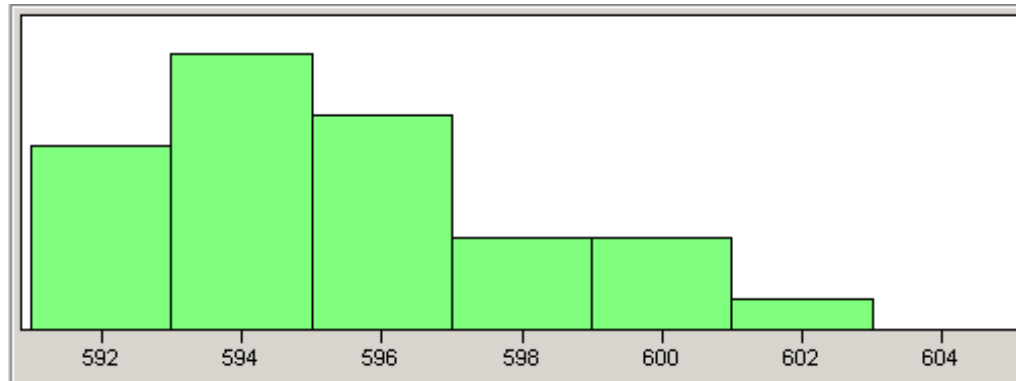


Figure 3: Histogram of Fill Levels (in ml) for 20 oz (591 ml) Bottles of Cola

Example "Right-skewed" histogram

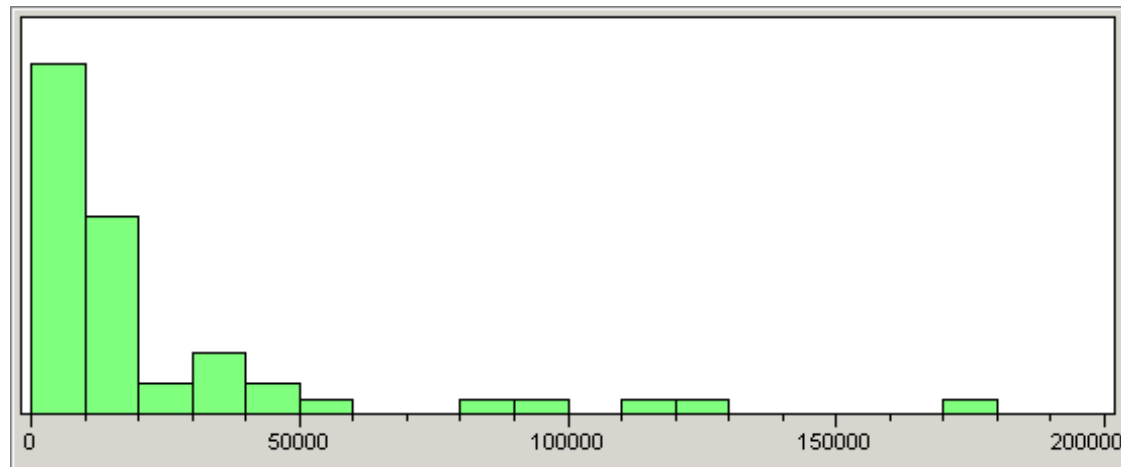


Figure 4: Histogram of AGI's from a Sample of Federal Form 1040s

Example Advantage of reduced variation/improved precision (probably at a cost)

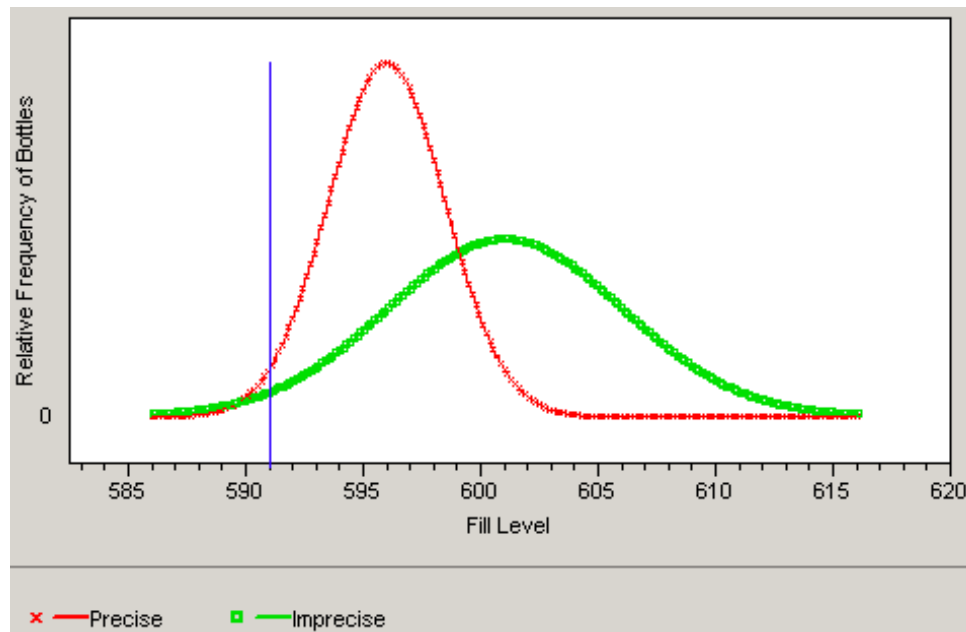


Figure 5: Two Possible Distributions of Fill Levels in "591 ml" Pop Bottles

Example Risk trade-off

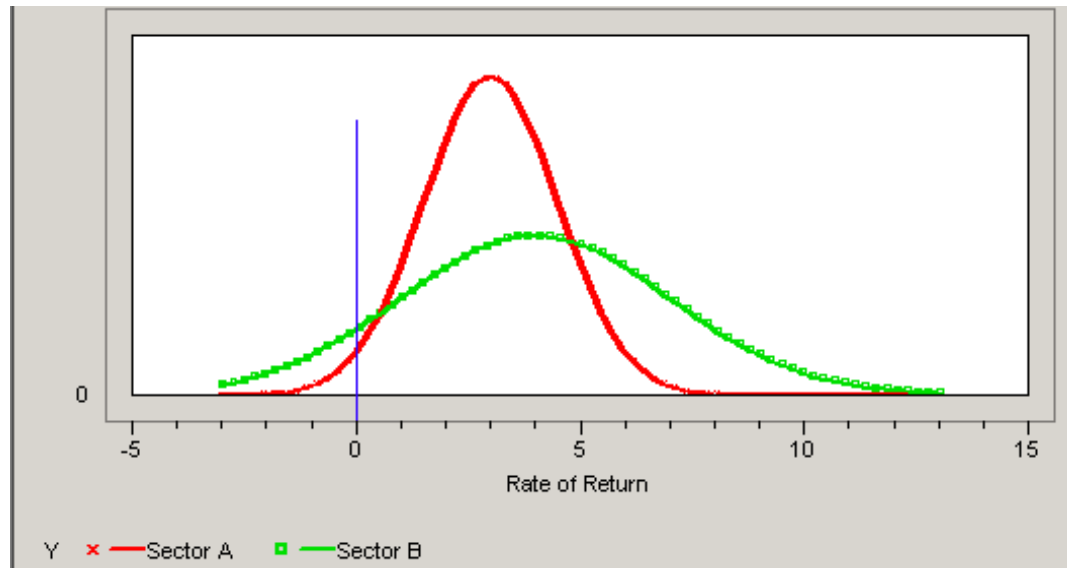


Figure 6: Rates of Return on Stocks in Two Different Sectors

Stem Plots/Stem-and-Leaf Plots These give much the same "shape" information as histograms, but do it without loss of information (retaining all individual values).

Example Stem Plot for the data of Table 1.1 of MMD&S

1 |
2 |
3 |
4 |
5 |
6 |
7 |
8 |

(See the pair of histograms on page 12 and the pair of stem plots on page 19 of MMD&S.)

Plotting Against Time Some data sets are essentially "snapshots" of a single time or condition. Others are meant to represent what happens over time (are "movies"), and in those cases we're wise to plot against time. The hope is to see trends or patterns (for the purposes of understanding temporal mechanisms and making forecasts, to the end of making wise business actions).

Example Exercise 1.12, page 22 MMD&S and JMP "overlay plot"

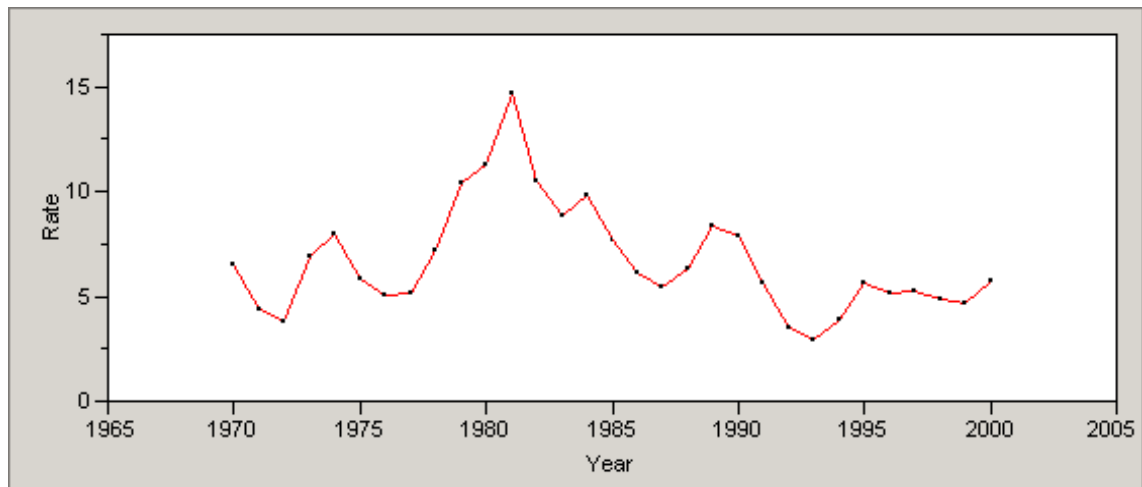


Figure 7: Plot of T-Bill Rates 1970 to 2000 (Exercise 1.12 MMD&S)

Simple Numerical Data Summaries

Measures of Center/Location/Average The (sample) (arithmetic) **mean** is

$$\bar{x} = \frac{1}{n} \sum x_i$$

The (sample) **median** is

$$M = \frac{n + 1}{2} \text{th ordered data value}$$

These are different measures of "center" with different properties. To be an intelligent consumer of statistical information, one must be aware of these.

Example Simple (fake) data set 1,1,2,2,9

$$\bar{x} = \frac{1}{5} (1 + 1 + 2 + 2 + 9) = \frac{15}{5} = 3$$

$$M = 3rd \left(\frac{5+1}{2} \text{th} \right) \text{ ordered data value} = 2$$

The mean is sensitive to a "few" extreme data points. The median is not.

Example Modified simple (fake) data set 1,1,2,2,1009 has

$$\bar{x} = 203$$

$$M = 2$$

\bar{x} is a "balance point" of a data set, while M "cuts the data set in half."

A right-skewed data set typically has a mean that is larger than its median.

Example Earlier sample of AGIs from Federal 1040s

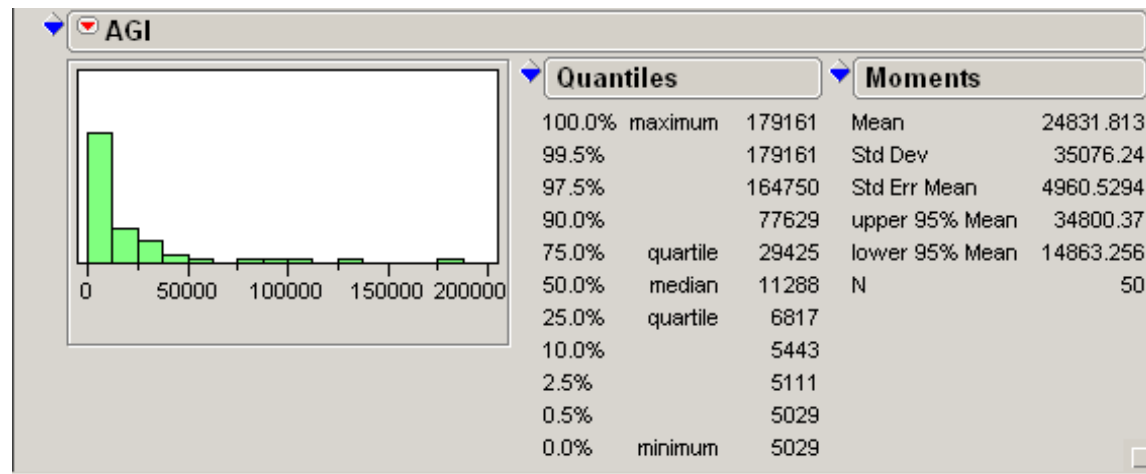


Figure 8: JMP Report Showing Mean and Median for a Strongly Right-skewed Distribution

5-Number Summaries and Boxplots A "5-number summary" of a data set consists of

minimum, Q_1 , M , Q_3 , maximum

Q_1 and Q_3 are the "first and third quartiles" ... roughly the 25% and 75% points of the data distribution. Various conventions are possible as to exactly how to get these from a small data. Two similar ones (not the only ones possible) are:

- Vardeman modification of MMD&S (the same as MMD&S for even n)

Q_1 = "1st quartile"

= median of the data values whose position in the ordered list is at or below that of M

Q_2 = "3rd quartile"

= median of the data values whose position in the ordered list is at or above that of M

- MMD&S convention, see page 26 of the text

Q_1 = "1st quartile"

= median of the data values whose position in the ordered list is below that of M

Q_2 = "3rd quartile"

= median of the data values whose position in the ordered list is above that of M

Example (small fake data set 1,1,2,2,9)

Vardeman version of 5-number summary is

$$\begin{aligned} \textit{minimum} &= 1 \\ Q_1 &= 1 \\ M &= 2 \\ Q_3 &= 2 \\ \textit{maximum} &= 9 \end{aligned}$$

MMD&S version of the 5-number summary is

$$\begin{aligned} \textit{minimum} &= 1 \\ Q_1 &= \frac{1 + 1}{2} = 1 \\ M &= 2 \\ Q_3 &= \frac{2 + 9}{2} = 5.5 \\ \textit{maximum} &= 9 \end{aligned}$$

The differences between various conventions for quartiles generally becomes negligible as n gets large. (JMP uses neither convention above, but yet something else.)

Boxplots display the 5-number summary "to scale" in graphical fashion.

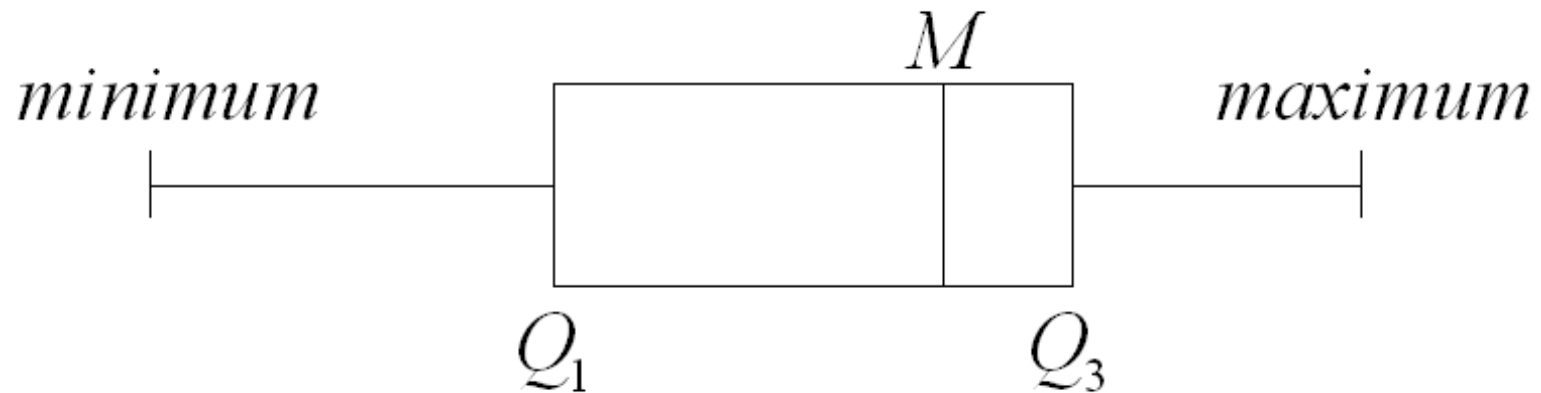


Figure 9: Schematic of MMD&S Boxplot

This kind of plot gives an indication of location, an indication of spread, and some information on shape/symmetry of a data set. It does so in a way that allows many to be put side by side on a single display for comparison purposes.

Example MMD&S style boxplot of AGI data.

$$\textit{minimum} = 5,029$$

$$Q_1 = 6,817$$

$$M = 11,288$$

$$Q_3 = 29,425$$

$$\textit{maximum} = 179,161$$

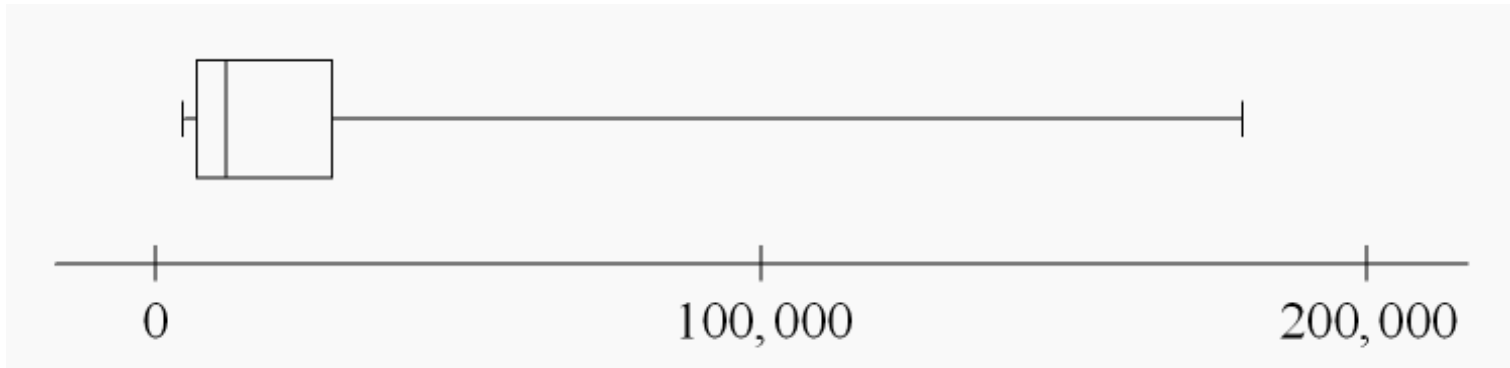


Figure 10: Boxplot of AGIs from a Sample of Federal Forms 1040

Examples See Figure 1.11, page 39 of MMD&S, plot of Wal-Mart data on page 75 of BPS

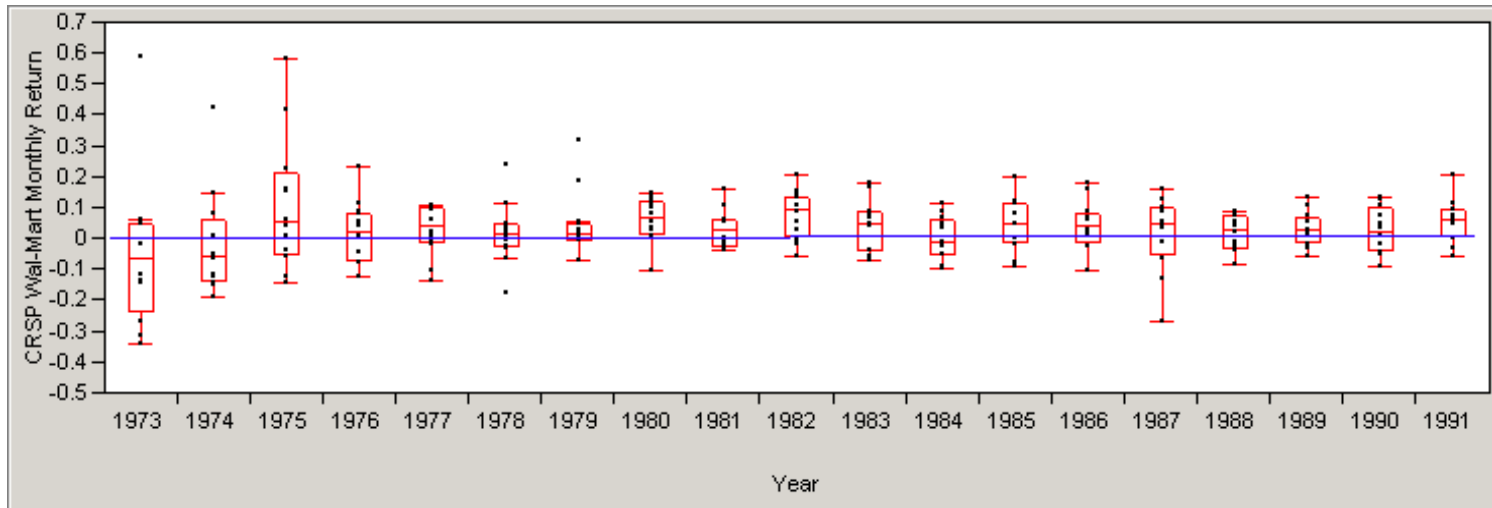


Figure 11: CRSP Wal-Mart Monthly Return (JMP Boxplots ... Made With a Different Convention than MMD&S)

Measuring Spread/Variability in a Data Set The most commonly used measure of spread/variability for a data set is the (sample) standard deviation. This is the square root of a kind of "average squared distance from data points to the 'center' of the data set." (The square root is required to "undo the squaring" and get a measure that has the same units as the original data values.) The sample standard deviation is

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The square of the sample standard deviation is called the (sample) variance and is written

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

This measure has units that are the square of the original units of the data values.

Example Simple (fake) data set 1,1,2,2,9

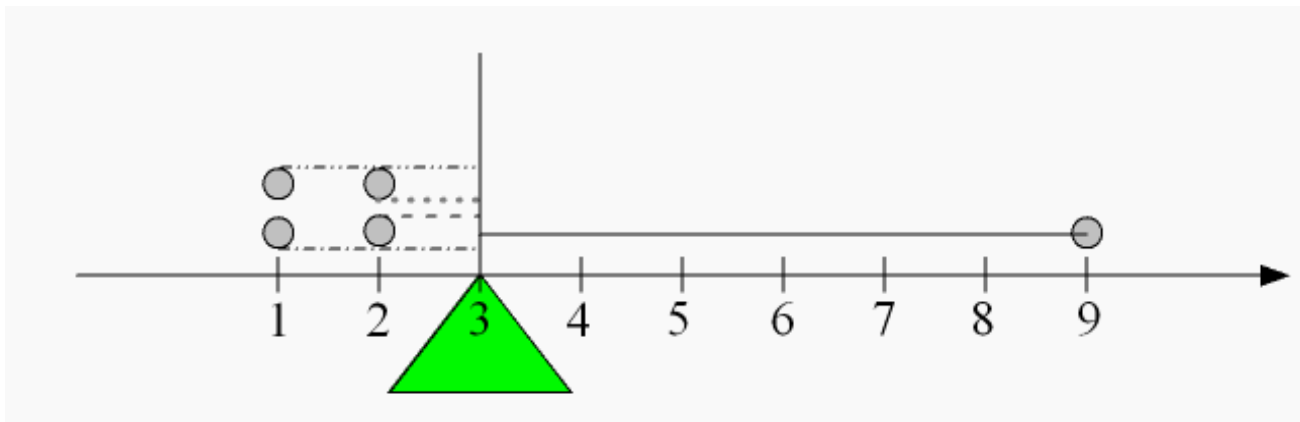


Figure 12: Computing and Motivating s^2 for the Fake Data Set

$$\begin{aligned}s^2 &= \frac{1}{4} \left((9 - 3)^2 + (2 - 3)^2 + (2 - 3)^2 + (1 - 3)^2 + (1 - 3)^2 \right) \\ &= \frac{46}{6}\end{aligned}$$

so that

$$s = \sqrt{\frac{46}{6}} = 3.39 \text{ (in the original units)}$$

NOTE: s^2 and s can be 0 ONLY when all data values are exactly the same ... they can NEVER be negative

Note also for future reference that the definition of s^2 implies that

$$\sum (x_i - \bar{x})^2 = (n - 1) s^2$$

(that is, since a good business calculator will compute s^2 automatically, by multiplying the sample variance by $n - 1$, one can easily compute $\sum (x_i - \bar{x})^2$)

Exercise For the small fake data sets below, find means and standard deviations "by hand."

Data set #2: 2,2,4,4,18

Data set #3: 51,51,52,52,59

The general story hinted at by the exercise is that for

x 's with mean \bar{x} and standard deviation s_x

if we make up values

$$y = a + bx$$

then these have

$$\text{mean } \bar{y} = a + b\bar{x}$$

and

$$\text{standard deviation } s_y = |b| s_x$$

BTW, smart statistical programmers make use of this ... (the people that did the programming of the EXCEL statistical add-ins were NOT smart). The data set

10000000000.1 10000000000.2 10000000000.3

has standard deviation that is .1 times the standard deviation of the data set

1 2 3

A bad programmer will instead get 0 because of round-off.

"Normal" Distribution Models (Bell-Shaped Idealized Histograms)

These are idealized/theoretical approximations to "bell-shaped" data sets/distributions. They are convenient because they are completely described by two "parameters" (two summary numbers). Given a measure of center, μ , (that is a theoretical

"mean") and a measure of spread, σ , (that is a theoretical "standard deviation") fractions of the distribution between any two values can be computed. (So for a given mean and standard deviation, one doesn't need to carry around an entire data set or relative frequency distribution ... approximate fractions can be computed "on the spot" based only on μ and σ .)

A crude summary of what a normal distribution says is provided by what MMD&S call the "68%-95%-99.7% rule." This says that

1. 68% of a normal distribution is within σ of μ
2. 95% of a normal distribution is within 2σ of μ
3. 99.7% of a normal distribution is within 3σ of μ

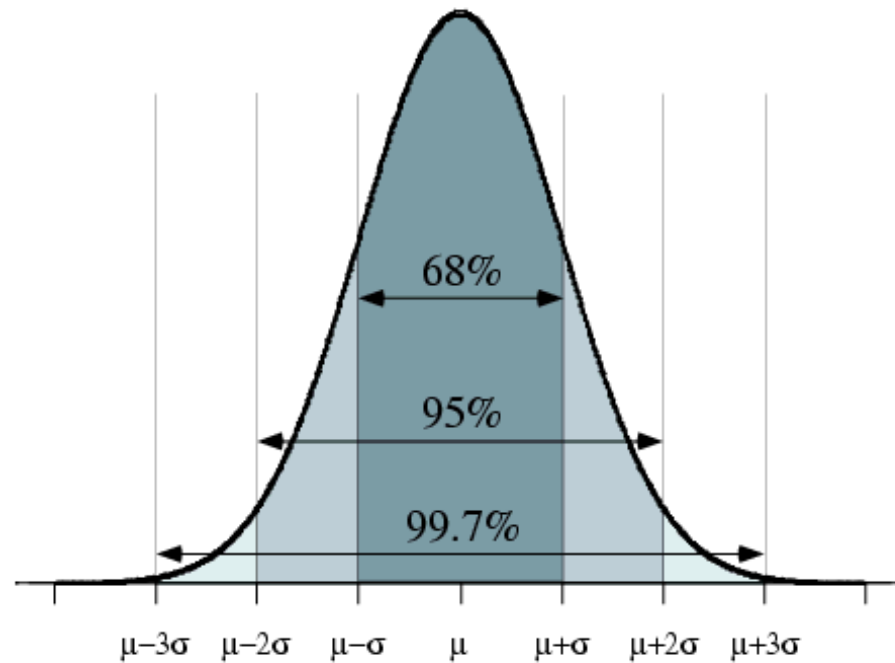


Figure 13: A "Normal" Idealized Histogram ... A "Model" for "Bell-Shaped" Data Distributions

Example Problem 1.64, page 59 MMD&S

WAI scores might be modeled as normal with $\mu = 110$ and $\sigma = 25$

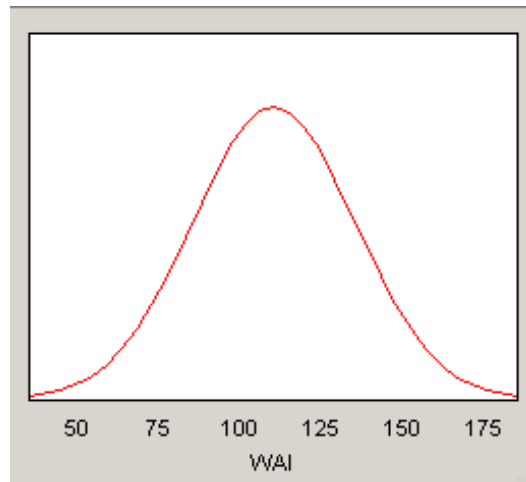


Figure 14: Normal Idealized Histogram for WAI Scores

- 68% of WAI scores are between 85 and 135

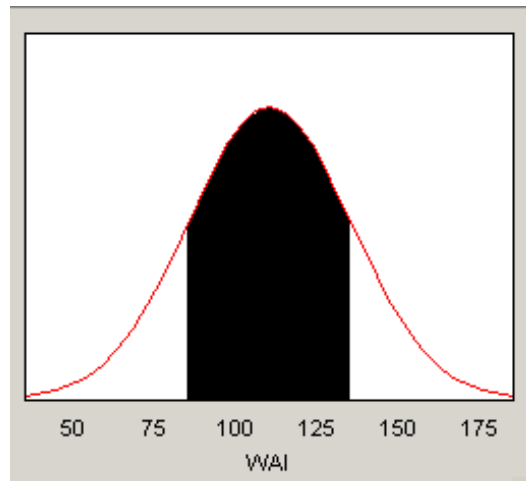


Figure 15: WAI Scores from 85 to 135

- the "middle 95%" of WAI scores are between 60 and 160

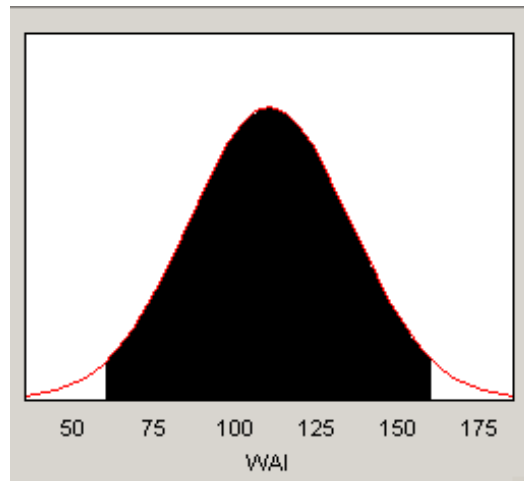


Figure 16: WAI Scores Between 60 and 160

- about 16% of WAI scores are above 135

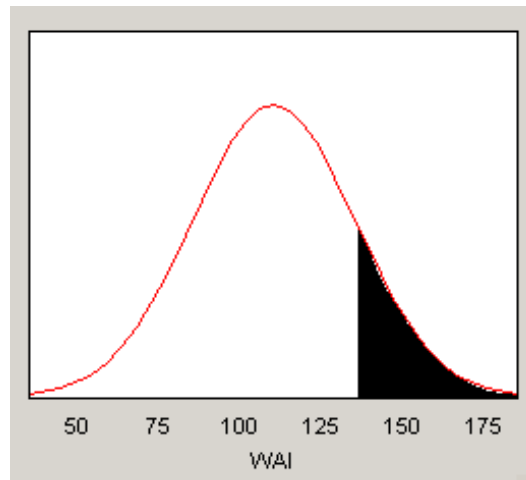


Figure 17: WAI Scores Above 135

What about more detail than the 68%-95%-99.7% rule provides? (For example, what fraction of WAI scores are above 140?) Two additional items are needed in order to fully detail a normal distribution

1. the notion of a " z -score" (that essentially converts to units of "standard deviations above the mean")

$$z = \frac{x - \mu}{\sigma}$$

2. a complete table of fractions of z -scores (fractions of a "standard normal" distribution ... one with mean 0 and standard deviation 1) to the left of any given z . See the table inside the front cover of MMD&S for this.

Example WAI scores again, $\mu = 110$ and $\sigma = 25$

$x = 135$ has z -score

$$z = \frac{135 - 110}{25} = 1$$

$x = 160$ has z -score

$$z = \frac{160 - 110}{25} = 2$$

$x = 85$ has z -score

$$z = \frac{85 - 110}{25} = \frac{-25}{25} = -1$$

NOTE: the sign is important! It tells you on which side of the mean the x of interest lies. It can not be dropped without error.

$x = 140$ has z -score

$$z = \frac{140 - 110}{25} = \frac{30}{25} = 1.2$$

A direct table look-up for $z = 1.20$ gives the number .8849. 88.49% of values from a normal distribution have z -scores less than 1.20 (a fraction .8849 of the area under a normal idealized histogram is to the left of the value with z -score 1.20 ... recall the remark earlier that on a histogram *area* \longleftrightarrow *fraction of a data set*).

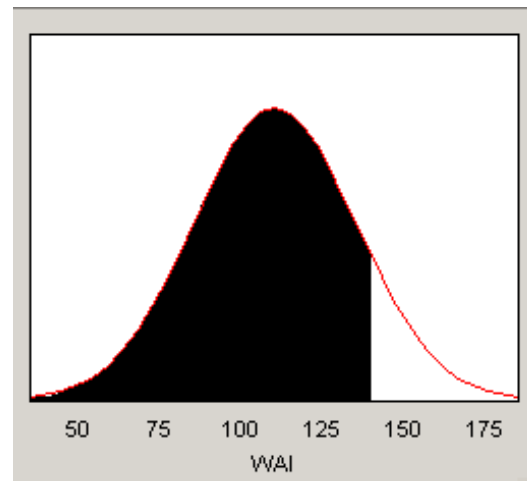


Figure 18: WAI Scores Below 140 (WAI Values With z -scores Below 1.2)

So using the normal model to describe WAI scores, we'd say

a fraction $.8849$ of scores are less than 140

and

a fraction $1 - .8849 = .1151$ of scores are more than 140

Example WAI scores again, $\mu = 110$ and $\sigma = 25$ What fraction of WAI scores are between 90 and 140?

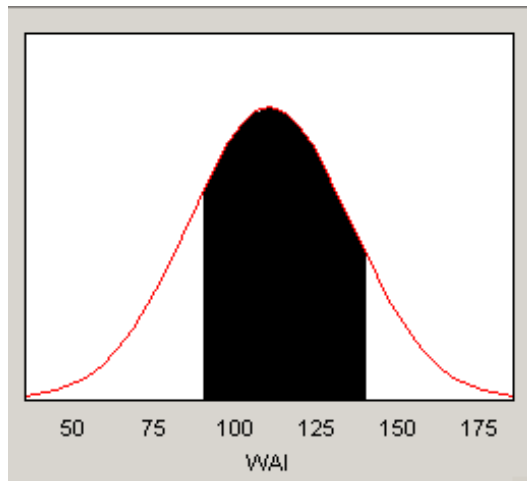


Figure 19: WAI Scores Between 90 and 140

$$z_1 = \frac{140 - 110}{25} = \frac{30}{25} = 1.2$$
$$z_2 = \frac{90 - 110}{25} = \frac{-20}{25} = -.8$$

Then two look-ups in the standard normal table give

$$area_1 = .8849 \text{ corresponding to } z_1 = 1.20$$
$$area_2 = .2119 \text{ corresponding to } z_2 = -.80$$

The fraction of the distribution between 90 and 140 is the difference

$$area_1 - area_2 = .8849 - .2119 = .6730$$

That is, according to the normal model, about 67% of WAI scores are between 90 and 140.

Example WAI scores again, $\mu = 110$ and $\sigma = 25$ Where are the lowest 25% of scores?

This is a different kind of normal problem ... we're given a fraction of the distribution (and therefore a z -score), the mean and the standard deviation and must come up with an x . To do this we use the table "backwards," running from body to margin.

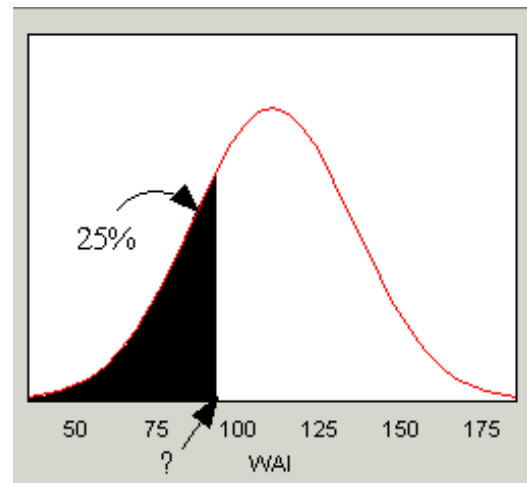


Figure 20: Lowest 25% of WAI Scores

We find that a left-tail area for a normal distribution corresponds to a z -score of about $-.675$. So we set

$$-.675 = z = \frac{x - 110}{25}$$

and try to solve for x . This gives

$$25(-.675) = x - 110$$

$$x = 110 + (-.675)25 = 93.1$$

We have illustrated 2 types of normal problems. In the first, we input x, μ, σ and computed z and thus a fraction of the distribution. In the second we input μ, σ , an fraction (and thus a z) and computed x . There are exactly 4 different types of normal problems ... starting from the basic

$$z = \frac{x - \mu}{\sigma}$$

one must be given 3 of the 4 entries and can then solve for the 4th. (So the 2 versions of the problem not illustrated here are ones where μ or σ can be solved for in terms of the other entries.)