Looking at Data—Distributions

IPS Chapter 1

- 1.1: Displaying distributions with graphs
- 1.2: Describing distributions with numbers
- 1.3: Density Curves and Normal Distributions

Looking at Data—Distributions

1.1 Displaying Distributions with Graphs

Objectives

1.1 Displaying distributions with graphs

- Variables
- Types of variables
- Graphs for categorical variables
  - Bar graphs
  - Pie charts
- Graphs for quantitative variables
  - Histograms
  - Stemplots
  - Stemplots versus histograms
- Interpreting histograms
- Time plots

Variables

In a study, we collect information—data—from **cases.** Cases can be individuals, companies, animals, plants, or any object of interest.

A **label** is a special variable used in some data sets to distinguish the different cases.

A **variable** is any characteristic of an case. A variable varies among cases.

Example: age, height, blood pressure, ethnicity, leaf length, first language

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

Two types of variables

- Variables can be either **quantitative…**
  - Something that takes numerical values for which arithmetic operations, such as adding and averaging, make sense.
  - Example: How tall you are, your age, your blood cholesterol level, the number of credit cards you own.
- … or **categorical.**
  - Something that falls into one of several categories. What can be counted is the count or proportion of cases in each category.
  - Example: Your blood type (A, B, AB, O), your hair color, your ethnicity, whether you paid income tax last tax year or not.

How do you know if a variable is categorical or quantitative?

Ask:

- What are the *n* cases/units in the sample (of size "*n*")?
- What is being recorded about those *n* cases/units?
- Is that a number (→ quantitative) or a statement (→ categorical)?

Label

Categorical
Each individual is assigned to one of several categories.

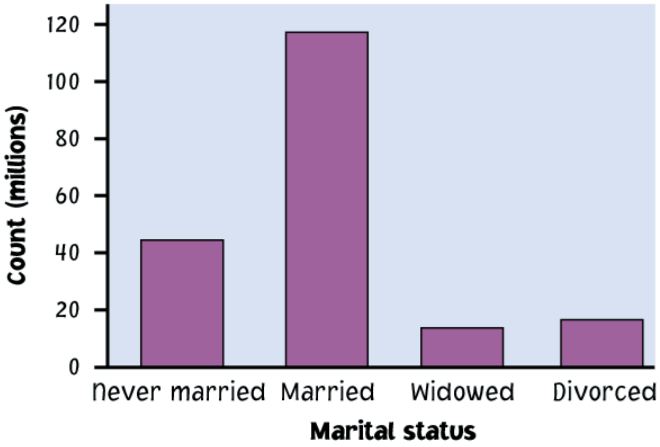
Quantitative
Each individual is attributed a numerical value.

Individuals in sample	DIAGNOSIS	AGE AT DEATH
Patient A	Heart disease	56
Patient B	Stroke	70
Patient C	Stroke	75
Patient D	Lung cancer	60
Patient E	Heart disease	80
Patient F	Accident	73
Patient G	Diabetes	69

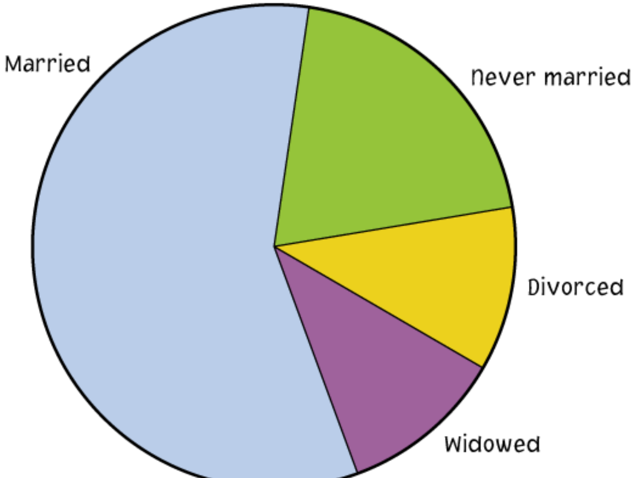
Ways to chart categorical data

Because the variable is categorical, the data in the graph can be ordered any way we want (alphabetical, by increasing value, by year, by personal preference, etc.)

- **Bar graphs**
  Each category is represented by a bar.



- **Pie charts**
  The slices must represent the parts of one whole.



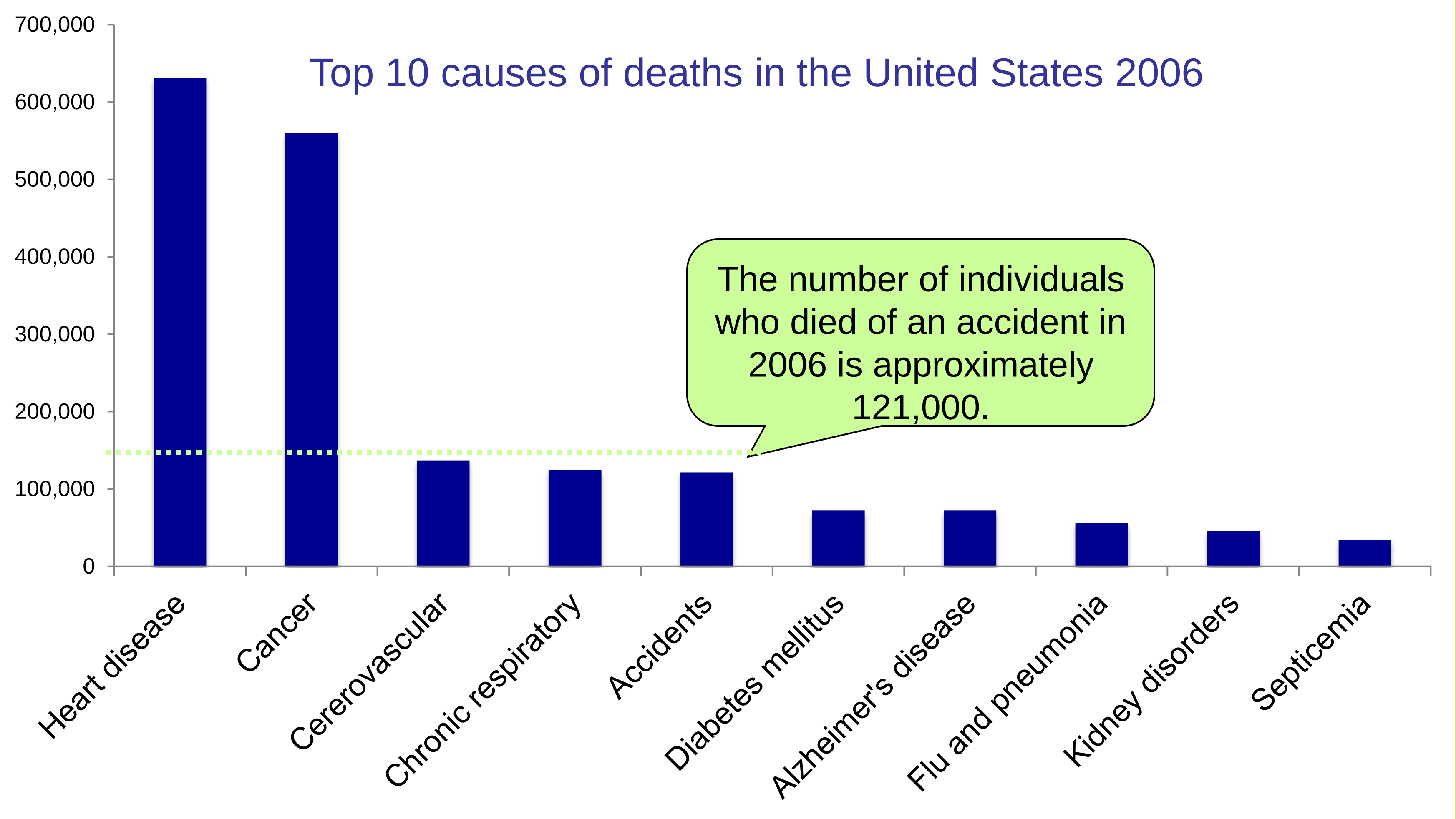
Example: Top 10 causes of death in the United States 2006

Rank	Causes of death	Counts	% of top 10s	% of total deaths
1	Heart disease	631,636	34%	*26%*
2	Cancer	559,888	30%	*23%*
3	Cerebrovascular	137,119	7%	*6%*
4	Chronic respiratory	124,583	7%	*5%*
5	Accidents	121,599	7%	*5%*
6	Diabetes mellitus	72,449	4%	*3%*
7	Alzheimer's disease	72,432	4%	*3%*
8	Flu and pneumonia	56,326	3%	*2%*
9	Kidney disorders	45,344	2%	*2%*
10	Septicemia	34,234	2%	*1%*
	All other causes	*570,654*		*24%*

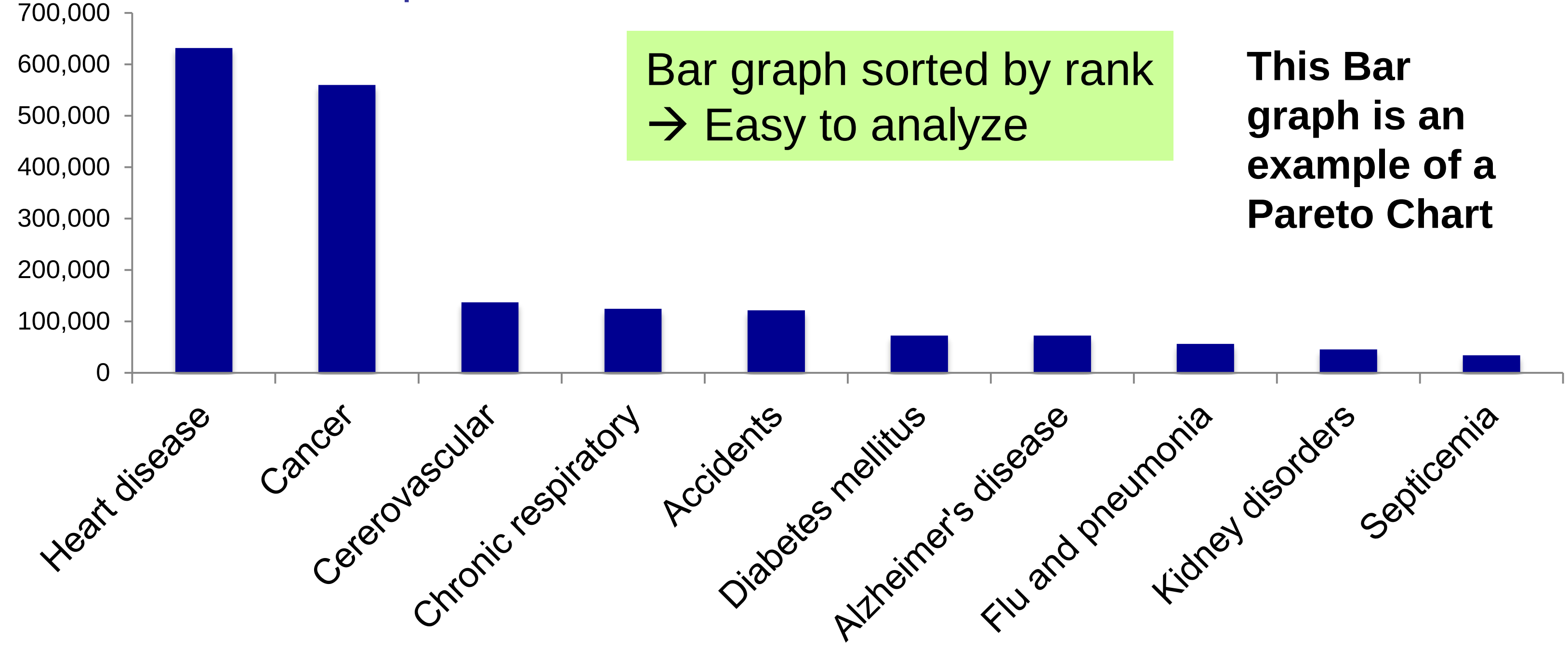
For each individual who died in the United States in 2006, we record what was the cause of death. The table above is a summary of that information.

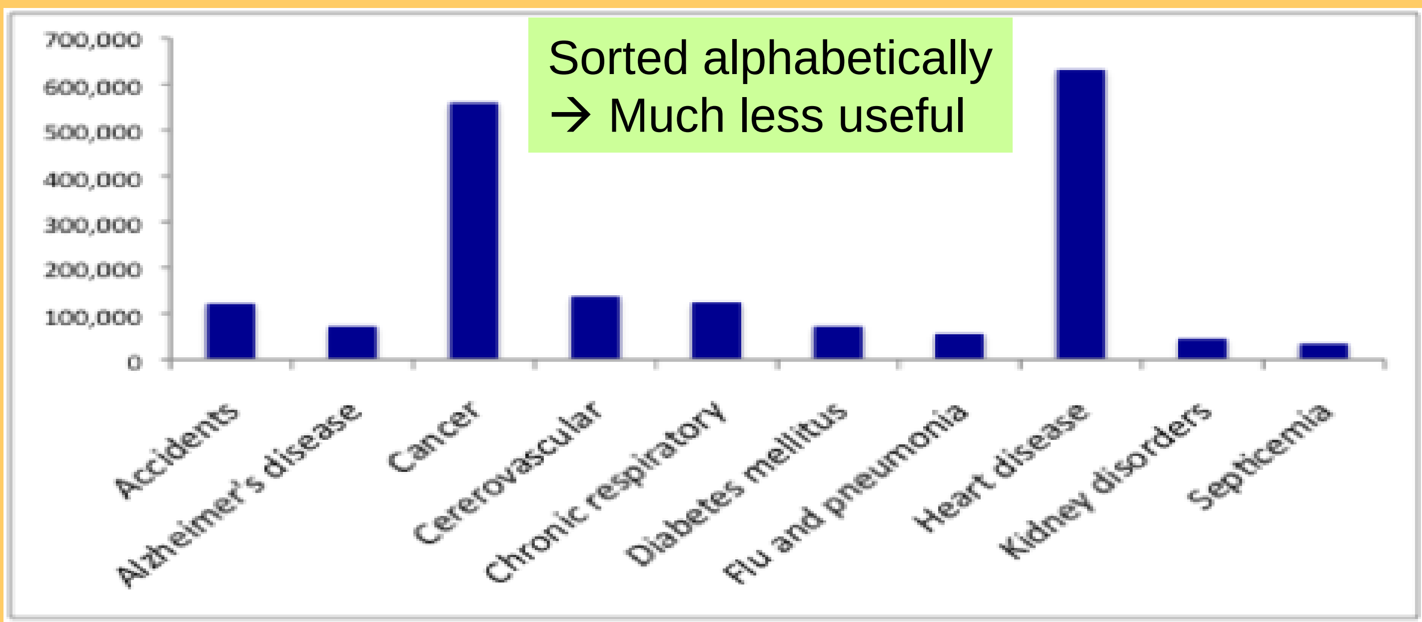
Bar graphs

Each category is represented by one bar. The bar's height shows the count (or sometimes the percentage) for that particular category.



Top 10 causes of deaths in the United States 2006

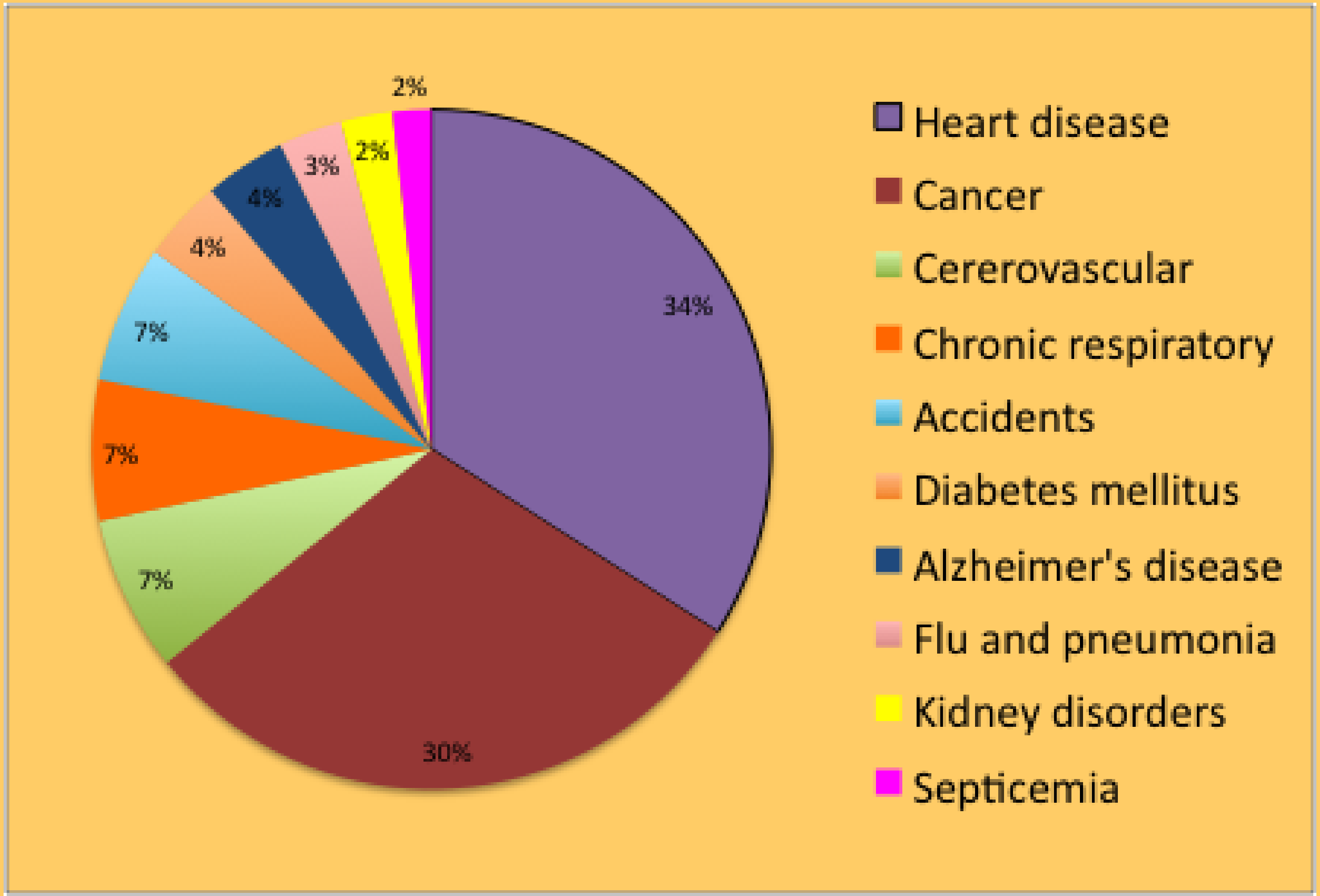


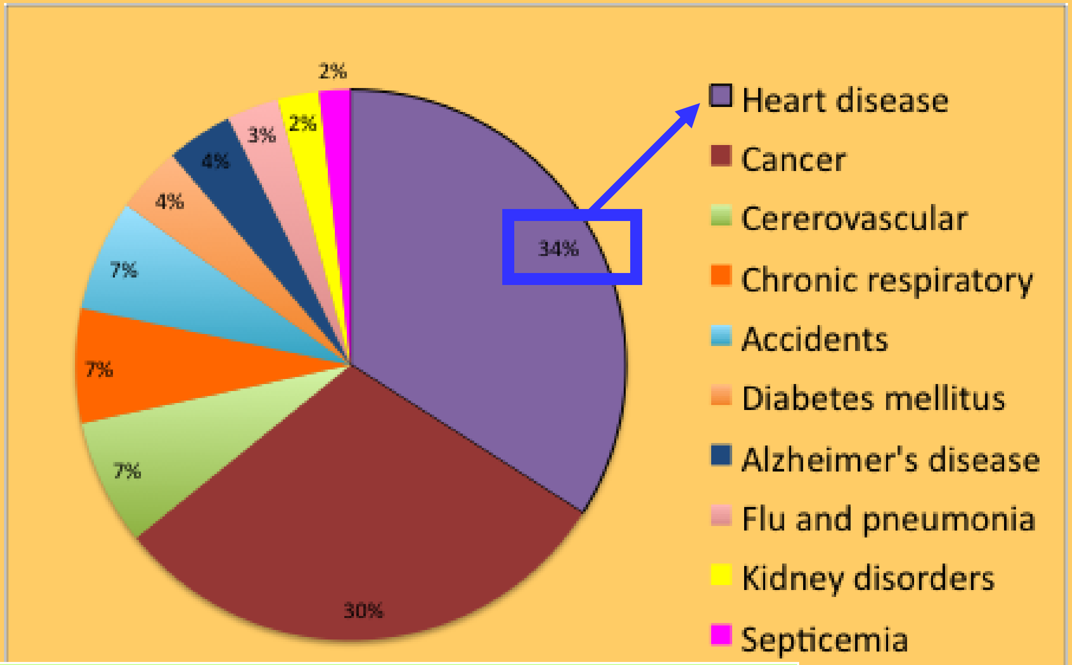


Pie charts

Each slice represents a piece of one whole. The size of a slice depends on what percent of the whole this category represents.

Percent of people dying from
top 10 causes of death in the United States in 2006

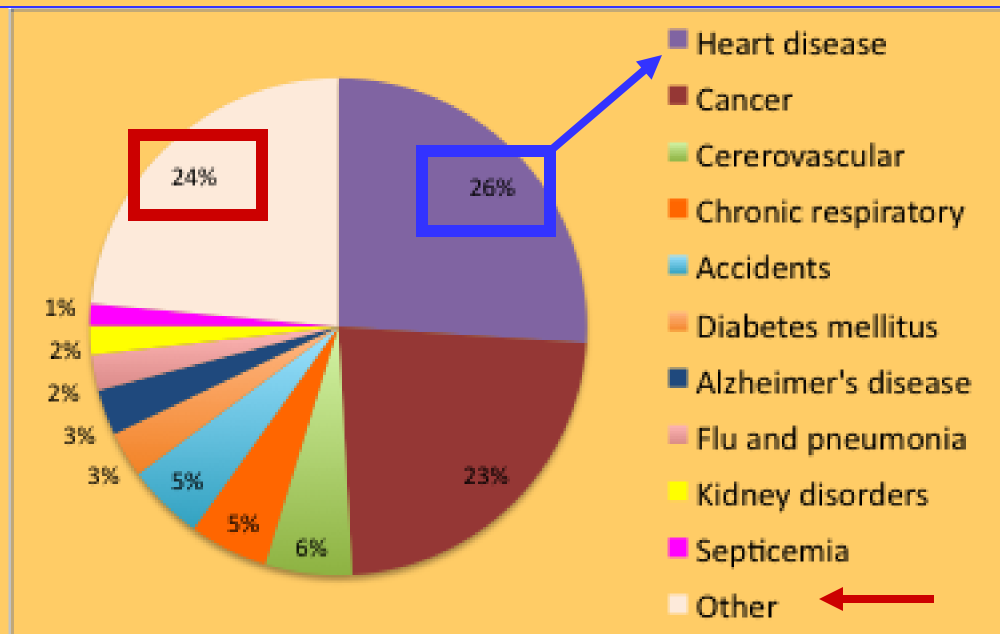




Percent of deaths from <u>top 10</u> causes

Make sure your labels match the data.

Make sure all percents add up to 100.



Percent of deaths from <u>all</u> causes

Child poverty before and after government intervention—UNICEF, 2005

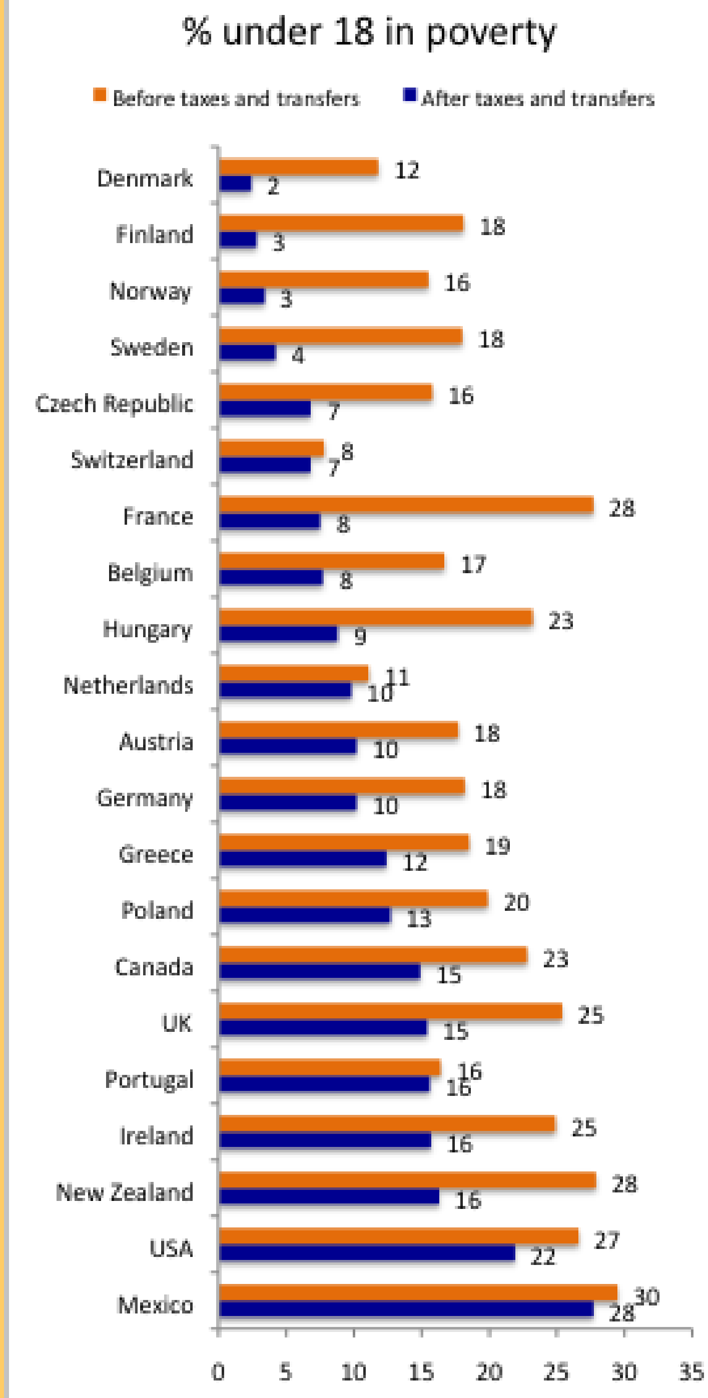
What does this chart tell you?

- The United States and Mexico have the highest rate of child poverty among OECD (Organization for Economic Cooperation and Development) nations (22% and 28% of under 18).
- Their governments do the least—through taxes and subsidies—to remedy the problem (size of orange bars and percent difference between orange/blue bars).

Identify the Pareto Chart in this graph.

Could you transform this bar graph to fit in 1 pie chart? In two pie charts? Why?

The poverty line is defined as 50% of national median income.



Ways to chart quantitative data

- Stemplots

  Also called a stem-and-leaf plot. Each observation is represented by a **stem**, consisting of all digits except the final one, which is the **leaf**.

- Histograms

  A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class.

- Line graphs: time plots

  A **time plot** of a variable plots each observation against the time at which it was measured.

Stem plots

How to make a **stemplot:**

- Separate each observation into a **stem,** consisting of all but the final (rightmost) digit, and a **leaf,** which is that remaining final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
- Write the stems in a vertical column with the smallest value at the top, and draw a vertical line at the right of this column.
- Write each leaf in the row to the right of its stem, in increasing order out from the stem.

STEM	LEAVES
0	9 9
1	
2	2
3	2 3 9 9
4	2 9
5	2 8
6	
7	0

Percent of Hispanic residents in each of the 50 states

State	Percent
Alabama	1.5
Alaska	4.1
Arizona	25.3
Arkansas	2.8
California	32.4
Colorado	17.1
Connecticut	9.4
Delaware	4.8
Florida	16.8
Georgia	5.3
Hawaii	7.2
Idaho	7.9
Illinois	10.7
Indiana	3.5
Iowa	2.8
Kansas	7
Kentucky	1.5
Louisiana	2.4
Maine	0.7
Maryland	4.3
Massachusett	6.8
Michigan	3.3
Minnesota	2.9
Mississippi	1.3
Missouri	2.1
Montana	2
Nebraska	5.5
Nevada	19.7
NewHampshi	1.7
NewJersey	13.3
NewMexico	42.1
NewYork	15.1
NorthCarolina	4.7
NorthDakota	1.2
Ohio	1.9
Oklahoma	5.2
Oregon	8
Pennsylvania	3.2
RhodeIsland	8.7
SouthCarolina	2.4
SouthDakota	1.4
Tennessee	2
Texas	32
Utah	9
Vermont	0.9
Virginia	4.7
Washington	7.2
WestVirginia	0.7
Wisconsin	3.6
Wyoming	6.4

Step 1:

Sort the data

State	Percent
Maine	0.7
WestVirginia	0.7
Vermont	0.9
NorthDakota	1.2
Mississippi	1.3
SouthDakota	1.4
Alabama	1.5
Kentucky	1.5
NewHampshi	1.7
Ohio	1.9
Montana	2
Tennessee	2
Missouri	2.1
Louisiana	2.4
SouthCarolina	2.4
Arkansas	2.8
Iowa	2.8
Minnesota	2.9
Pennsylvania	3.2
Michigan	3.3
Indiana	3.5
Wisconsin	3.6
Alaska	4.1
Maryland	4.3
NorthCarolina	4.7
Virginia	4.7
Delaware	4.8
Oklahoma	5.2
Georgia	5.3
Nebraska	5.5
Wyoming	6.4
Massachusett	6.8
Kansas	7
Hawaii	7.2
Washington	7.2
Idaho	7.9
Oregon	8
RhodeIsland	8.7
Utah	9
Connecticut	9.4
Illinois	10.7
NewJersey	13.3
NewYork	15.1
Florida	16.8
Colorado	17.1
Nevada	19.7
Arizona	25.3
Texas	32
California	32.4
NewMexico	42.1

Step 2:

Assign the values to stems and leaves

0	779
1	2345579
2	00144889
3	2356
4	13778
5	235
6	48
7	0229
8	07
9	04
10	7
11	
12	
13	3
14	
15	1
16	8
17	1
18	
19	7
21	
22	
23	
24	
25	3

These entries are 6.4% and 6.8%

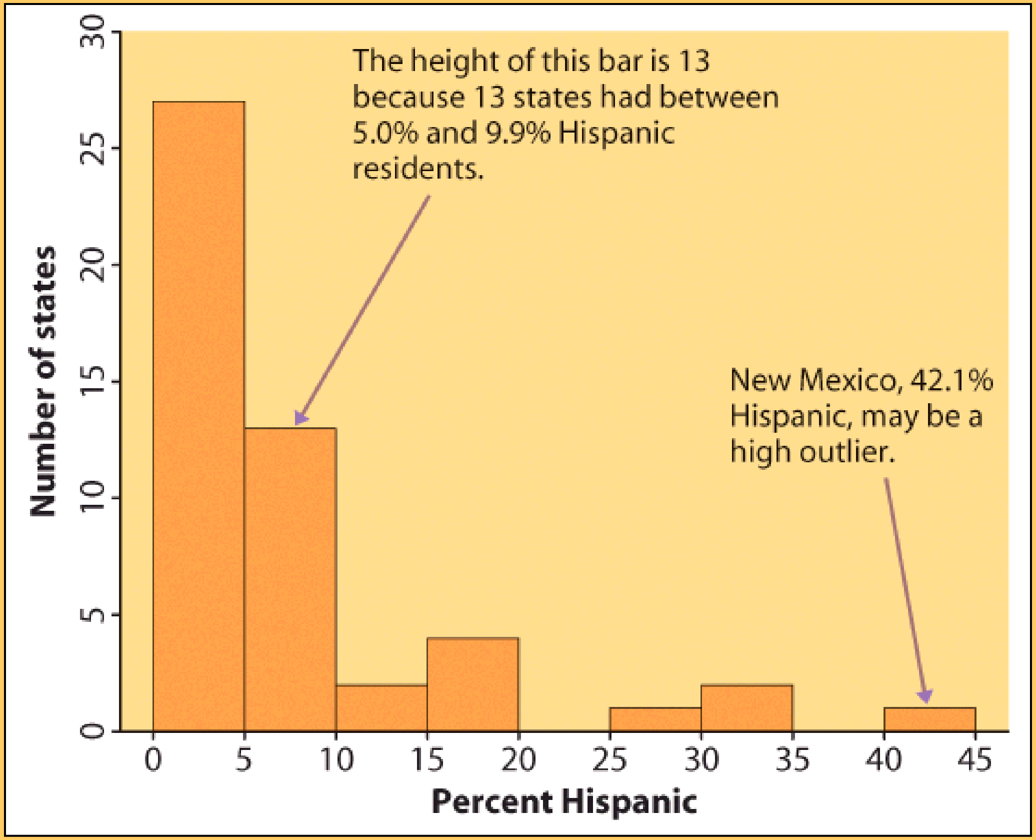
Stem Plot

- To compare two related distributions, a back-to-back stem plot with common stems is useful.
- Stem plots do not work well for large datasets.
- When the observed values have too many digits, trim the numbers before making a stem plot.
- When plotting a moderate number of observations, you can split each stem.

Histograms

The range of values that a variable can take is divided into equal size intervals.

The histogram shows the number of individual data points that fall in each interval.



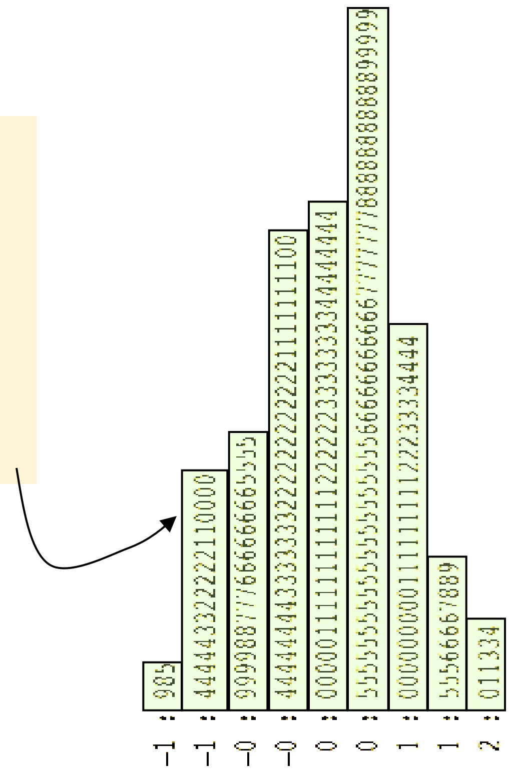
The first column represents all states with a Hispanic percent in their population between 0% and 4.99%. The height of the column shows how many states (27) have a percent in this range.

The last column represents all states with a Hispanic percent in their population between 40% and 44.99%. There is only one such state: New Mexico, at 42.1% Hispanics.

Stemplots versus histograms

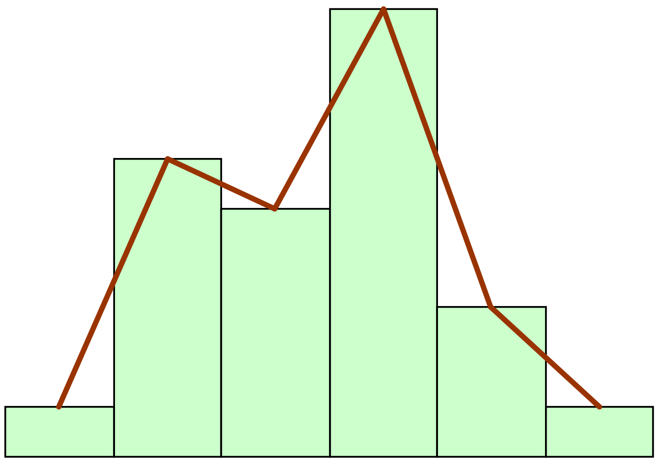
Stemplots are quick and dirty histograms that can easily be done by hand, and therefore are very convenient for back of the envelope calculations. However, they are rarely found in scientific or laymen publications.

```
—1  :  985
—1  :  44444332222110000
—0  :  999988777666666665555
—0  :  444444443333333222222222222221111111100
 0  :  0000011111111111122222223333333344444444
 0  :  55555555555555555555566666666666777777788888888888899999
 1  :  00000000011111111122233334444
 1  :  55566667889
 2  :  011334
```

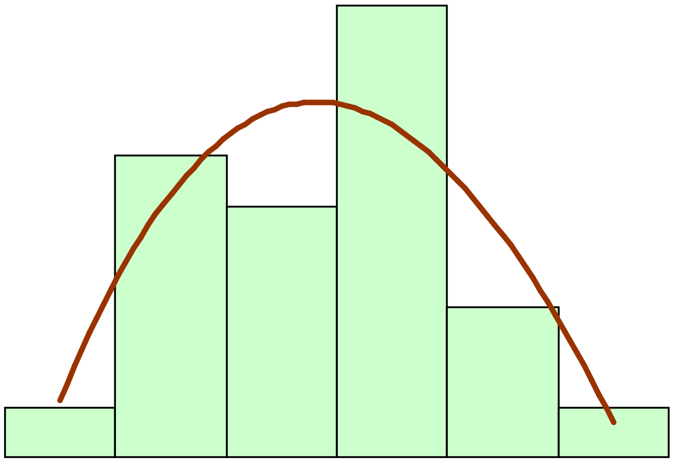


Interpreting histograms

When describing the distribution of a quantitative variable, we look for the overall pattern and for striking deviations from that pattern. We can describe the *overall* pattern of a histogram by its **shape, center,** and **spread.**



Histogram with a line connecting each column → too detailed

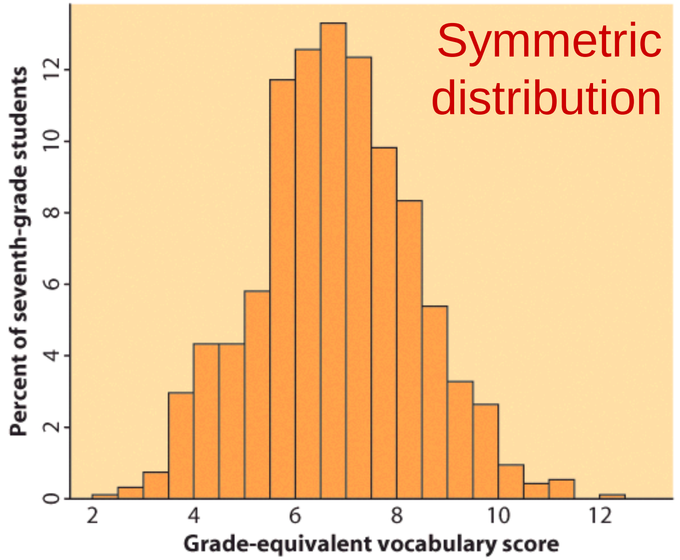


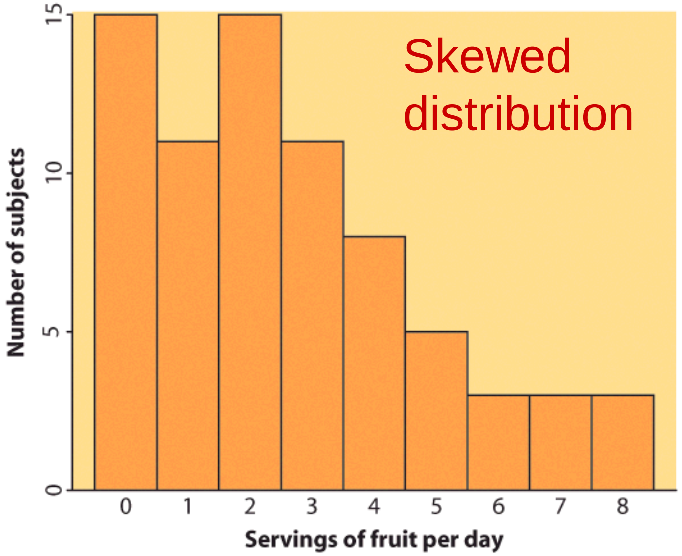
Histogram with a smoothed curve highlighting the overall pattern of the distribution

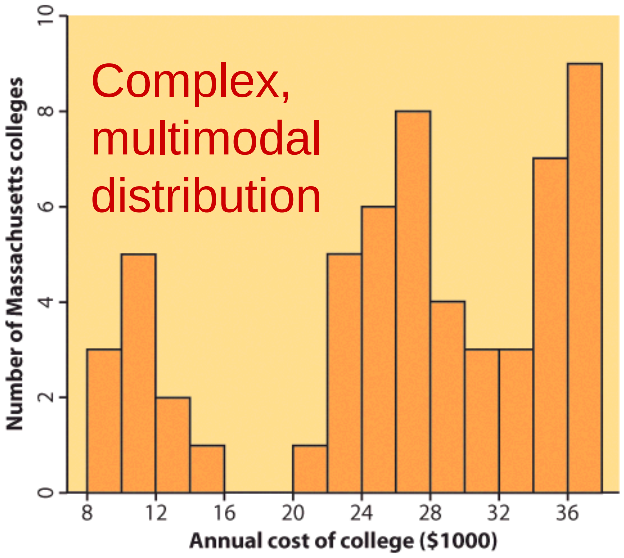
Most common distribution shapes

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.







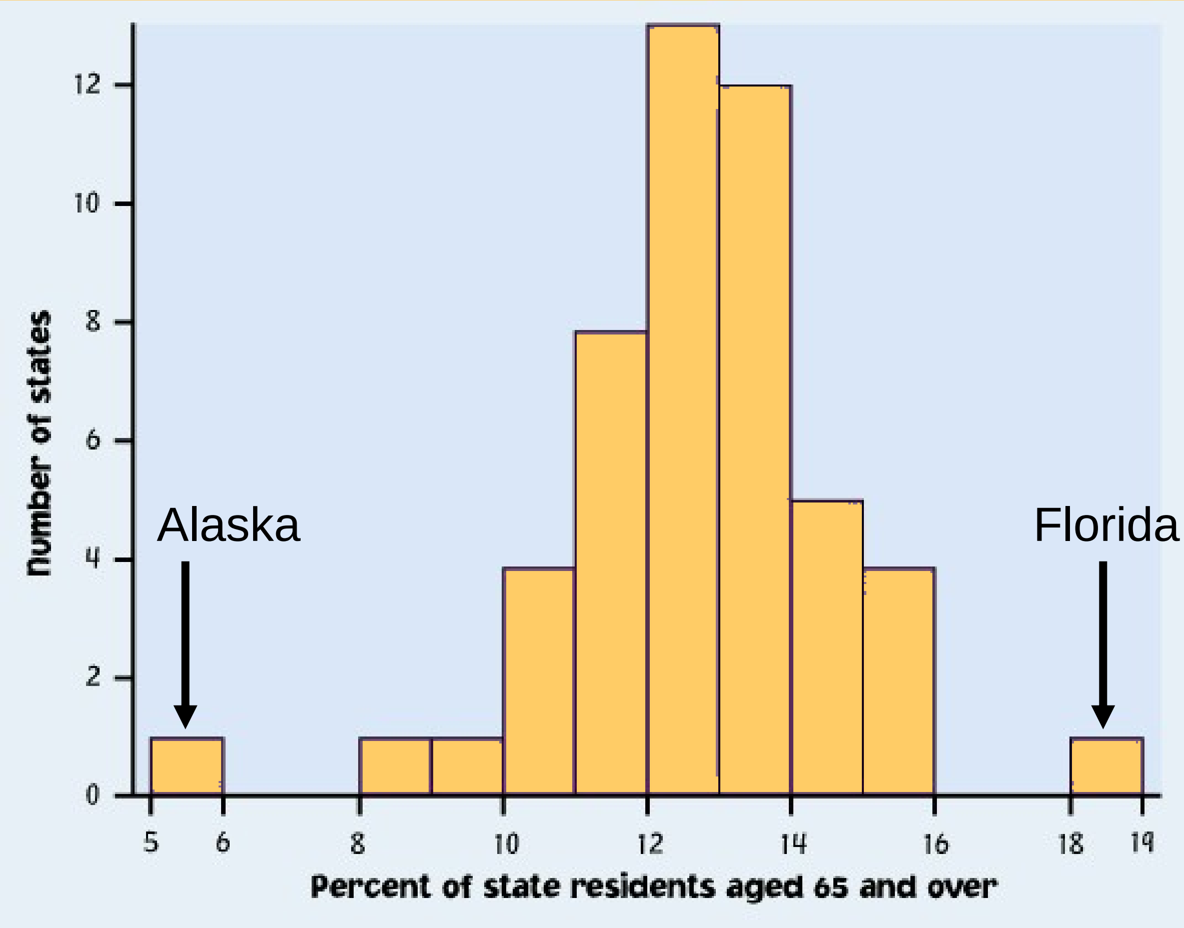
- Not all distributions have a simple overall shape, especially when there are few observations.

Outliers

An important kind of deviation is an **outlier.** Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

The overall pattern is fairly symmetrical except for 2 states that clearly do not belong to the main trend. Alaska and Florida have unusual representation of the elderly in their population.

A large gap in the distribution is typically a sign of an outlier.



How to create a histogram

It is an iterative process – try and try again.

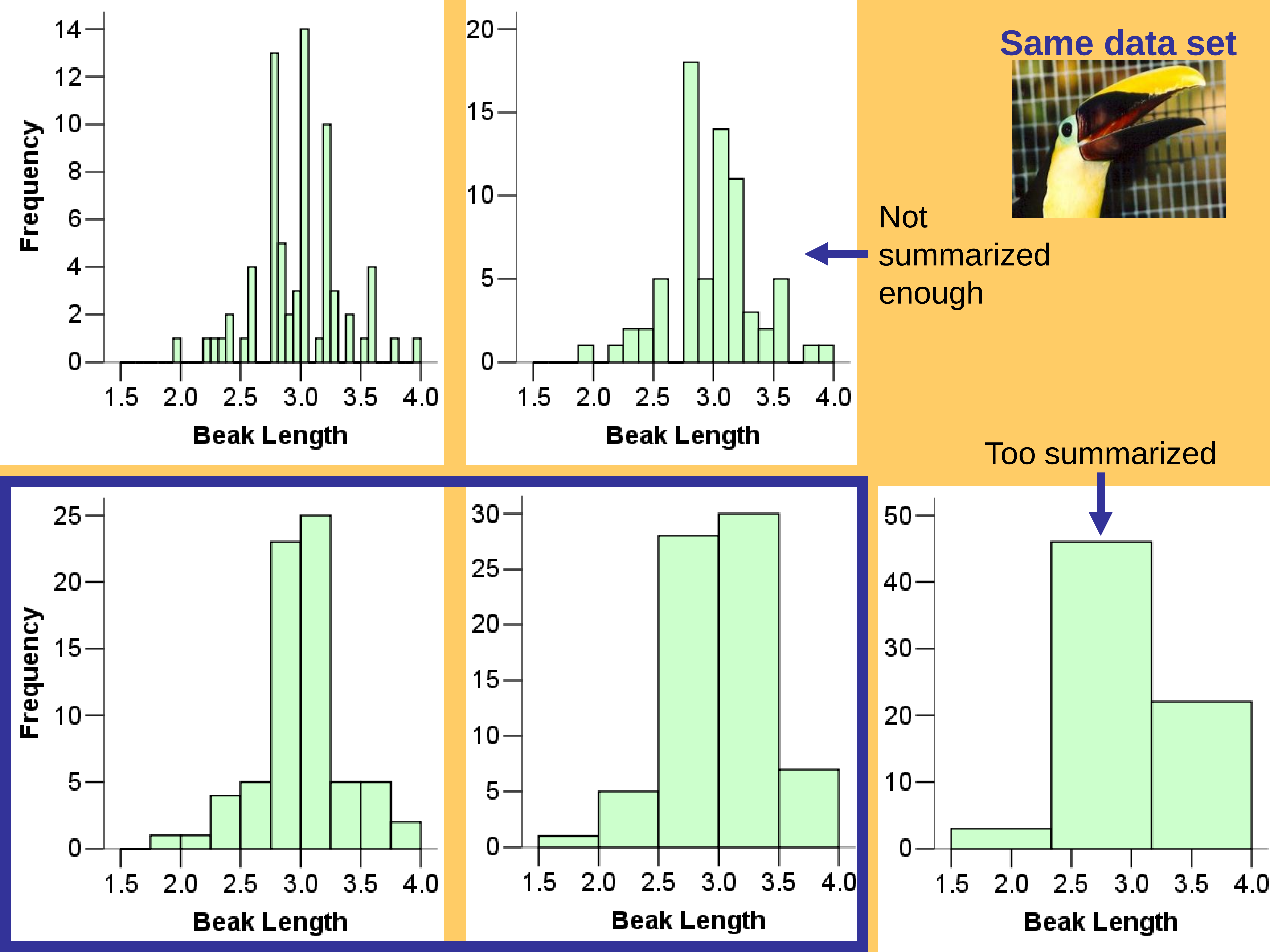
What bin size should you use?

- Not too many bins with either 0 or 1 counts
- Not overly summarized that you lose all the information
- Not so detailed that it is no longer summary

➔ rule of thumb: start with 5 to 10 bins

Look at the distribution and refine your bins

(There isn't a unique or "perfect" solution)

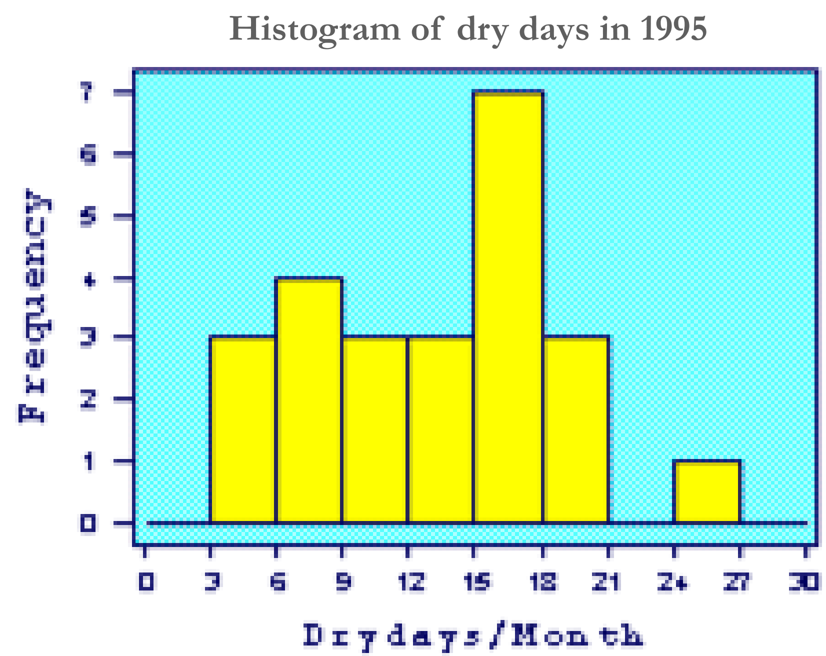


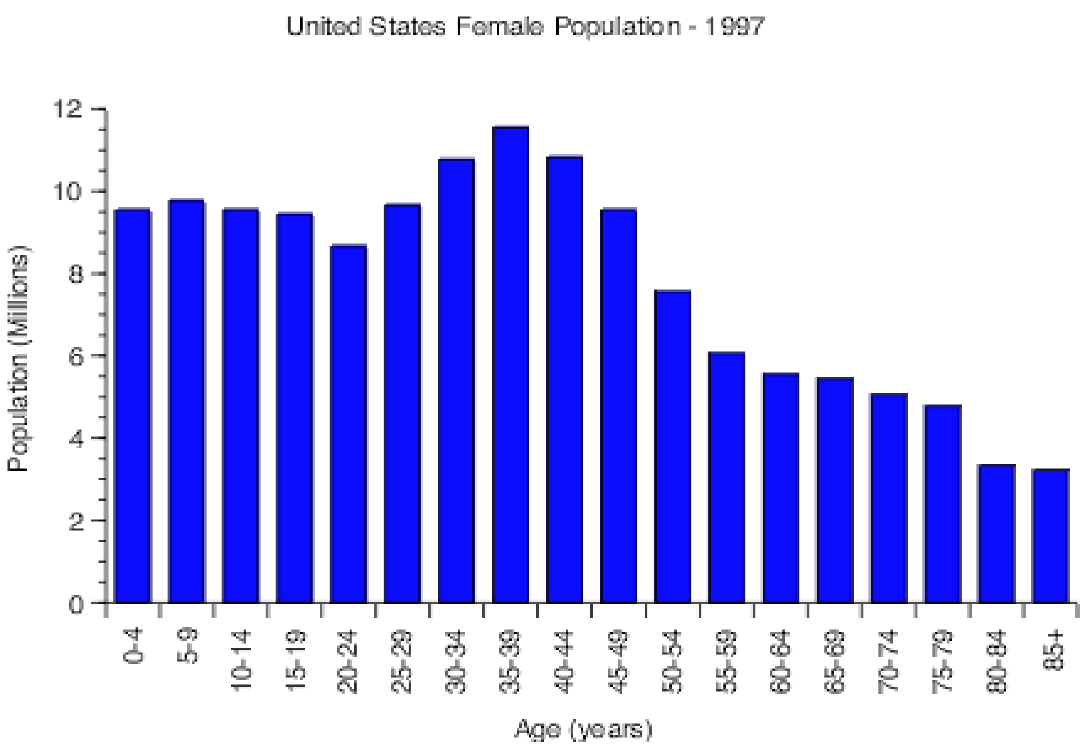
Same data set
Not summarized enough
Too summarized
Frequency
Beak Length
1.5
2.0
2.5
3.0
3.5
4.0
0
2
4
6
8
10
12
14
5
15
20
25
30
10
40
50

IMPORTANT NOTE:

Your data are the way they are.

Do not try to force them into a particular shape.



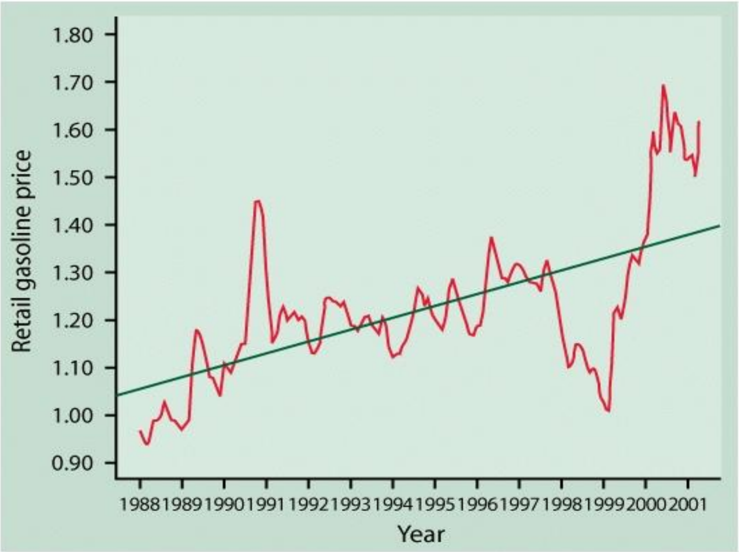


It is a common misconception that if you have a large enough data set, the data will eventually turn out nice and symmetrical.

Line graphs: time plots

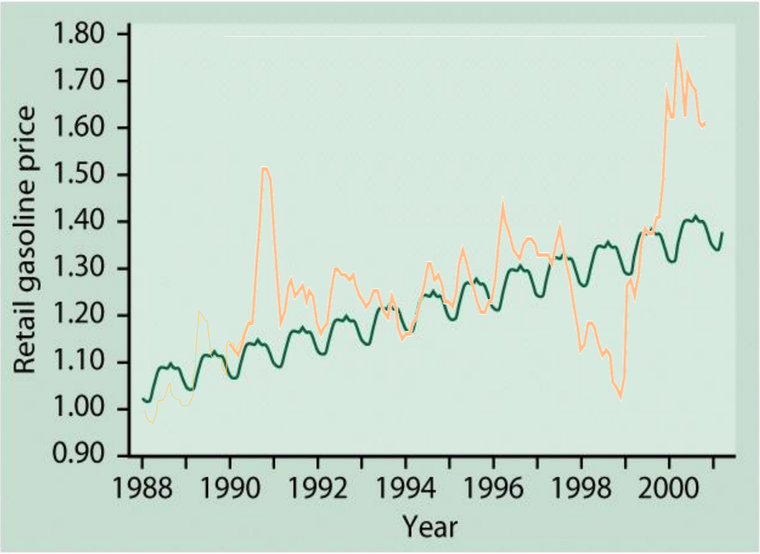
In a time plot, time always goes on the horizontal, x axis.

We describe time series by looking for an overall pattern and for striking deviations from that pattern. In a time series:

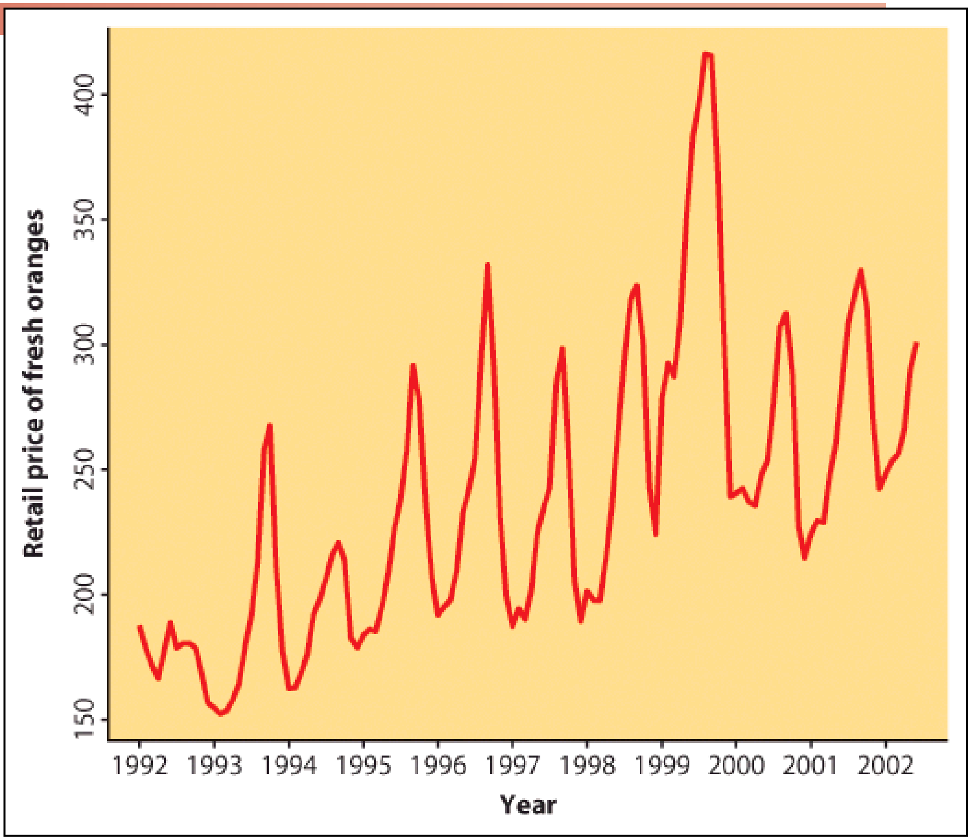


A **trend** is a rise or fall that persists over time, despite small irregularities.

A pattern that repeats itself at regular intervals of time is called **seasonal variation**.



Retail price of fresh oranges over time



Time is on the horizontal, x axis.

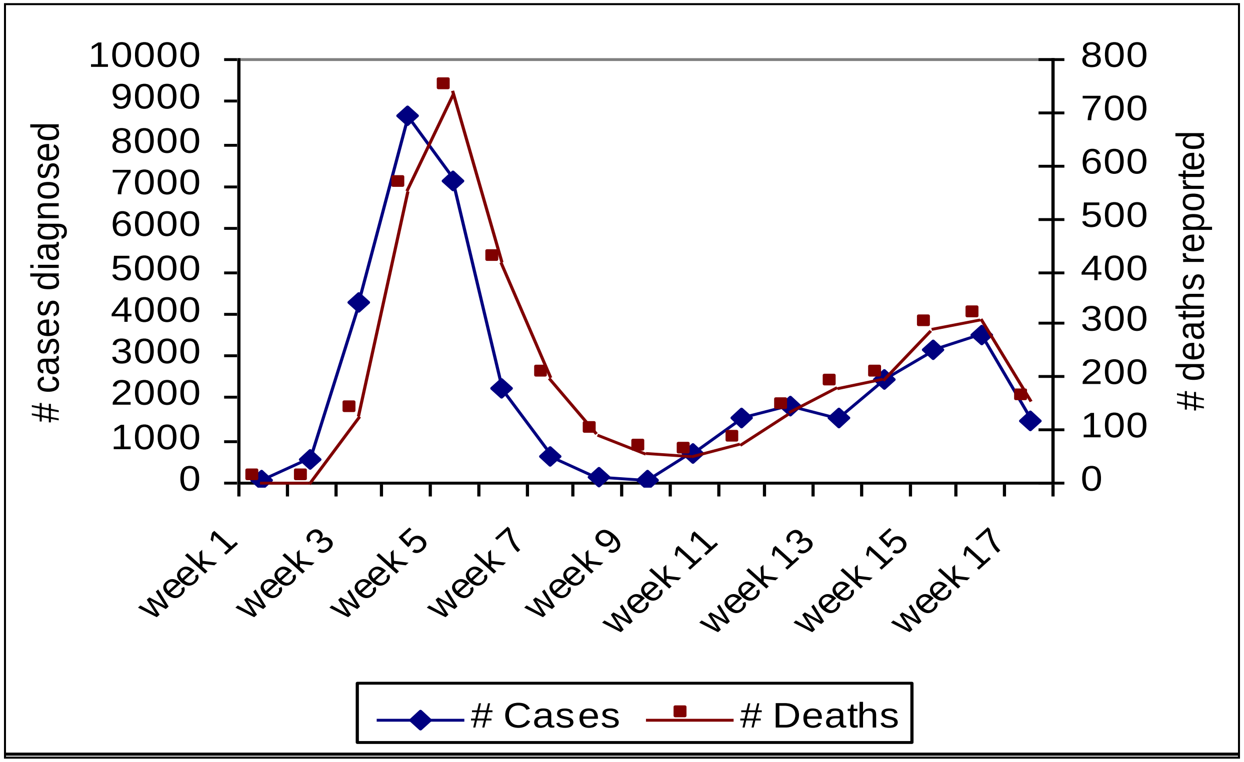
The variable of interest—here “retail price of fresh oranges”—goes on the vertical, y axis.

This time plot shows a regular pattern of yearly variations. These are seasonal variations in fresh orange pricing most likely due to similar seasonal variations in the production of fresh oranges.

There is also an overall upward trend in pricing over time. It could simply be reflecting inflation trends or a more fundamental change in this industry.

A time plot can be used to compare two or more data sets covering the same time period.

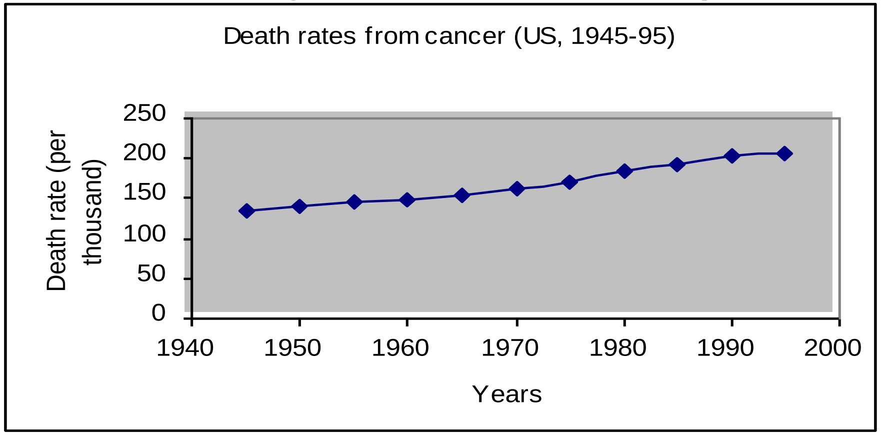
1918 influenza epidemic		
Date	**# Cases**	**# Deaths**
week 1	36	0
week 2	531	0
week 3	4233	130
week 4	8682	552
week 5	7164	738
week 6	2229	414
week 7	600	198
week 8	164	90
week 9	57	56
week 10	722	50
week 11	1517	71
week 12	1828	137
week 13	1539	178
week 14	2416	194
week 15	3148	290
week 16	3465	310
week 17	1440	149

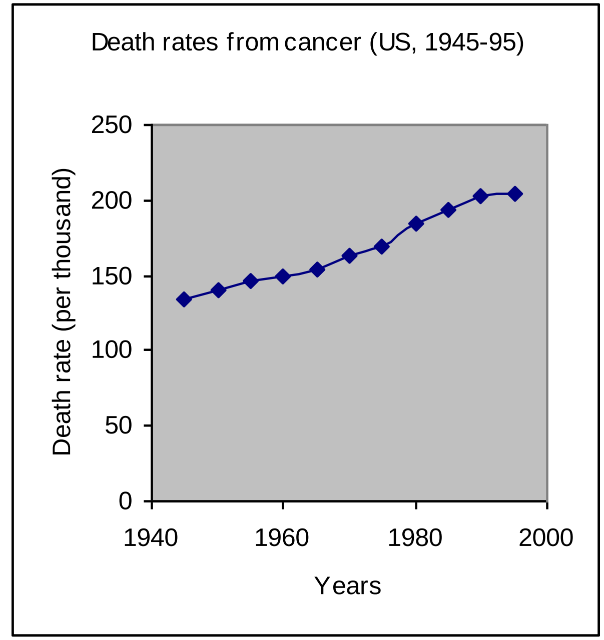


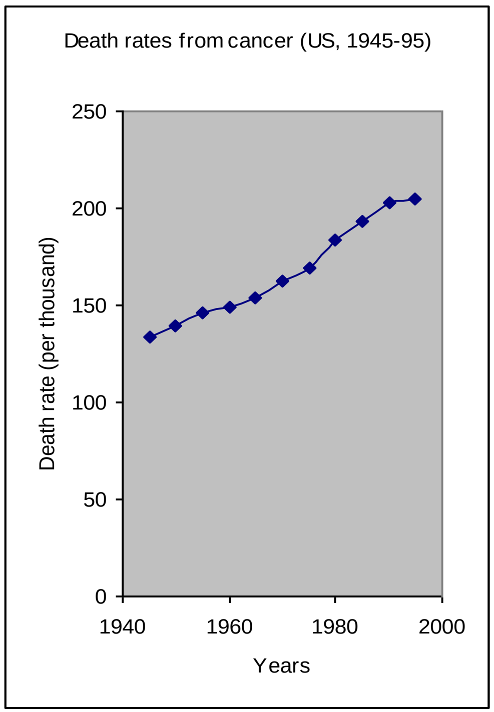
The pattern over time for the number of flu diagnoses closely resembles that for the number of deaths from the flu, indicating that about 8% to 10% of the people diagnosed that year died shortly afterward, from complications of the flu.

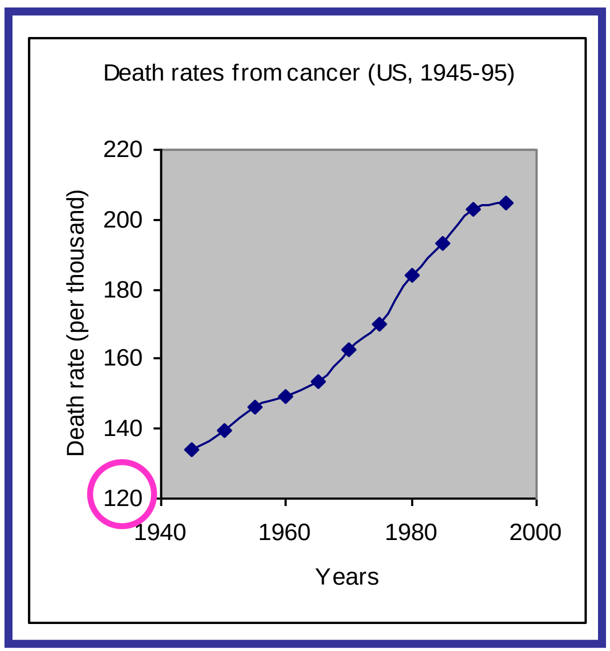
Scales matter

How you stretch the axes and choose your scales can give a different impression.







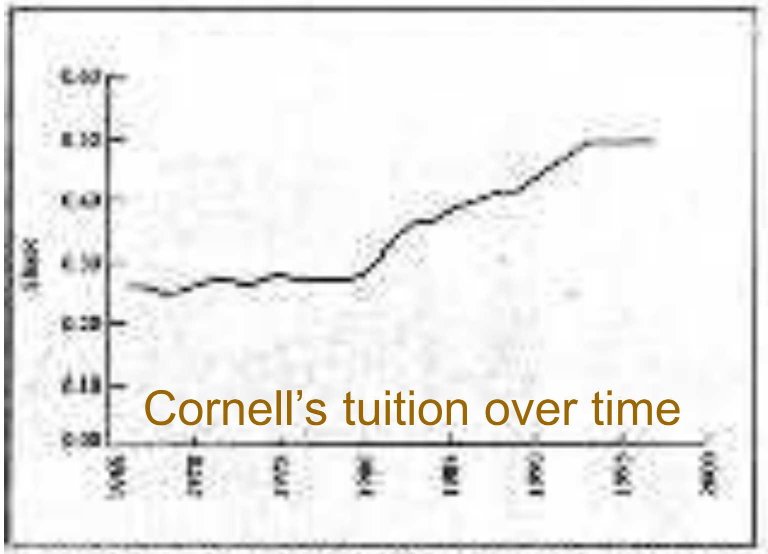


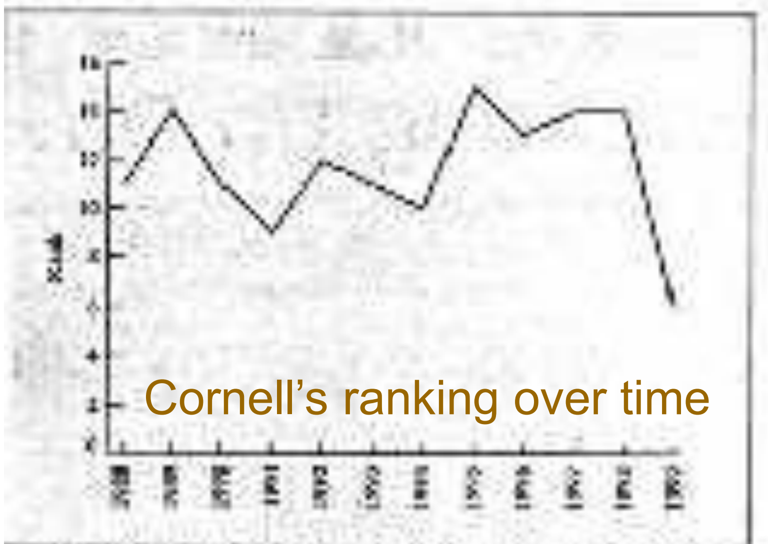
A picture is worth a thousand words,

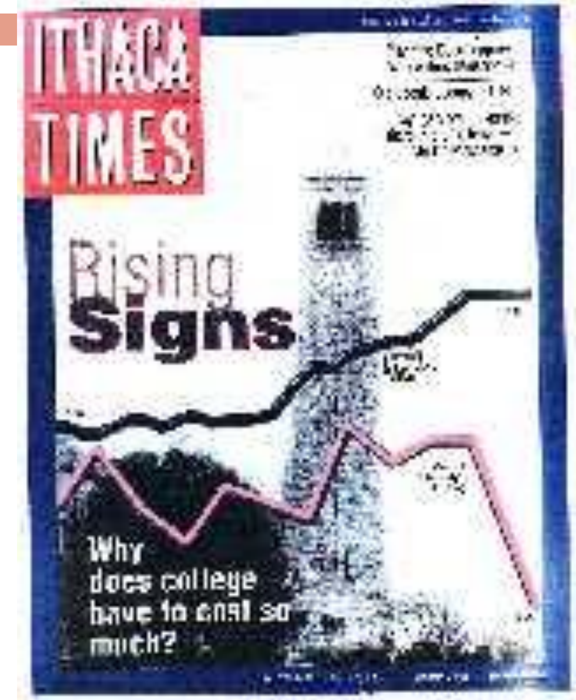
BUT

There is nothing like hard numbers.
→ **Look at the scales.**

Why does it matter?







What's wrong with these graphs?

Careful reading reveals that:

1. The ranking graph covers an 11-year period, the tuition graph 35 years, yet they are shown comparatively on the cover and without a horizontal time scale.

2. Ranking and tuition have very different units, yet both graphs are placed on the same page without a vertical axis to show the units.

3. The impression of a recent sharp "drop" in the ranking graph actually shows that Cornell's rank has IMPROVED from 15th to 6th ...

Looking at Data—Distributions

1.2 Describing distributions with numbers

Objectives

1.2 Describing distributions with numbers

- Measures of center: **mean, median**
- Mean versus median
- Measures of spread: **quartiles, standard deviation**
- Five-number summary and **boxplot**
- Choosing among summary statistics
- Changing the unit of measurement

Measure of center: the **mean**

The mean or arithmetic average

To calculate the *average,* or **mean,** add all values, then divide by the number of cases. It is the “center of mass.”

Sum of heights is 1598.3
divided by 25 women = 63.9 inches

58.2	64.0
59.5	64.5
60.7	64.1
60.9	64.8
61.9	65.2
61.9	65.7
62.2	66.2
62.2	66.7
62.4	67.1
62.9	67.8
63.9	68.9
63.1	69.6
63.9	

woman (i)	height (x)		woman (i)	height (x)
i = 1	x_1= 58.2		i = 14	x_{14}= 64.0
i = 2	x_2= 59.5		i = 15	x_{15}= 64.5
i = 3	x_3= 60.7		i = 16	x_{16}= 64.1
i = 4	x_4= 60.9		i = 17	x_{17}= 64.8
i = 5	x_5= 61.9		i = 18	x_{18}= 65.2
i = 6	x_6= 61.9		i = 19	x_{19}= 65.7
i = 7	x_7= 62.2		i = 20	x_{20}= 66.2
i = 8	x_8= 62.2		i = 21	x_{21}= 66.7
i = 9	x_9= 62.4		i = 22	x_{22}= 67.1
i = 10	x_{10}= 62.9		i = 23	x_{23}= 67.8
i = 11	x_{11}= 63.9		i = 24	x_{24}= 68.9
i = 12	x_{12}= 63.1		i = 25	x_{25}= 69.6
i = 13	x_{13}= 63.9		**n=25**	**Σ=1598.3**

Mathematical notation:

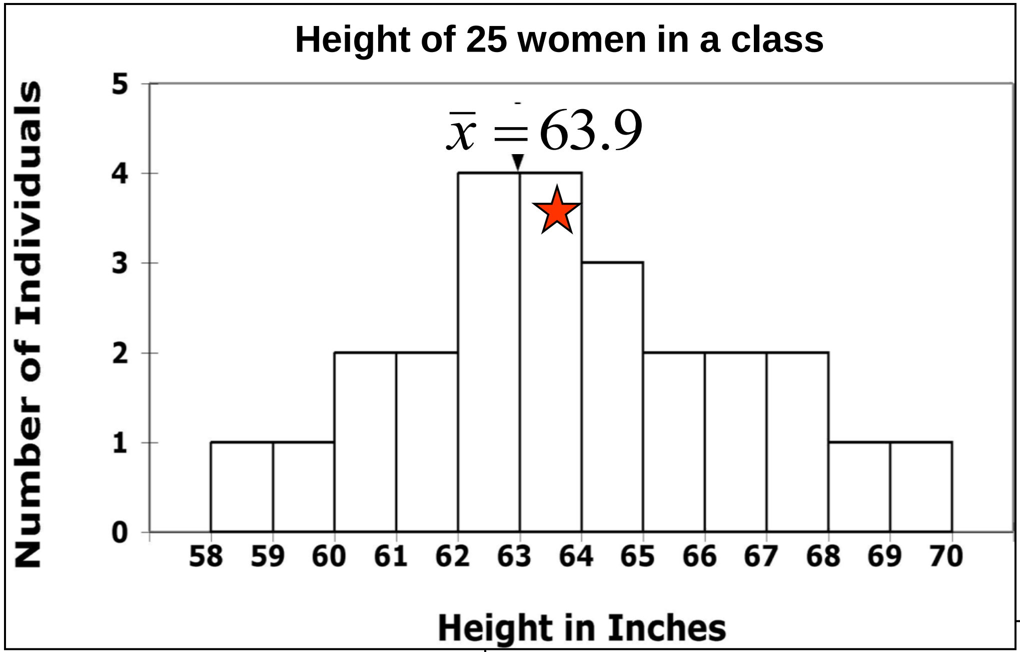
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

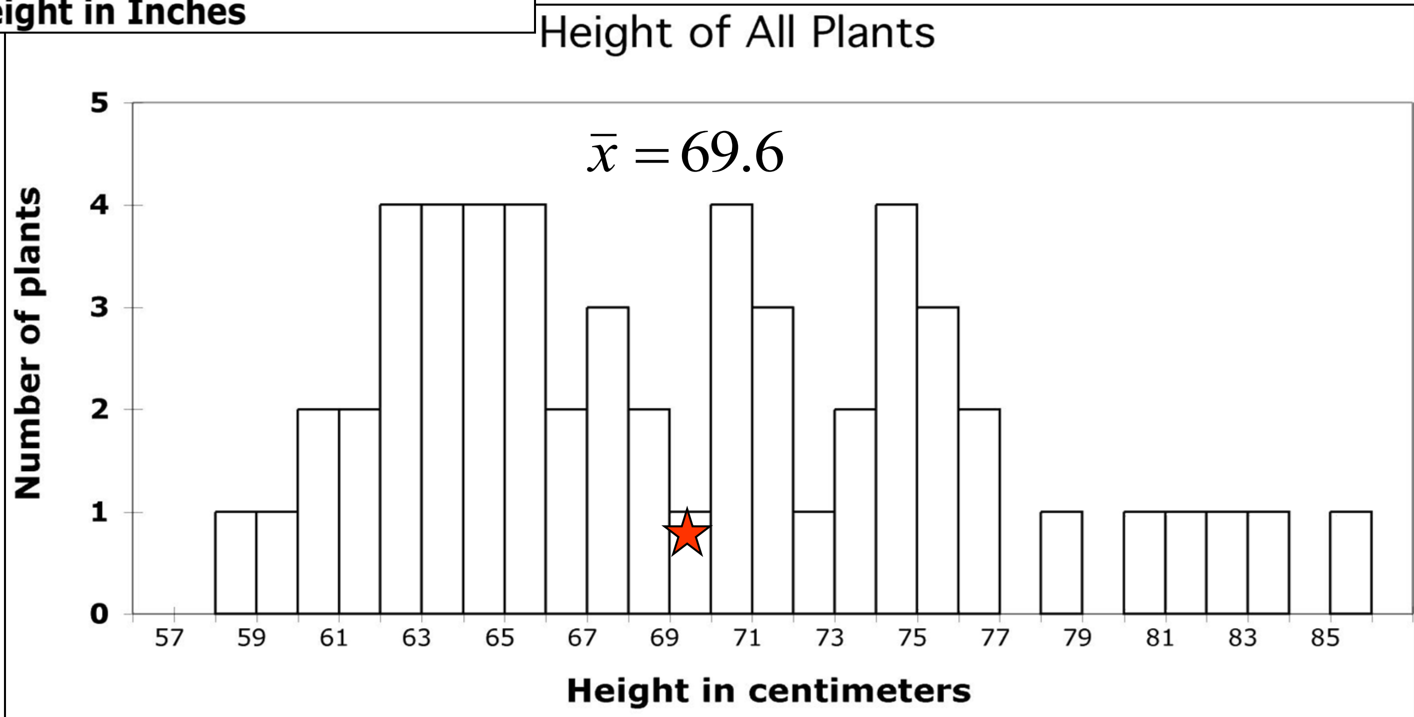
$$\bar{x} = \frac{1598.3}{25} = 63.9$$

Learn right away how to get the mean using your calculators.

Your numerical summary must be meaningful.

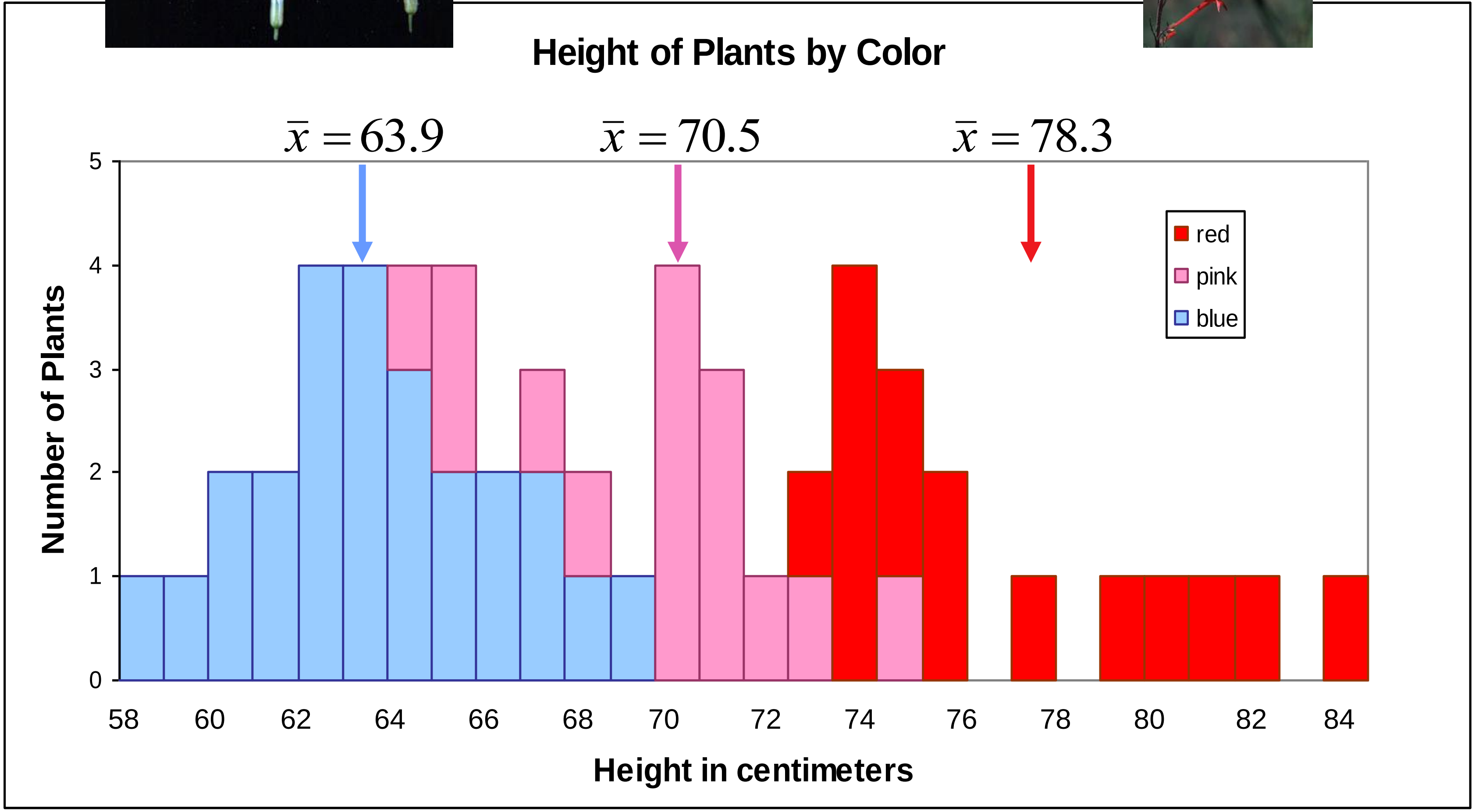


The distribution of women's heights appears coherent and symmetrical. The mean is a good numerical summary.



Here the shape of the distribution is wildly irregular. Why?

Could we have more than one plant species or phenotype?



A single numerical summary here would not make sense.

Measure of center: the **median**

The **median** is the midpoint of a distribution—the number such that half of the observations are smaller and half are larger.

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12	12	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6
25	**12**	**6.1**

1. Sort observations by size.
n = number of observations

2.a. If n is **odd,** the median is observation $(n+1)/2$ down the list

← $n = 25$
$(n+1)/2 = 26/2 = 13$
Median = 3.4

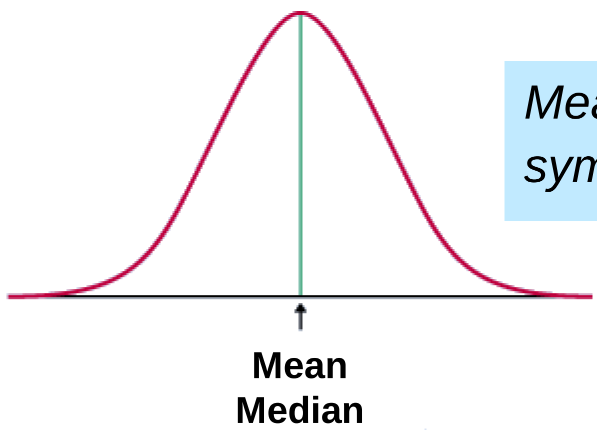
2.b. If n is **even,** the median is the mean of the two middle observations.

$n = 24$ →
$n/2 = 12$
Median = (3.3+3.4) /2 = 3.35

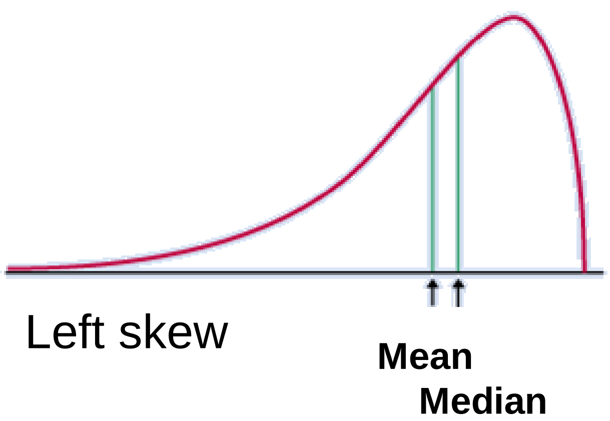
1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12		3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6

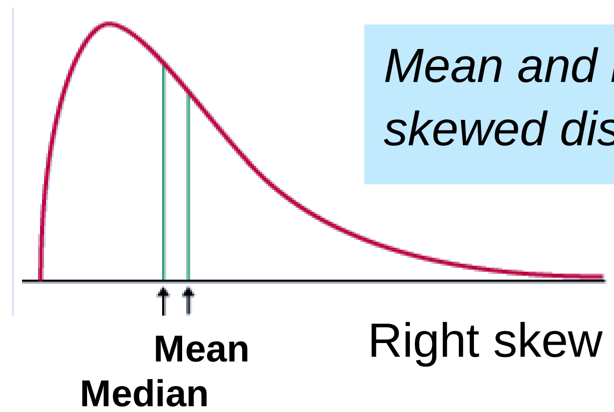
Comparing the mean and the median

The mean and the median are the same only if the distribution is symmetrical. The median is a measure of center that is resistant to skew and outliers. The mean is not.



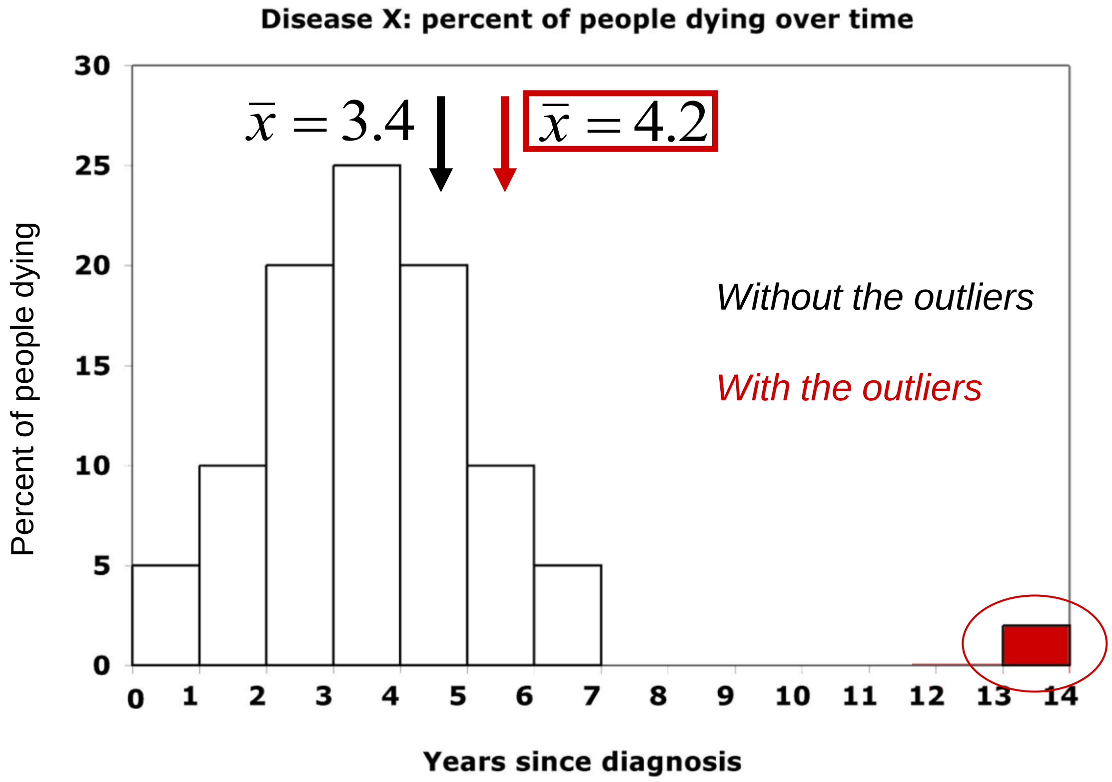
Mean and median for a symmetric distribution





Mean and median for skewed distributions

Mean and median of a distribution with outliers



The mean is pulled to the right a lot by the outliers (from 3.4 to 4.2).

The median, on the other hand, is only slightly pulled to the right by the outliers (from 3.4 to 3.6).

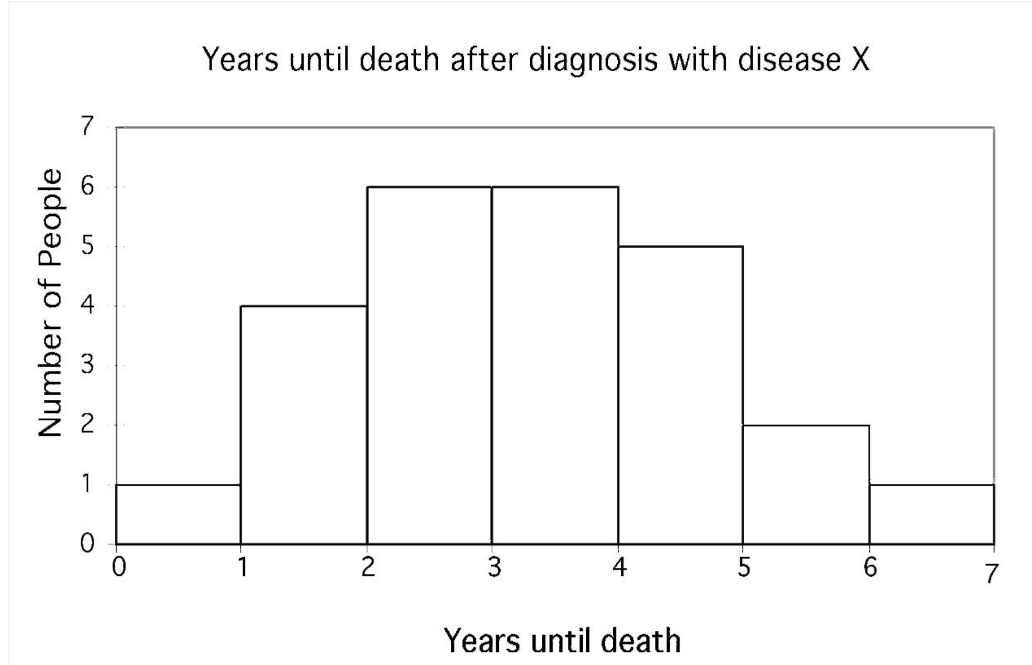
Impact of skewed data

Symmetric distribution…

Disease X: $\bar{x} = 3.4$

$M = 3.4$

Mean and median are the same.

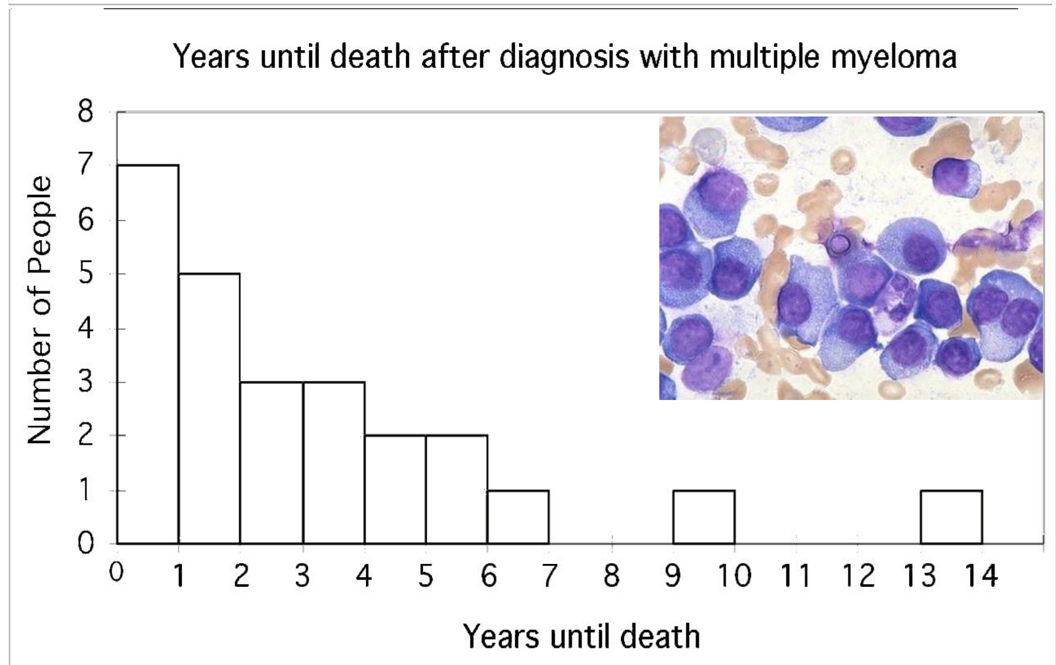


… and a right-skewed distribution

Multiple myeloma: $\bar{x} = 3.4$

$M = 2.5$

The mean is pulled toward the skew.



Measure of spread: the **quartiles**

The **first quartile, Q_1,** is the value in the sample that has 25% of the data at or below it (⇔ it is the median of the lower half of the sorted data, excluding *M*).

The **third quartile, Q_3,** is the value in the sample that has 75% of the data at or below it (⇔ it is the median of the upper half of the sorted data, excluding *M*).

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	1	2.3
9	2	2.5
10	3	2.8
11	4	2.9
12	5	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	1	4.7
22	2	4.9
23	3	5.3
24	4	5.6
25	5	6.1

Q_1= first quartile = 2.2

M = median = *3.4*

Q_3= third quartile = 4.35

Five-number summary and boxplot

25	6	6.1
24	5	5.6
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

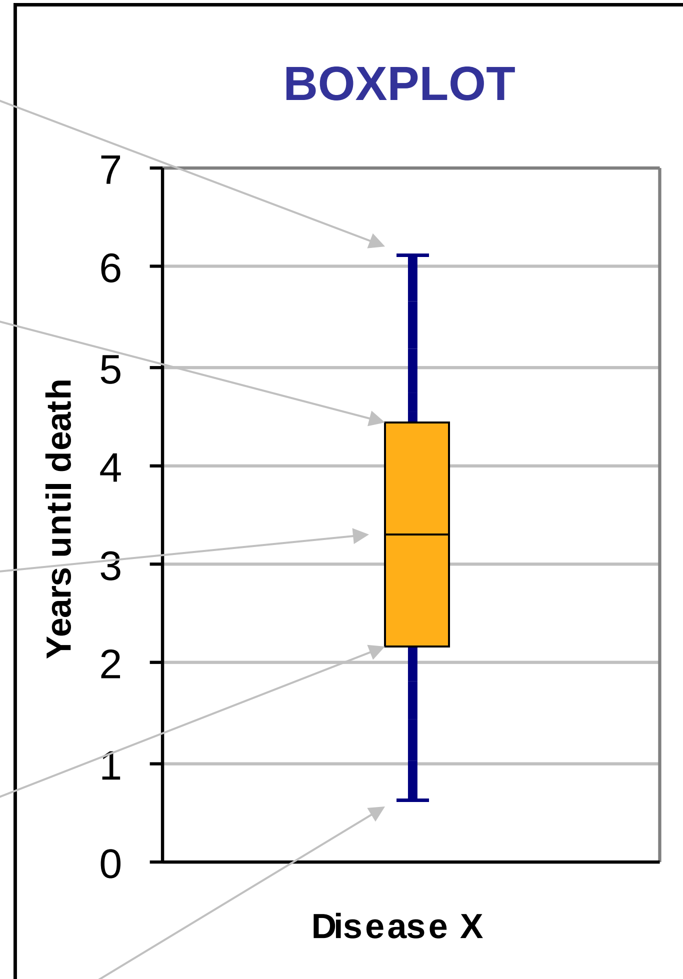
Largest = max = 6.1

Q_3= third quartile
= 4.35

M = median = 3.4

Q_1= first quartile
= 2.2

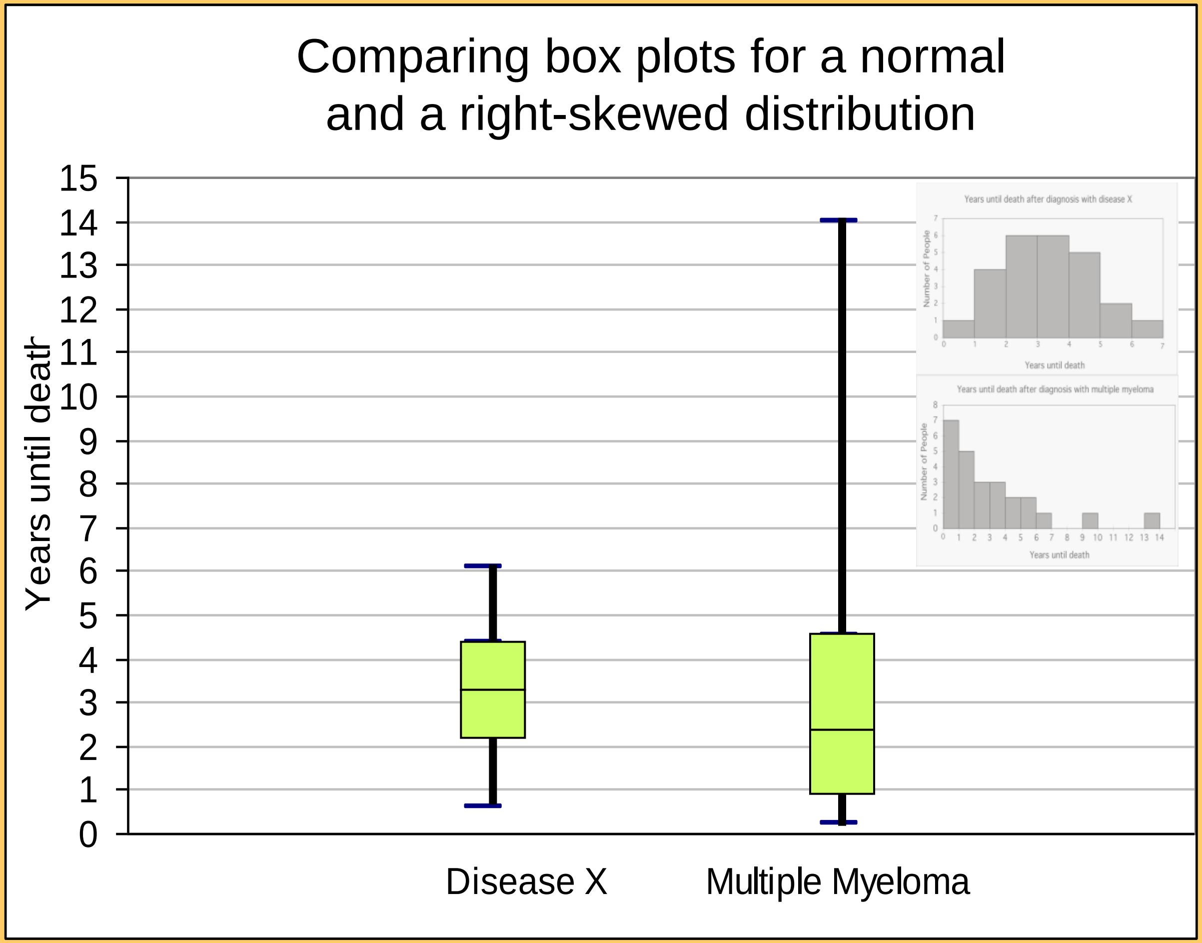
Smallest = min = 0.6



Five-number summary:

min Q_1 M Q_3 max

Boxplots for skewed data



Boxplots remain true to the data and depict clearly symmetry or skew.

Suspected outliers

Outliers are troublesome data points, and it is important to be able to identify them.

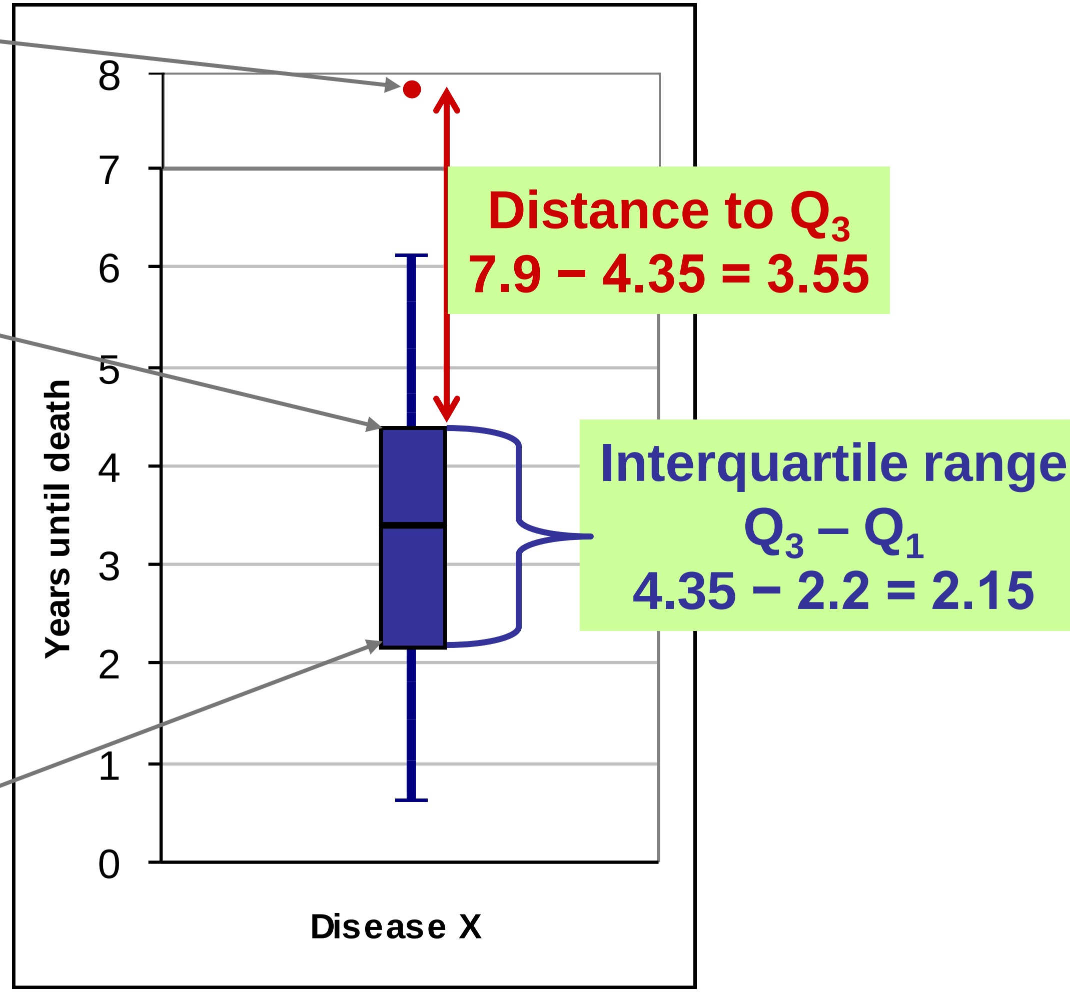
One way to raise the flag for a suspected outlier is to compare the distance from the suspicious data point to the nearest quartile (Q_1 or Q_3). We then compare this distance to the **interquartile range** (distance between Q_1 and Q_3).

We call an observation a **suspected outlier** if it falls more than 1.5 times the size of the interquartile range (IQR) below the first quartile or above the third quartile. This is called the "**1.5 * IQR rule for outliers.**"

25	6	**7.9**
24	5	6.1
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

$Q_3 = 4.35$

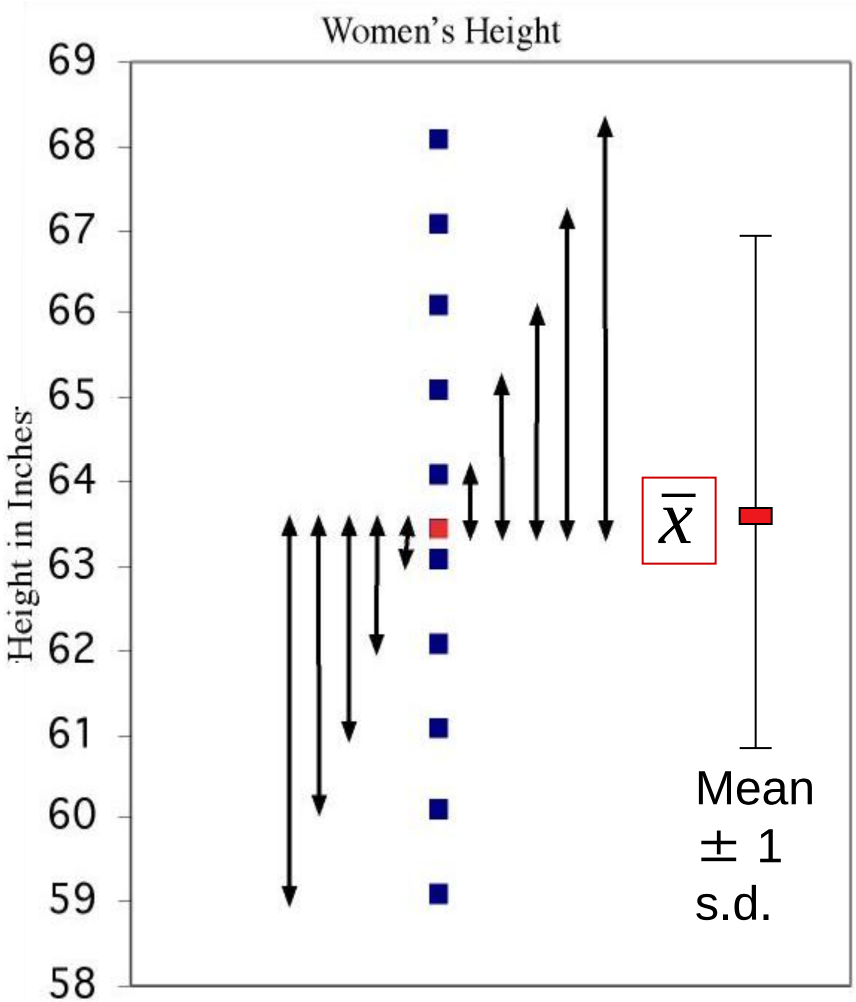
$Q_1 = 2.2$



Individual #25 has a value of 7.9 years, which is 3.55 years above the third quartile. This is more than 3.225 years, 1.5 * IQR. Thus, individual #25 is a suspected outlier.

Measure of spread: the **standard deviation**

The standard deviation "*s*" is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.



1. First calculate the **variance s^2**.

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

2. Then take the square root to get the **standard deviation *s***.

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

Calculations …

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

Mean = 63.4

Sum of squared deviations from mean = 85.2

Degrees freedom (df) = (n − 1) = 13

s^2 = variance = 85.2/13 = 6.55 inches squared

s = standard deviation = √6.55 = 2.56 inches

Women's height (inches)

i	x_i	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
1	59	63.4	-4.4	19.0
2	60	63.4	-3.4	11.3
3	61	63.4	-2.4	5.6
4	62	63.4	-1.4	1.8
5	62	63.4	-1.4	1.8
6	63	63.4	-0.4	0.1
7	63	63.4	-0.4	0.1
8	63	63.4	-0.4	0.1
9	64	63.4	0.6	0.4
10	64	63.4	0.6	0.4
11	65	63.4	1.6	2.7
12	66	63.4	2.6	7.0
13	67	63.4	3.6	13.3
14	68	63.4	4.6	21.6
	Mean 63.4		Sum 0.0	Sum 85.2

We'll never calculate these by hand, so make sure to know how to get the standard deviation using your calculator or software.

Variance and Standard Deviation

- *Why do we square the deviations?*
  - The sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be.
  - The sum of the deviations of any set of observations from their mean is always zero.

- *Why do we emphasize the standard deviation rather than the variance?*
  - s, not s^2, is the natural measure of spread for Normal distributions.
  - s has the same unit of measurement as the original observations.

- *Why do we average by dividing by $n - 1$ rather than n in calculating the variance?*
  - The sum of the deviations is always zero, so only $n - 1$ of the squared deviations can vary freely.
  - The number $n - 1$ is called the **degrees of freedom**.

Properties of Standard Deviation

- *s* measures spread about the mean and should be used only when the mean is the measure of center.
- *s* = 0 only when all observations have the same value and there is no spread. Otherwise, *s* > 0.
- *s* is not resistant to outliers.
- *s* has the same units of measurement as the original observations.

Software output for summary statistics:

Excel - From Menu: Tools/Data Analysis/ Descriptive Statistics

Give common statistics of your sample data.

Microsoft Excel - ta01-01.dat

	A	B	C	D
1	*Pct Hisp*			
2				
3	Mean	7.73	QUARTILE(B2:B51,1)	2.175
4	Standard Error	1.2604	QUARTILE(B2:B51,3)	8.525
5	Median	4.7		
6	Mode	1.5		
7	Standard Deviation	8.9125		
8	Sample Variance	79.4332		
9	Kurtosis	5.1186		
10	Skewness	2.2450		
11	Range	41.4		
12	Minimum	0.7		
13	Maximum	42.1		
14	Sum	386.5		
15	Count	50		

Sheet2 | Sheet1 | ta01-01

Session

Descriptive Statistics: PctHisp

Variable	N	Mean	Median	TrMean	StDev	SE Mean
PctHisp	50	7.73	4.70	6.31	8.91	1.26

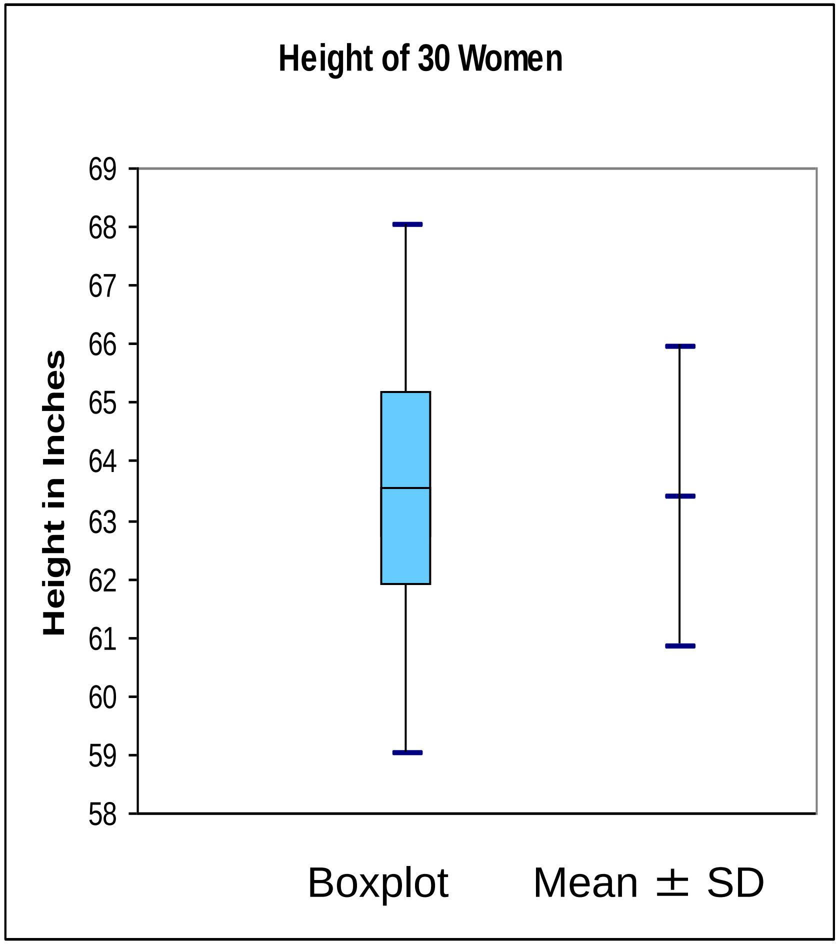
Variable	Minimum	Maximum	Q1	Q3
PctHisp	0.70	42.10	2.08	8.77

Minitab

Choosing among summary statistics

- Because the **mean** is not resistant to outliers or skew, use it to describe distributions that are fairly symmetrical and don't have outliers.
  → Plot the mean and use the standard deviation for error bars.

- Otherwise use the **median** in the five number summary which can be plotted as a boxplot.



What should you use, when, and why?

Arithmetic mean or median?

- Middletown is considering imposing an income tax on citizens. City hall wants a numerical summary of its citizens' income to estimate the total tax base.
  - **Mean**: Although income is likely to be right-skewed, the city government wants to know about the total tax base.
- In a study of standard of living of typical families in Middletown, a sociologist makes a numerical summary of family income in that city.
  - **Median**: The sociologist is interested in a "typical" family and wants to lessen the impact of extreme incomes.

Changing the unit of measurement

Variables can be recorded in different units of measurement. Most often, one measurement unit is a **linear transformation** of another measurement unit: $x_{new} = a + bx$.

Temperatures can be expressed in degrees Fahrenheit or degrees Celsius.
$\text{Temperature}^{\text{Fahrenheit}} = 32 + (9/5)^* \text{ Temperature}^{\text{Celsius}}$ ➔ $a + bx$.

Linear transformations do not change the basic shape of a distribution (skew, symmetry, multimodal). But they do change the measures of center and spread:

- Multiplying each observation by a positive number *b* multiplies both measures of center (mean, median) and spread (IQR, *s*) by *b*.
- Adding the same number *a* (positive or negative) to each observation adds *a* to measures of center and to quartiles but it does not change measures of spread (IQR, *s*).

Looking at Data—Distributions

1.3 Density Curves and Normal Distributions

Objectives

1.3 Density curves and Normal distributions

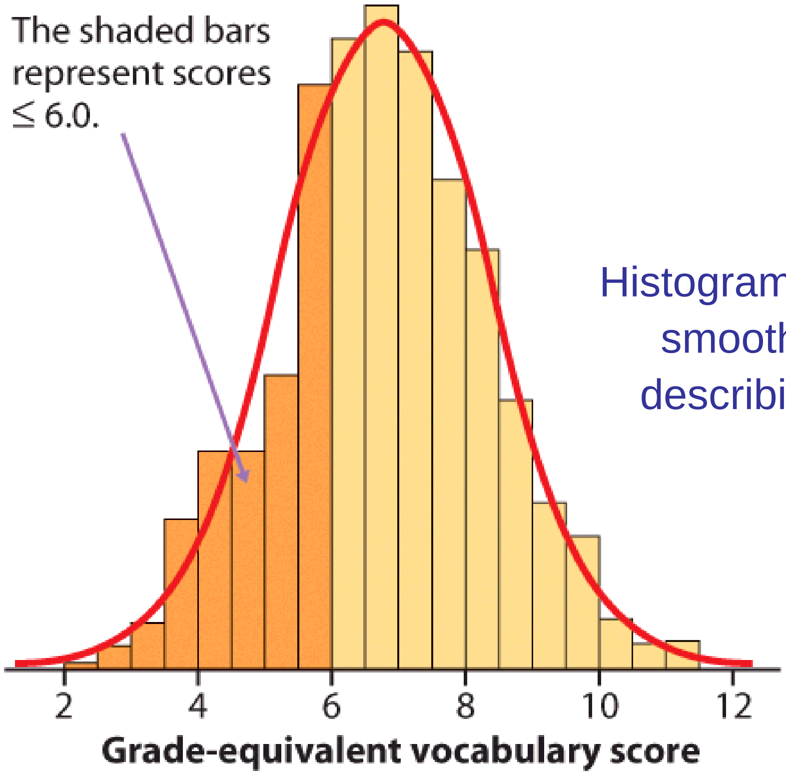
- Density curves
- Measuring center and spread for density curves
- Normal distributions
- The 68-95-99.7 rule
- Standardizing observations
- Using the standard Normal Table
- Inverse Normal calculations
- Normal quantile plots

Density curves

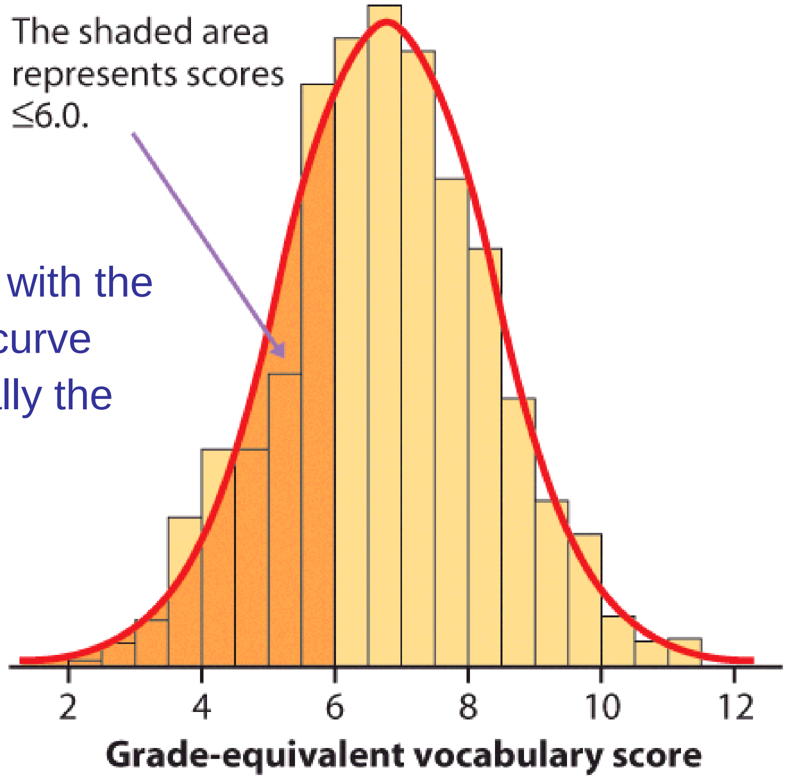
A **density curve** is a mathematical model of a distribution.

The total area under the curve, by definition, is equal to 1, or 100%.

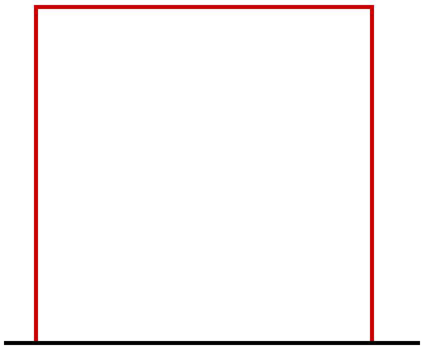
The area under the curve for a range of values is the proportion of all observations for that range.

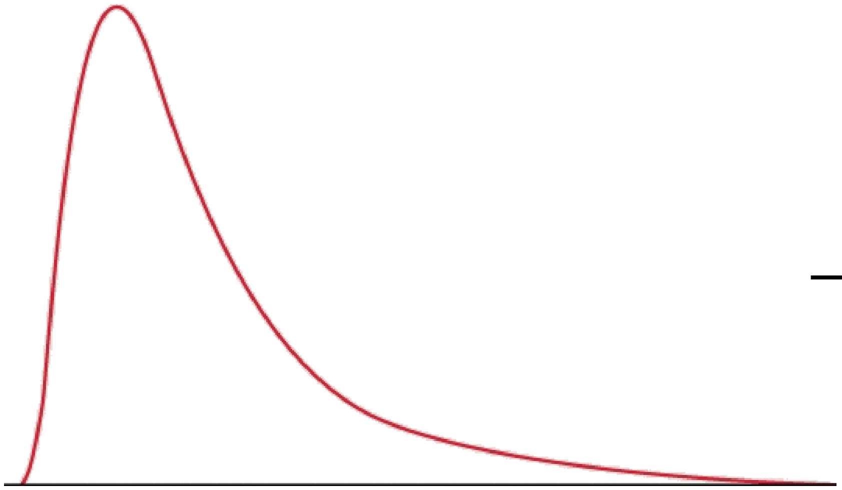


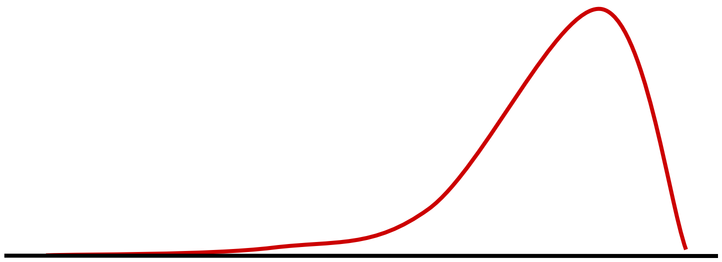
Histogram of a sample with the smoothed, density curve describing theoretically the population.

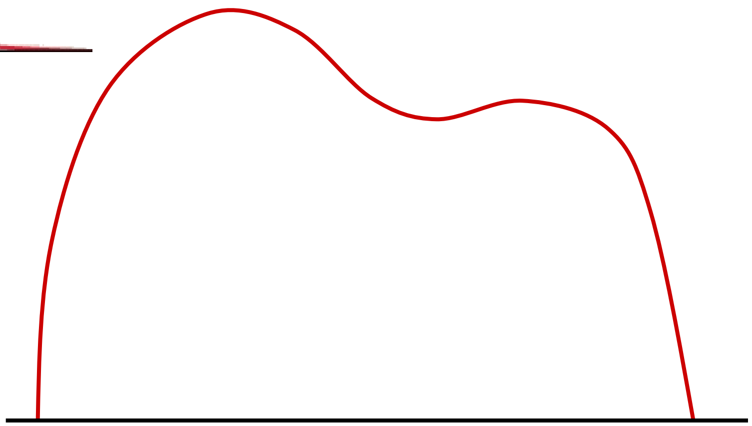


Density curves come in any imaginable shape.







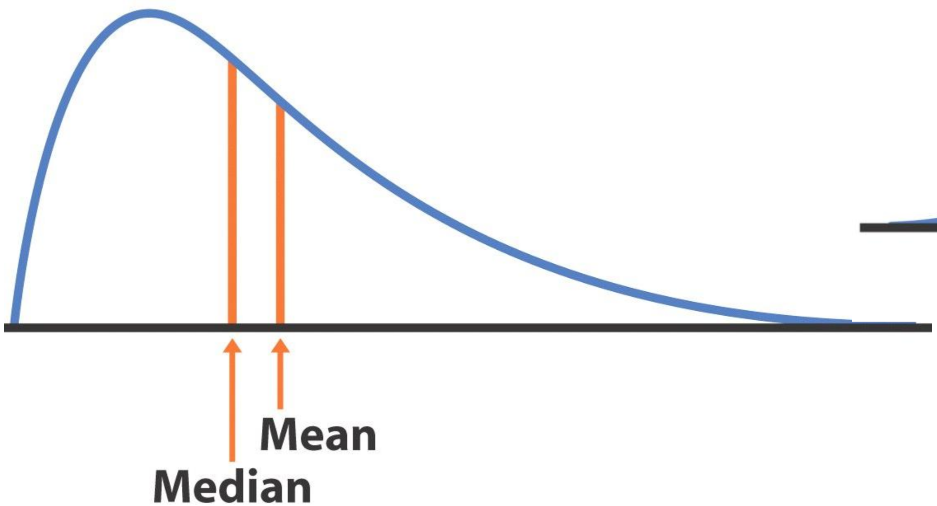


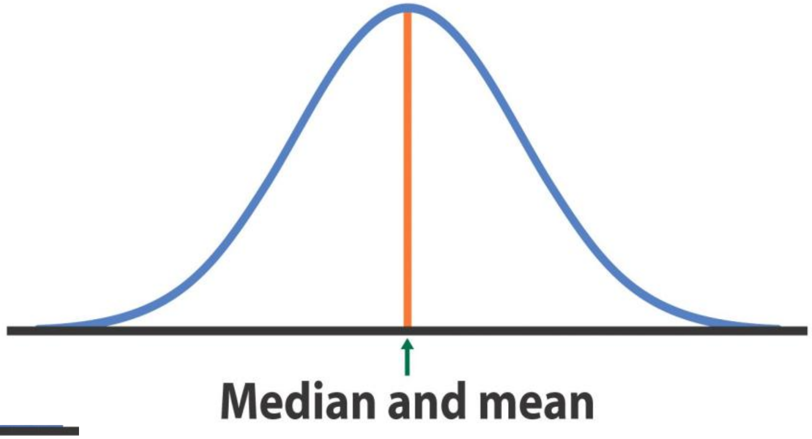
Some are well known mathematically and others aren’t.

Median and mean of a density curve

The **median** of a density curve is the equal-areas point: the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if it were made of solid material.



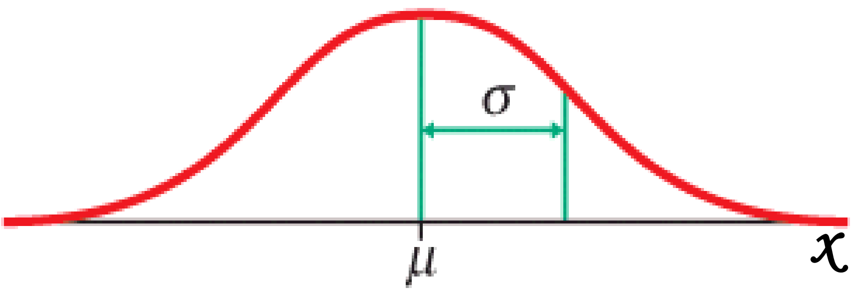


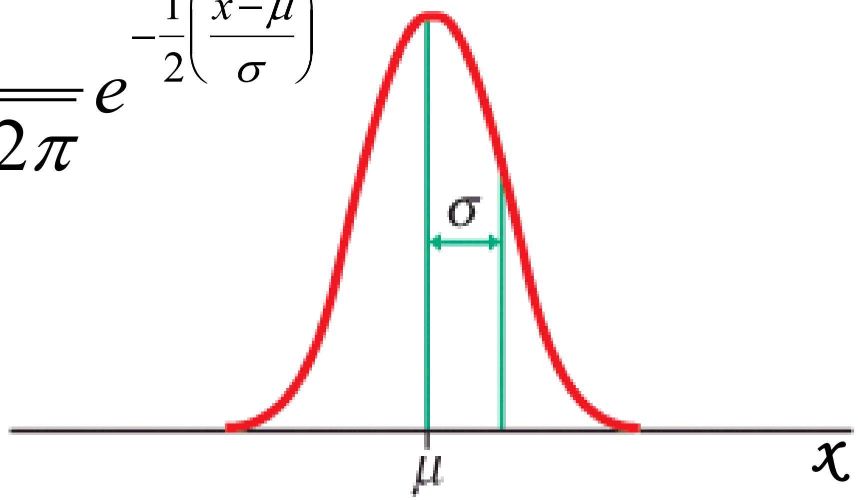
The median and mean are the same for a symmetric density curve.
The mean of a skewed curve is pulled in the direction of the long tail.

Normal distributions

Normal – or Gaussian – distributions are a family of symmetrical, bell-shaped density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*) : N(μ,σ).

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

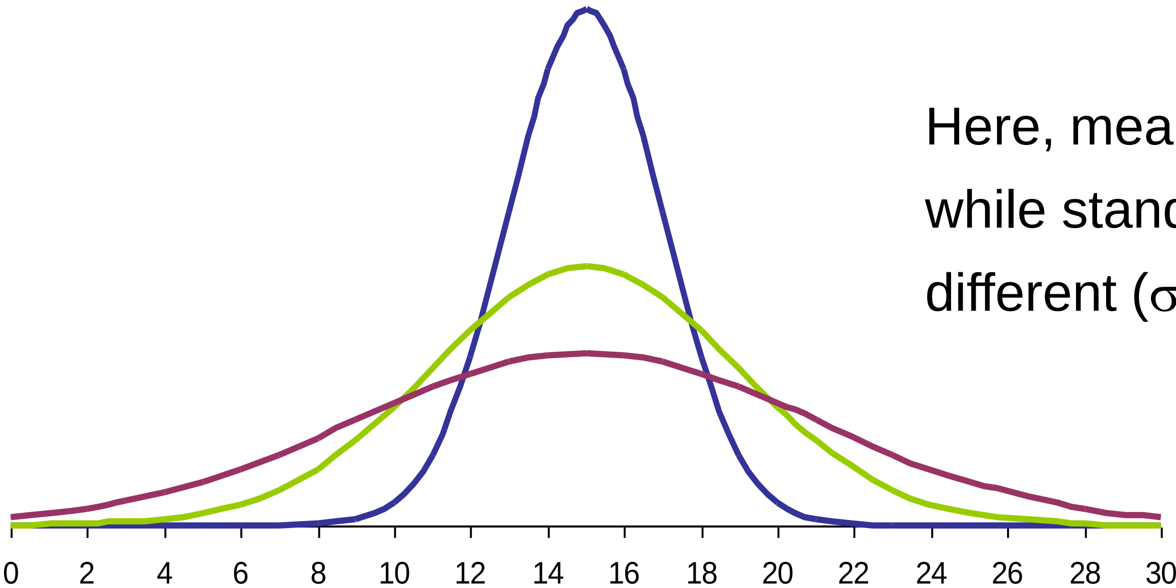




e = 2.71828… The base of the natural logarithm

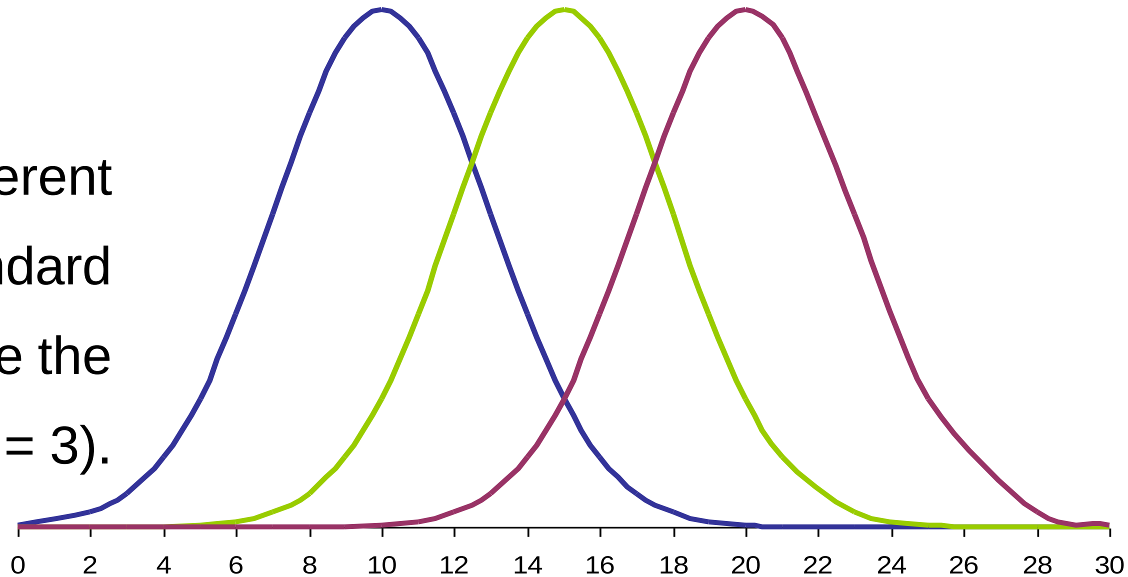
π = pi = 3.14159…

A family of density curves



Here, means are the same (μ = 15) while standard deviations are different (σ = 2, 4, and 6).

Here, means are different (μ = 10, 15, and 20) while standard deviations are the same (σ = 3).

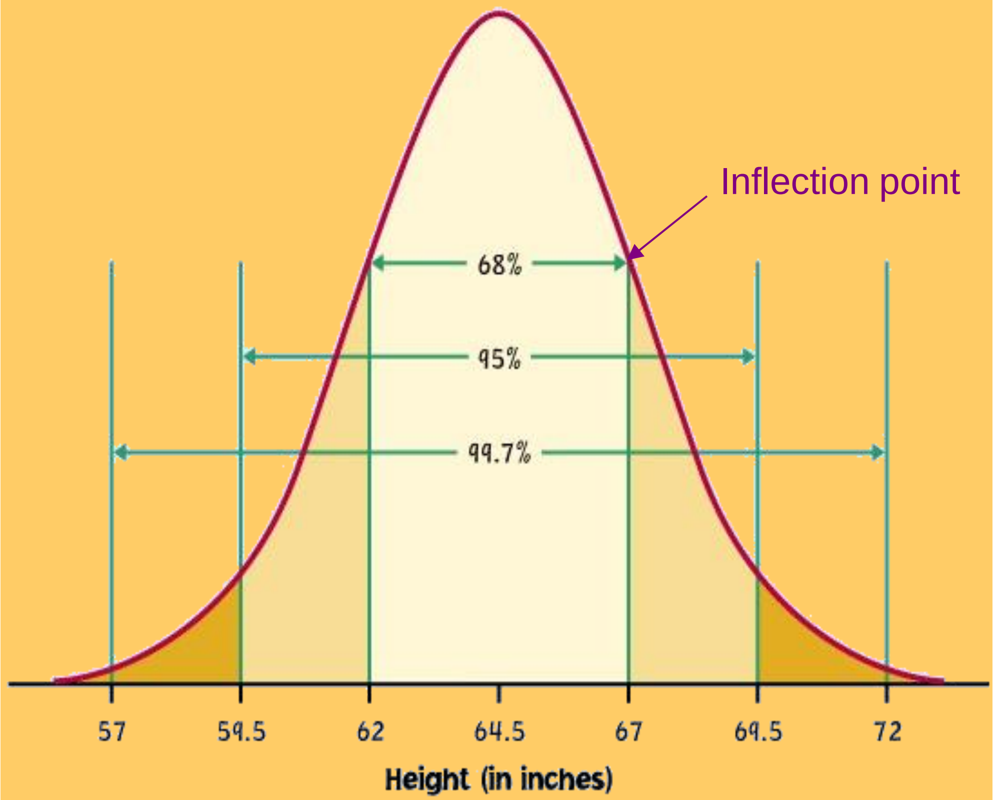


The 68-95-99.7% Rule for Normal Distributions

- About 68% of all observations are within 1 standard deviation (σ) of the mean (μ).

- About 95% of all observations are within 2 σ of the mean μ.

- Almost all (99.7%) observations are within 3 σ of the mean.



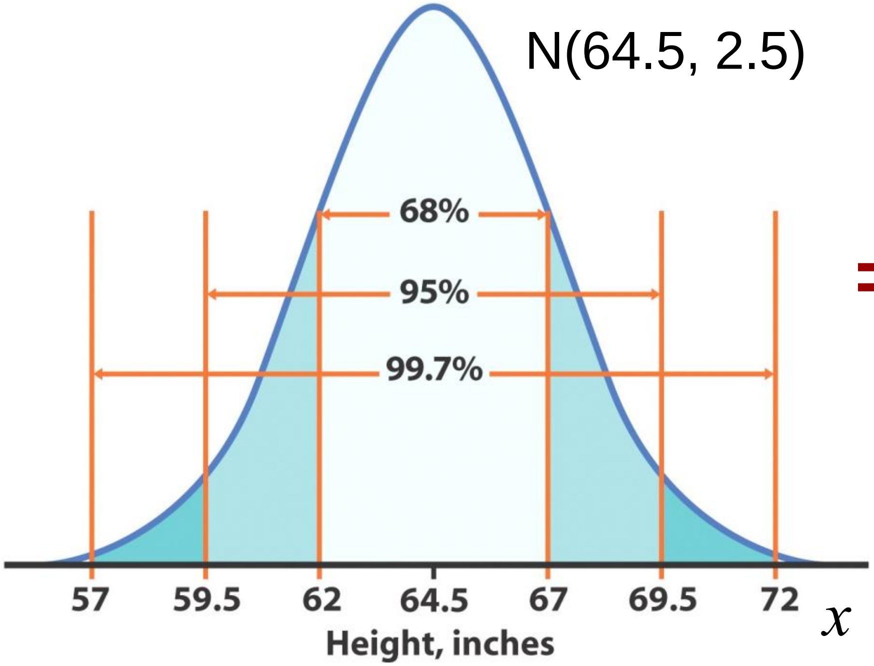
mean μ = 64.5 standard deviation σ = 2.5

$N(\mu, \sigma)$ = N(64.5, 2.5)

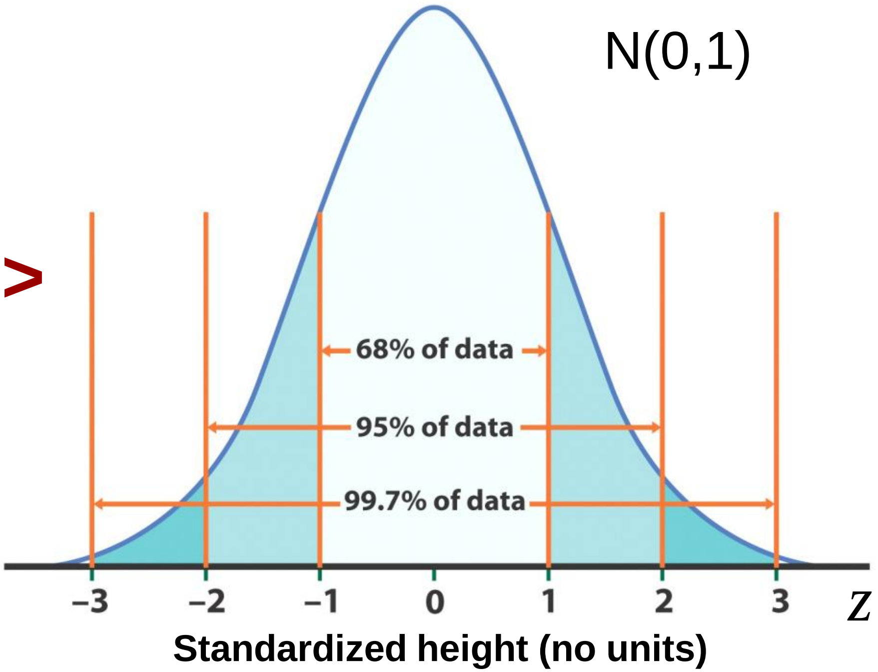
<u>Reminder</u>: μ (mu) is the mean of the idealized curve, while $\bar{x}$ is the mean of a sample.
σ (sigma) is the standard deviation of the idealized curve, while s is the s.d. of a sample.

The standard Normal distribution

Because all Normal distributions share the same properties, we can **standardize** our data to transform any Normal curve N(μ,σ) into the standard Normal curve N(0,1).



=>



For each x we calculate a new value, z (called a **z-score**).

Standardizing: calculating *z*-scores

A ***z*-score** measures the number of standard deviations that a data value *x* is from the mean μ.

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

When x is 1 standard deviation larger than the mean, then z = 1.

$$\text{for } x = \mu + \sigma, \quad z = \frac{\mu + \sigma - \mu}{\sigma} = \frac{\sigma}{\sigma} = 1$$

When x is 2 standard deviations larger than the mean, then z = 2.

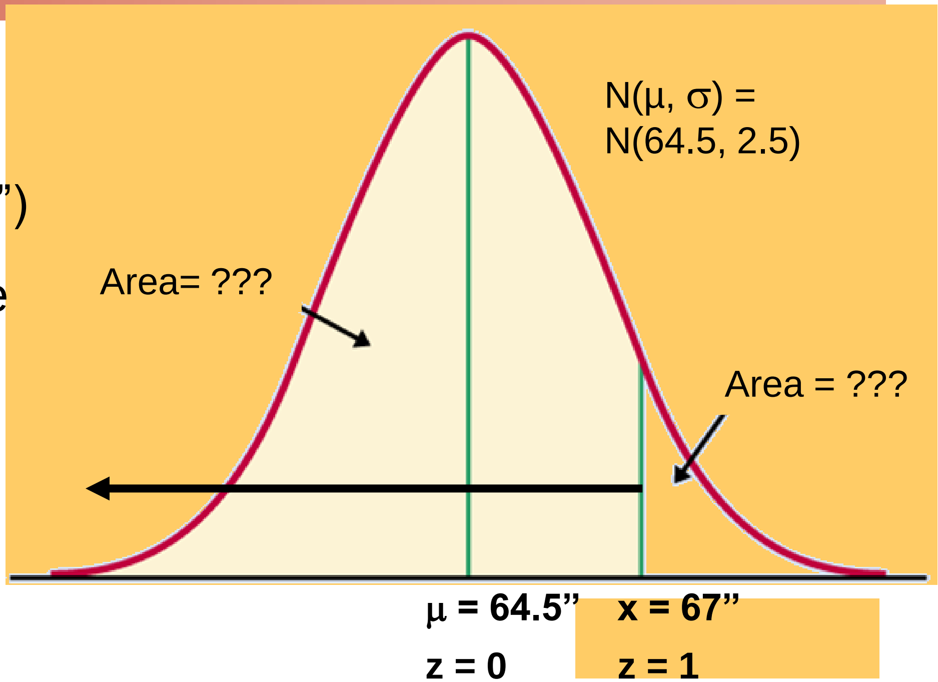
$$\text{for } x = \mu + 2\sigma, \quad z = \frac{\mu + 2\sigma - \mu}{\sigma} = \frac{2\sigma}{\sigma} = 2$$

When x is larger than the mean, z is positive.

When x is smaller than the mean, z is negative.

Ex. Women heights

Women's heights follow the N(64.5",2.5") distribution. What percent of women are shorter than 67 inches tall (that's 5'6")?



mean μ = 64.5"
standard deviation σ = 2.5"
x (height) = 67"

We calculate *z*, the standardized value of *x*:

$$z = \frac{(x-\mu)}{\sigma}, \quad z = \frac{(67-64.5)}{2.5} = \frac{2.5}{2.5} = 1 => 1 \text{ stand. dev. from mean}$$

Because of the 68-95-99.7 rule, we can conclude that the percent of women shorter than 67" should be, approximately, .68 + half of (1 - .68) = .84 or 84%.

Using the standard Normal table

Table A gives the area under the standard Normal curve to the left of any z value.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
[illegible]	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	[illegible]	.0146	.0143
−2.0	[illegible]	.0222	.0217	.0212	.0207	.0202	.0197	.0192	[illegible]	.0183
−1.9	[illegible]	.0281	.0274	.0268	.0262	.0256	.0250	.0244	[illegible]	.0233
−1.8	[illegible]	.0351	.0344	.0336	.0329	.0322	.0314	.0307	[illegible]	[illegible]
[illegible]	[illegible]	.0436	.0427	.0418	.0409	.0401	.0392	.0384	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	.0526	.0516	.0505	.0495	.0485	.0475	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	.0643	.0630	.0618	.0606	.0594	.0582	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	.0778	.0764	.0749	.0735	.0721	.0708	[illegible]	[illegible]

(…)

.0082 is the area under N(0,1) left of z = -2.40

.0080 is the area under N(0,1) left of z = -2.41

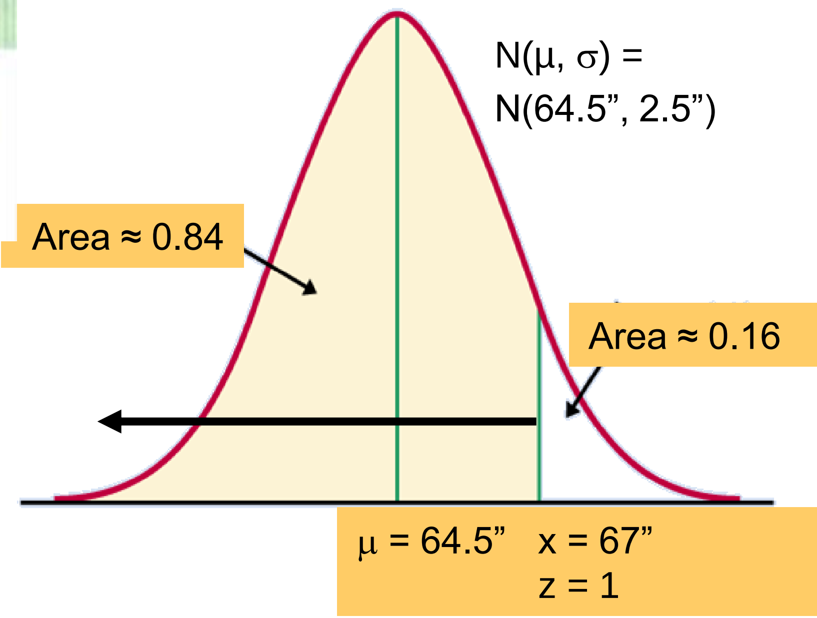
0.0069 is the area under N(0,1) left of z = -2.46

Percent of women shorter than 67”

TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389
0.7	.7580	.7611	.7642	.7673	.7704
0.8	.7881	.7910	.7939	.7967	.7995
0.9	.8159	.8186	.8212	.8238	.8264
1.0	.8413	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729
1.2	.8849	.8869	.8888	.8907	.8925
1.3	.9032	.9049	.9066	.9082	.9099

For $z = 1.00$, the area under the standard Normal curve to the left of z is 0.8413.



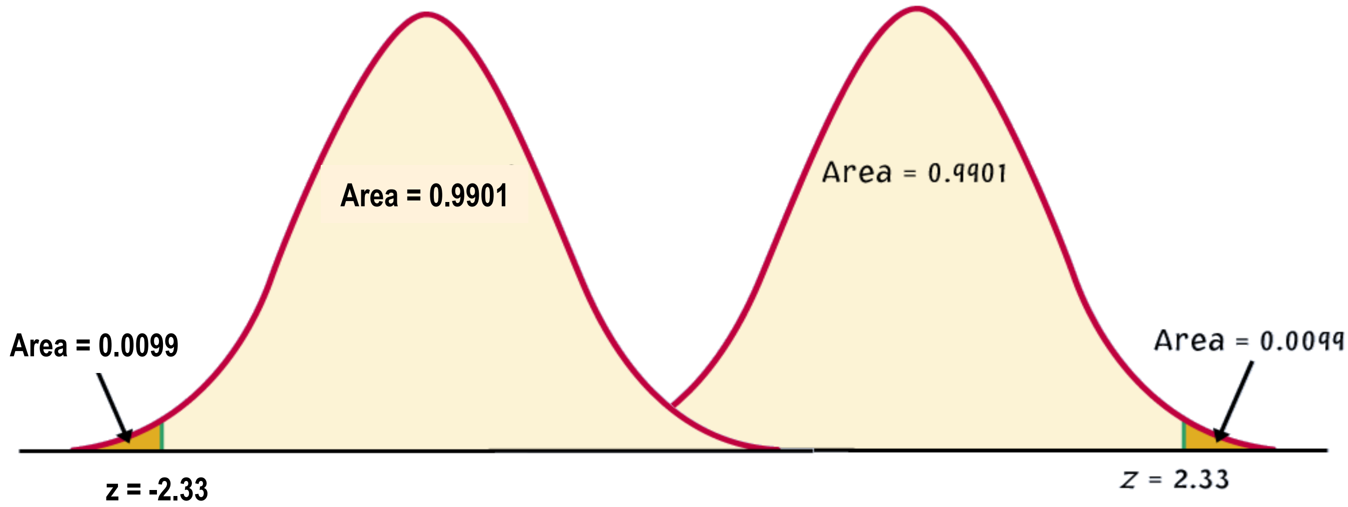
Conclusion:

84.13% of women are shorter than 67”.

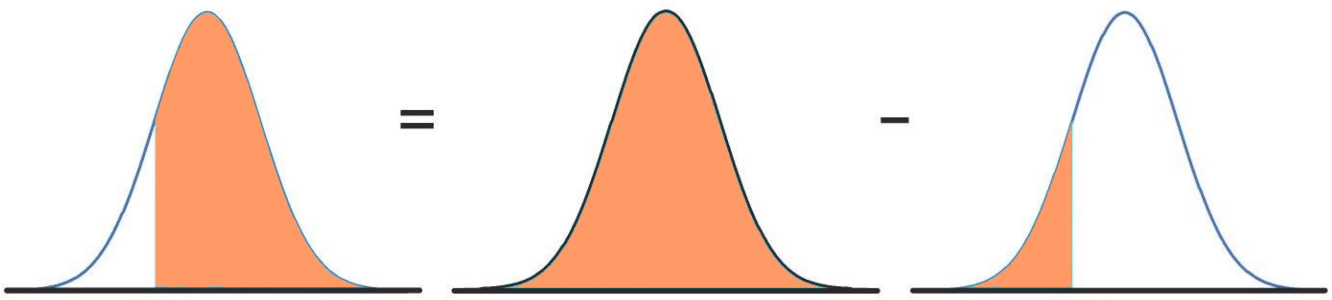
By subtraction, 1 - 0.8413, or 15.87% of women are taller than 67".

Tips on using Table A

Because the Normal distribution is symmetrical, there are 2 ways that you can calculate the area under the standard Normal curve to the right of a z value.



area right of z = area left of $-z$



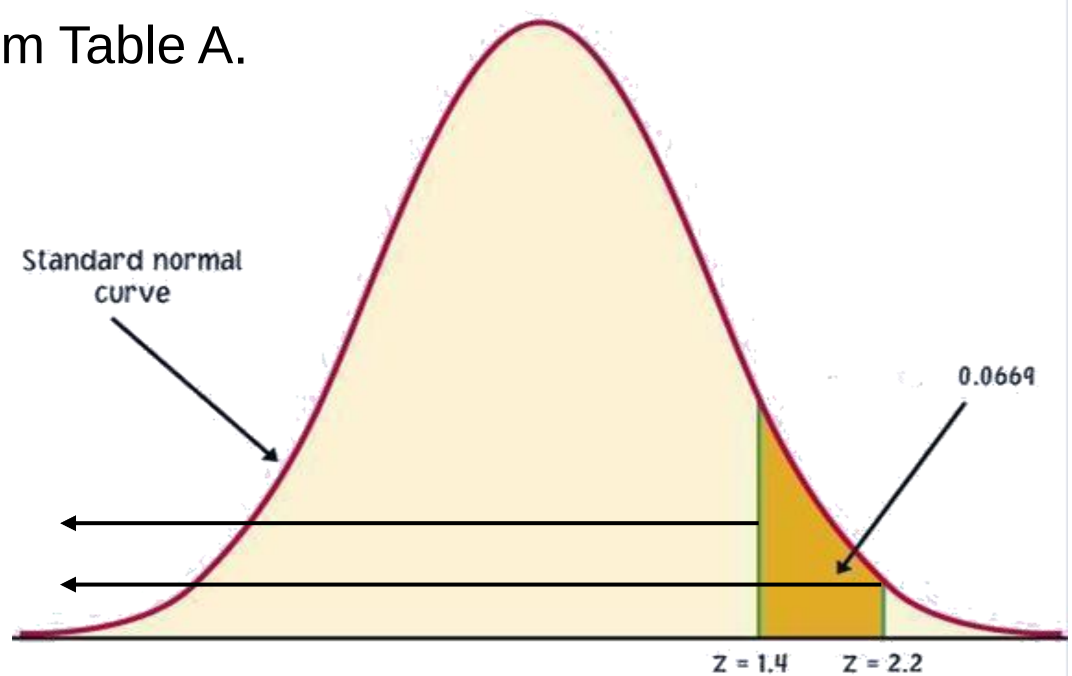
area right of z = 1 - area left of z

Tips on using Table A

To calculate the area between 2 *z*- values, first get the area under N(0,1) to the left for each *z*-value from Table A.

Then subtract the smaller area from the larger area.

A common mistake made by students is to subtract both z values. But the Normal curve is not uniform.



area between z_1 and z_2 =
area left of z_1 – area left of z_2

➔ **The area under N(0,1) for a single value of *z* is zero.**

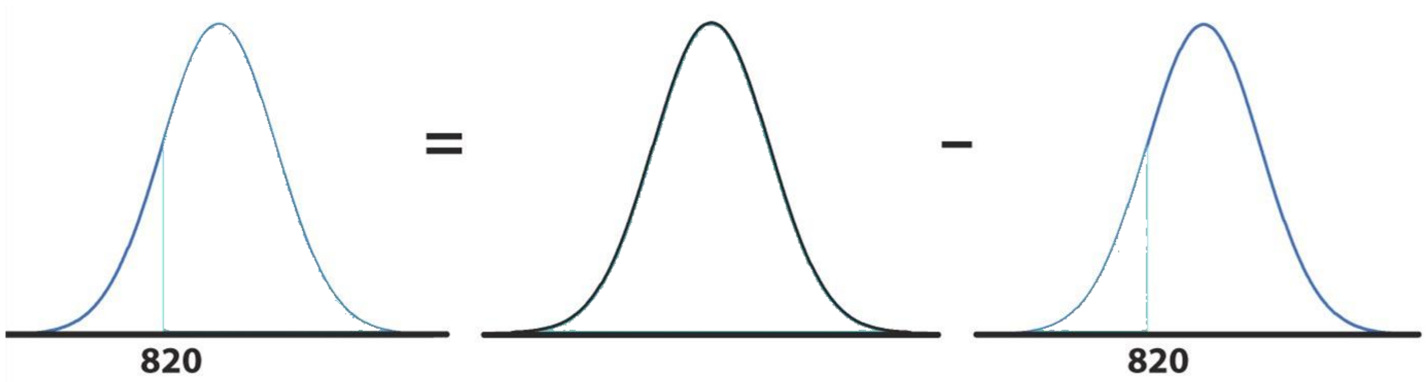
(Try calculating the area to the left of z minus that same area!)

The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined math and verbal SAT exam to compete in their first college year. The SAT scores of 2003 were approximately normal with mean 1026 and standard deviation 209.

What proportion of all students would be NCAA qualifiers (SAT ≥ 820)?

$$x = 820$$
$$\mu = 1026$$
$$\sigma = 209$$
$$z = \frac{(x - \mu)}{\sigma}$$
$$z = \frac{(820 - 1026)}{209}$$
$$z = \frac{-206}{209} \approx -0.99$$

Table A : area under N(0,1) to the left of z = -0.99 is 0.1611 or approx. 16%.



area right of 820 = total area - area left of 820

= 1 - 0.1611

≈ 84%

Note: The actual data may contain students who scored exactly 820 on the SAT. However, the proportion of scores exactly equal to 820 is 0 for a normal distribution is a consequence of the idealized smoothing of density curves.

The NCAA defines a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, with a combined SAT score of at least 720.

What proportion of all students who take the SAT would be partial qualifiers? That is, what proportion have scores between 720 and 820?

$$x = 720$$

$$\mu = 1026$$

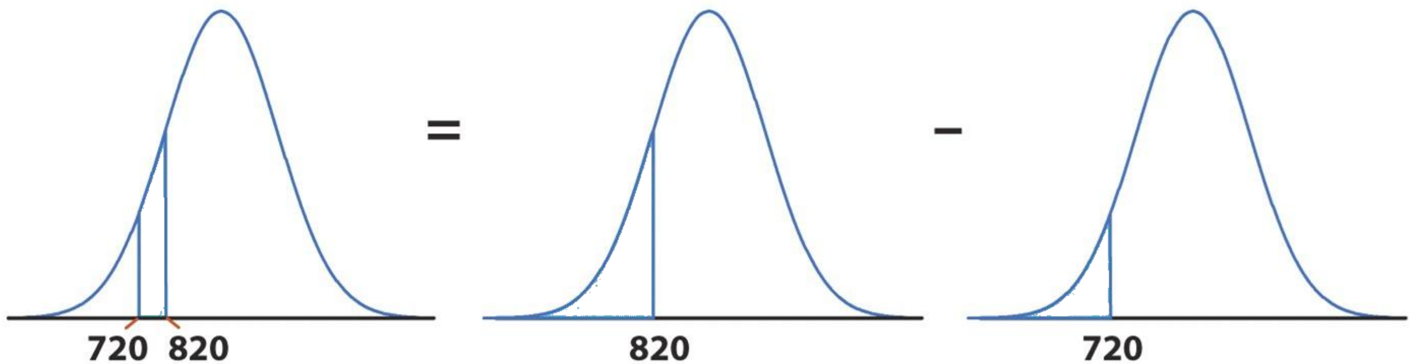
$$\sigma = 209$$

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

$$z = \frac{(720-1026)}{209}$$

$$z = \frac{-306}{209} \approx -1.46$$

Table A : area under N(0,1) to the left of z = -1.46 is 0.0721 or approx. 7%.



area between 720 and 820 = area left of 820 - area left of 720

= 0.1611 - 0.0721

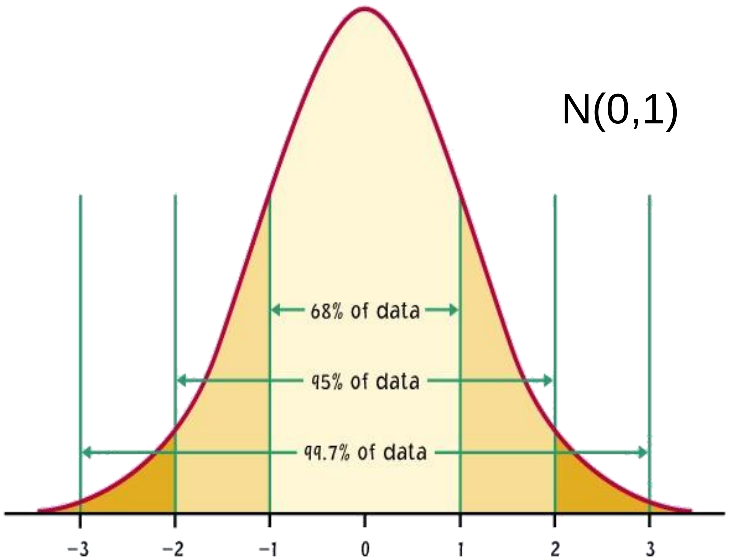
≈ 9%

About 9% of all students who take the SAT have scores between 720 and 820.

The cool thing about working with normally distributed data is that we can manipulate it, and then find answers to questions that involve comparing seemingly non-comparable distributions.

We do this by “standardizing” the data. All this involves is changing the scale so that the mean now = 0 and the standard deviation =1. If you do this to different distributions it makes them comparable.

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

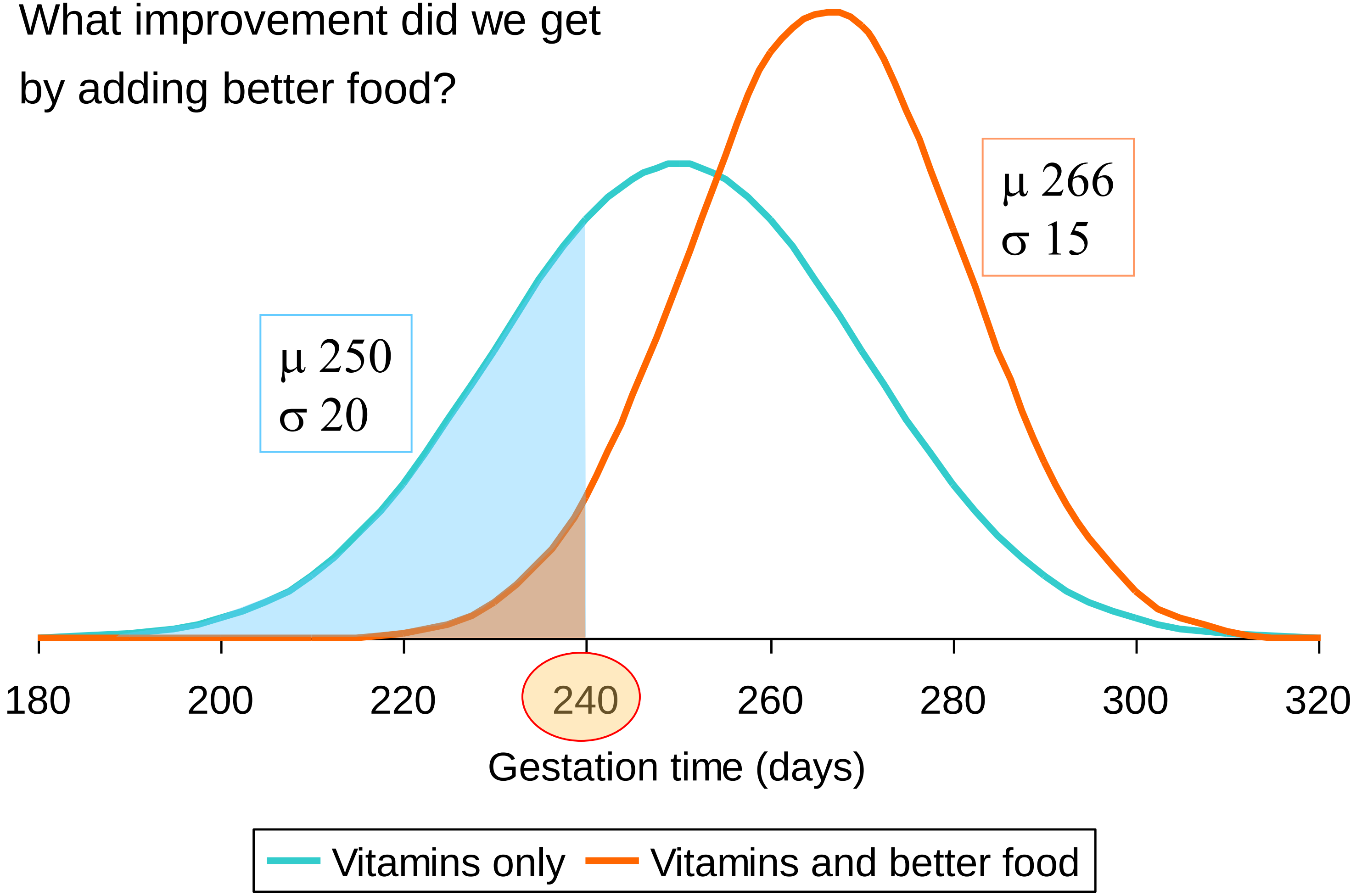


Ex. Gestation time in malnourished mothers

What is the effect of better maternal care on gestation time and preemies?

The goal is to obtain pregnancies 240 days (8 months) or longer.

What improvement did we get by adding better food?



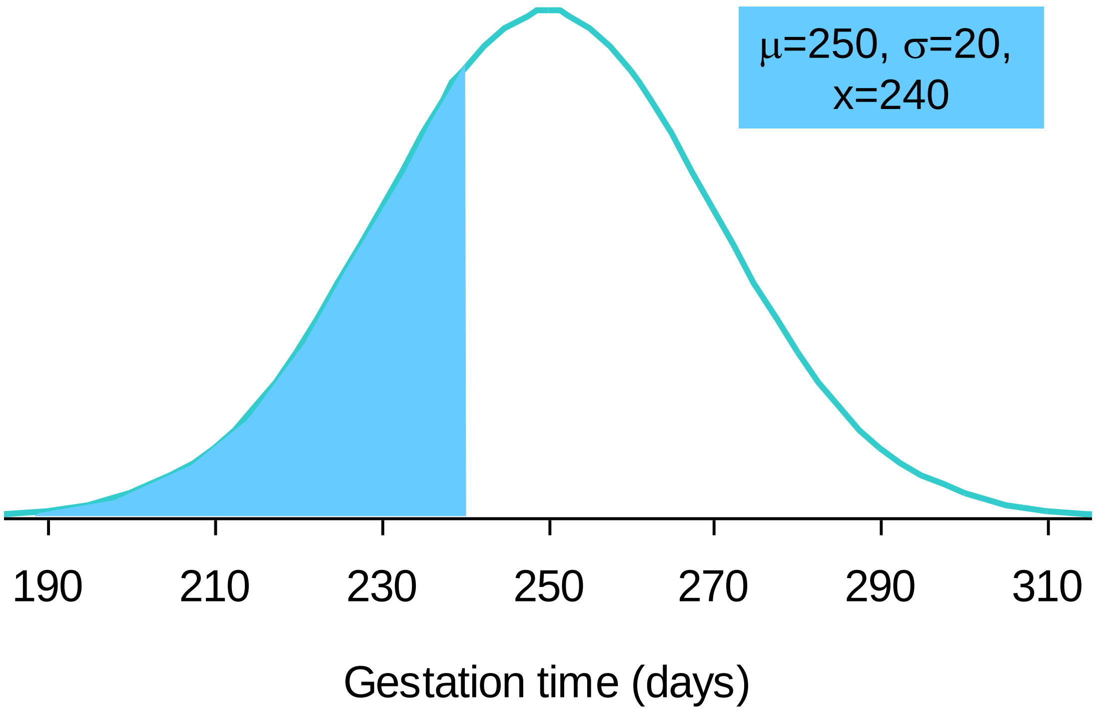
Under **each treatment**, what percent of mothers failed to carry their babies at least 240 days?

Vitamins Only

$$x = 240$$
$$\mu = 250$$
$$\sigma = 20$$
$$z = \frac{(x - \mu)}{\sigma}$$
$$z = \frac{(240 - 250)}{20}$$
$$z = \frac{-10}{20} = -0.5$$

(half a standard deviation)

Table A : area under N(0,1) to the left of z = -0.5 is 0.3085.



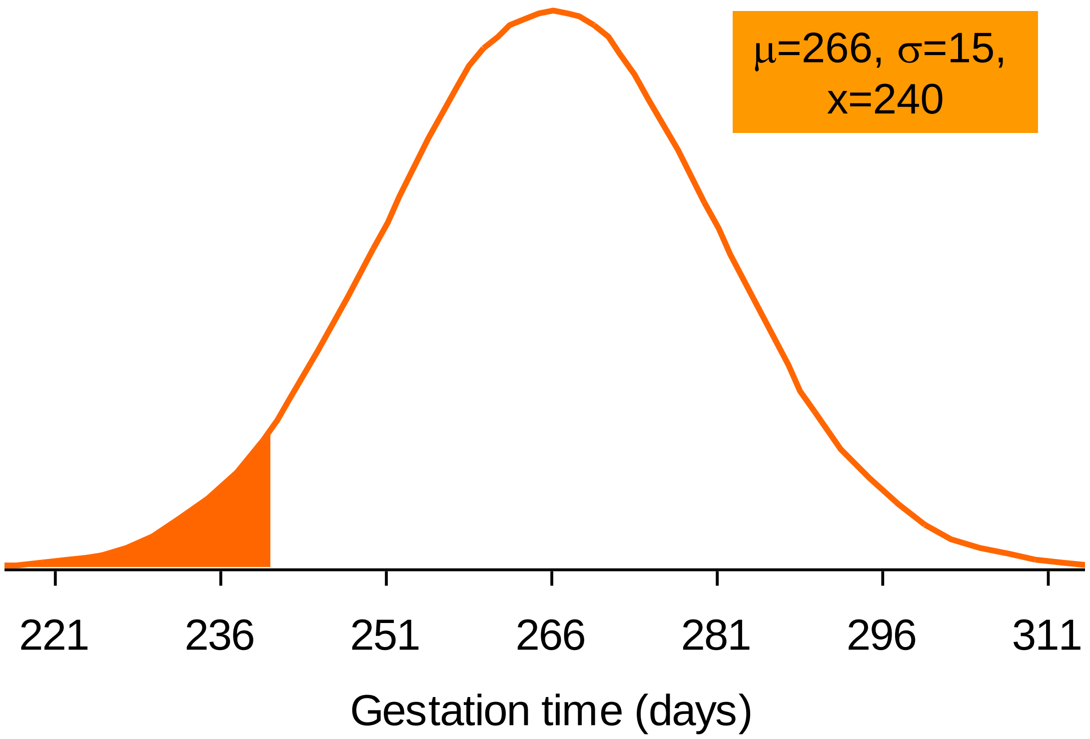
<u>Vitamins only</u>: 30.85% of women would be expected to have gestation times shorter than 240 days.

Vitamins and better food

$$x = 240$$
$$\mu = 266$$
$$\sigma = 15$$
$$z = \frac{(x - \mu)}{\sigma}$$
$$z = \frac{(240 - 266)}{15}$$
$$z = \frac{-26}{15} = -1.73$$

(almost 2 sd from mean)

Table A : area under N(0,1) to the left of z = -1.73 is 0.0418.



Vitamins and better food: 4.18% of women would be expected to have gestation times shorter than 240 days.

Compared to vitamin supplements alone, vitamins and better food resulted in a much smaller percentage of women with pregnancy terms below 8 months (4% vs. 31%).

Inverse normal calculations

We may also want to find the observed range of values that correspond to a given proportion/ area under the curve.

For that, we use Table A backward:

- we first find the desired area/ proportion in the body of the table,
- we then read the corresponding *z-value* from the left column and top row.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

For an area to the left of 1.25 % (0.0125), the *z*-value is -2.24

Vitamins and better food

How long are the longest 75% of pregnancies when mothers with malnutrition are given vitamins and better food?

$\mu = 266$

$\sigma = 15$

$\text{upper area} = 75\%$

$\text{lower area} = 25\%$

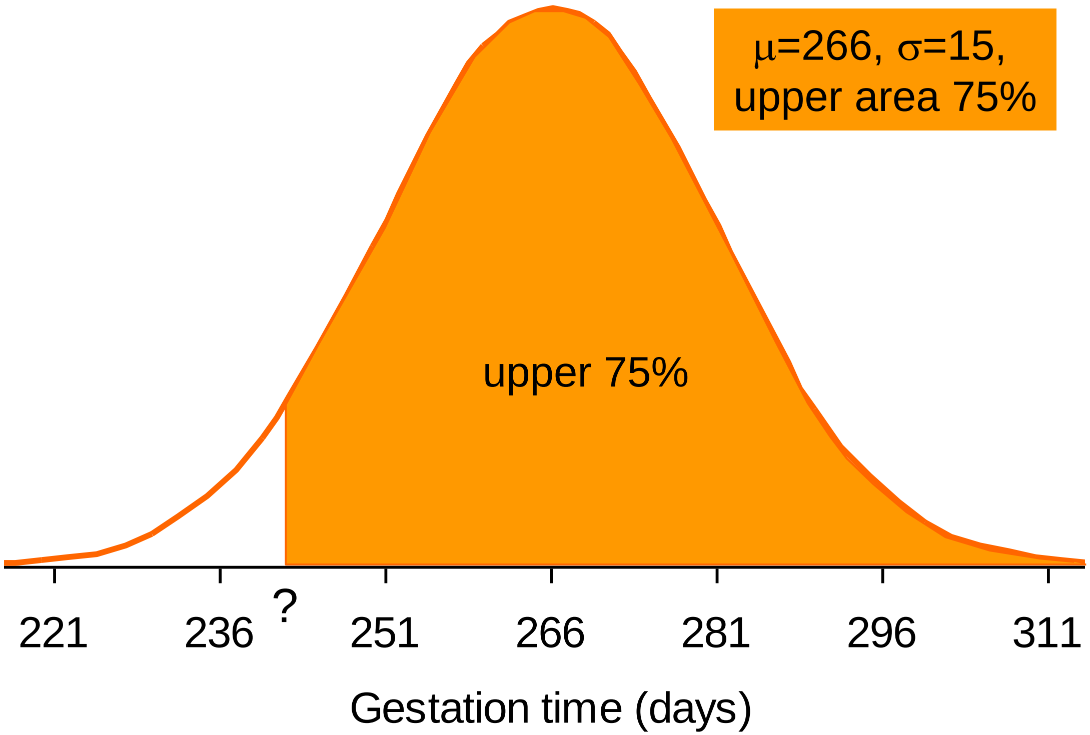
$x = ?$

Table A : z value for the lower area 25% under N(0,1) is about -0.67.

$$z = \frac{(x-\mu)}{\sigma} \Leftrightarrow x = \mu + (z * \sigma)$$

$$x = 266 + (-0.67 * 15)$$

$$x = 255.95 \approx 256$$



Remember that Table A gives the area to the left of *z*. Thus, we need to search for the lower 25% in Table A in order to get *z*.

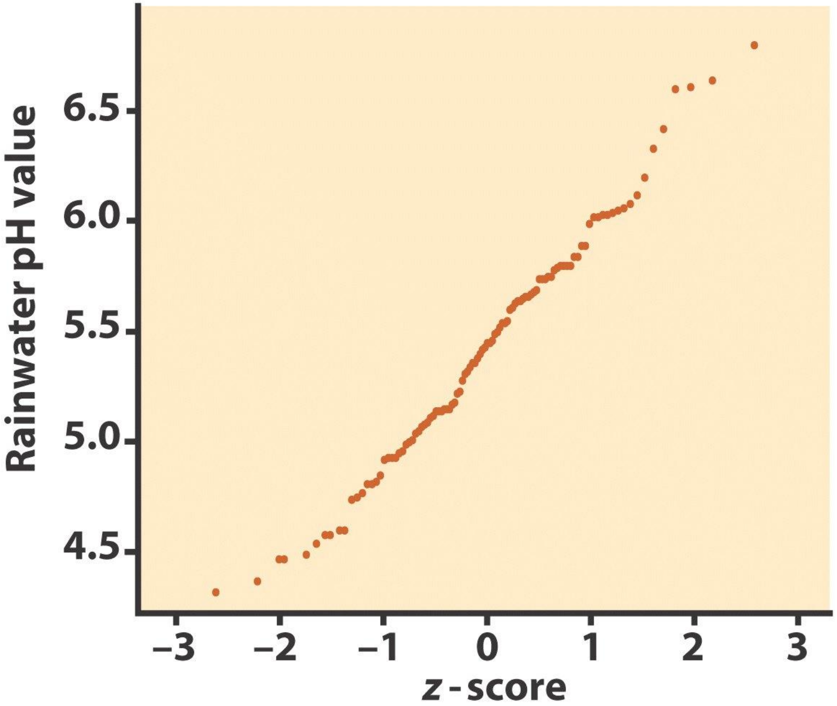
➔ The 75% longest pregnancies in this group are about 256 days or longer.

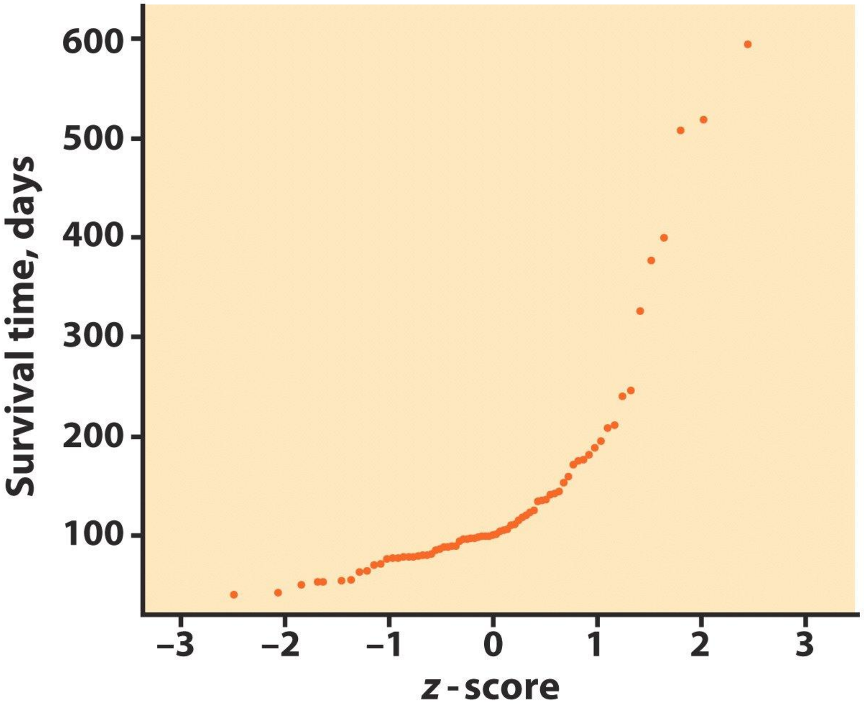
Normal quantile plots

One way to assess if a distribution is indeed approximately normal is to plot the data on a **normal quantile plot**.

The data points are ranked and the percentile ranks are converted to *z*-scores with Table A. The *z*-scores are then used for the x axis against which the data are plotted on the y axis of the normal quantile plot.

- If the distribution is indeed normal the plot will show a straight line, indicating a good match between the data and a normal distribution.
- Systematic deviations from a straight line indicate a non-normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.





Good fit to a straight line: the distribution of rainwater pH values is close to normal.

Curved pattern: the data are not normally distributed. Instead, it shows a right skew: a few individuals have particularly long survival times.

Normal quantile plots are complex to do by hand, but they are standard features in most statistical software.

Alternate Slide

The following slide offers alternate software output data and examples for this presentation.

Software output for summary statistics:

JMP- From Menu: ➔ Analyze/Distribution/ Heights ➔ Y, Columns/OK

Give common statistics of your sample data.

Womens_Height_inches

Statistic	Result
N	14
Minimum	59
First Quartile	61
Median	63
Third Quartile	65.5000
Maximum	68
Mean	63.3571
Standard Deviation	2.5603

Distributions

Women's Height (inches)

Quantiles

100.0%	maximum	68
99.5%		68
97.5%		68
90.0%		67.5
75.0%	quartile	65.25
50.0%	median	63
25.0%	quartile	61.75
10.0%		59.5
2.5%		59
0.5%		59
0.0%	minimum	59

Moments

Mean	63.357143
Std Dev	2.5602627
Std Err Mean	0.684259
Upper 95% Mean	64.835395
Lower 95% Mean	61.878891
N	14

🡐 CrunchIt!

Stat ➔ Summary Statistics ➔ Columns
Women’s Heights (inches)
n, Min, Q1, Median, Q3, Max, Mean, Std. Dev.
OK

(Control Click or Click Shift to select several statistics.)