Math 140
Introductory Statistics

Professor Silvia Fernández

Chapter 2
Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

Visualizing Distributions

- Recall the definition:

  The values of a summary statistic (e.g. the average age of the laid-off workers) and how often they occur.

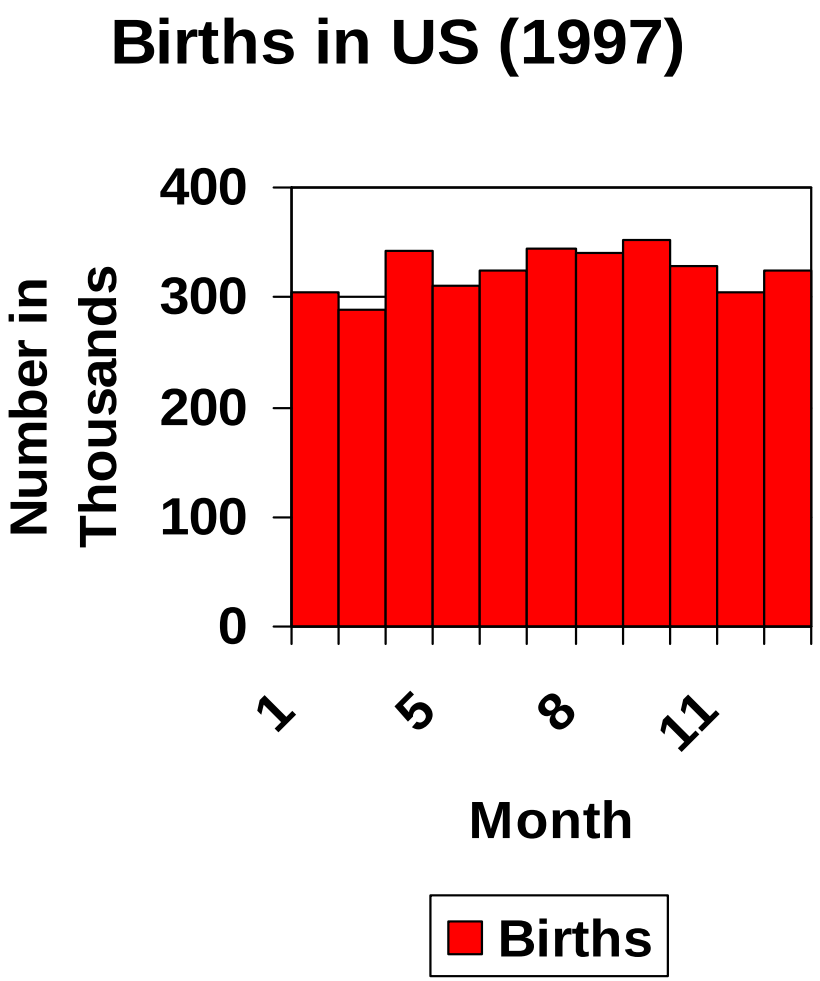
- Four of the most common basic **shapes**:
  - Uniform or Rectangular
  - Normal
  - Skewed
  - Bimodal (Multimodal)

Uniform (or Rectangular) Distribution

- Each outcome occurs roughly the same number of times.
- Examples.
  - Number of U.S. births per month in a particular year (see Page 25)
  - Computer generated random numbers on a particular interval.
  - Number of times a fair die is rolled on a particular number.

Month	Births (in thousands)	Deaths (in thousands)
1	305	218
2	289	191
3	313	198
4	342	189
5	311	195
6	324	182
7	345	192
8	341	178
9	353	176
10	329	193
11	304	189
12	324	192

Uniform (or Rectangular) Distribution



Month	Births (in thousands)	Deaths (in thousands)
1	305	218
2	289	191
3	313	198
4	342	189
5	311	195
6	324	182
7	345	192
8	341	178
9	353	176
10	329	193
11	304	189
12	324	192

Normal Distributions

- These distributions arise from
  - Variations in measurements.

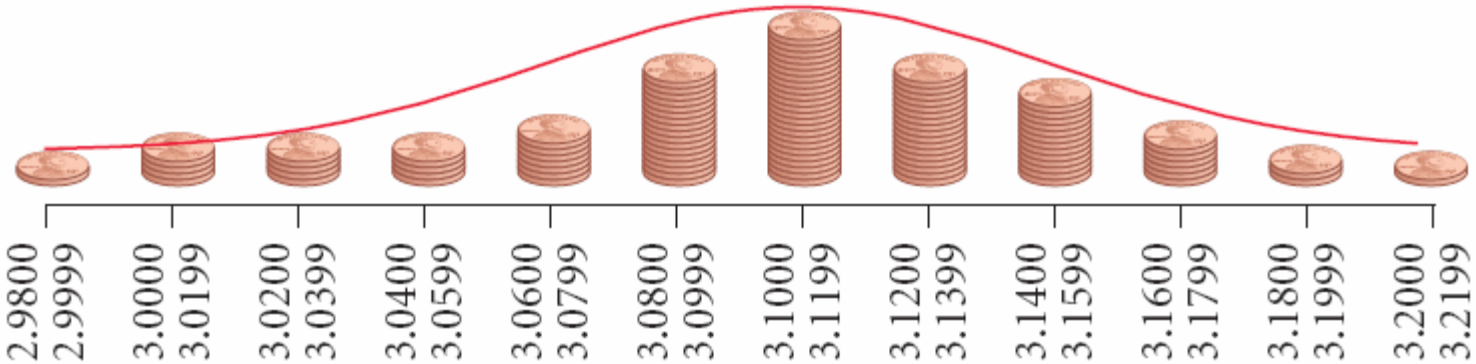
    (e.g. pennies example, see 2.3 page 31)
  - Natural variations in population sizes

    (e.g. weight of a set of people)
  - Variations in averages of random samples.

    (e.g. Average age of 3 workers out of 10, see 1.10 in page 14)

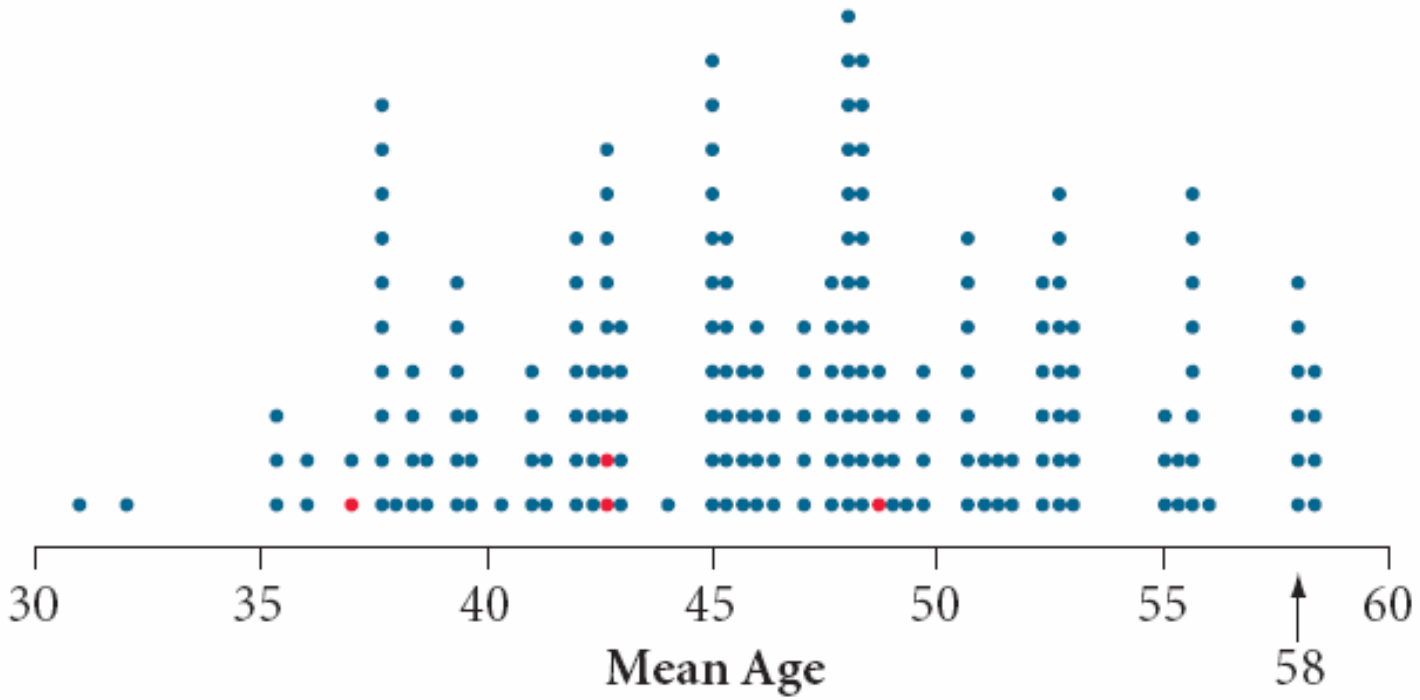
Pennies example

Pennies minted in the United States are supposed to weigh 3.110 g, but a tolerance of 0.130 g is allowed in either direction. Display 2.3 shows a plot of the weights of 100 pennies.



Display 2.3 Weights of pennies. [*Source:* W. J. Youden, *Experimentation and Measurement* (National Science Teachers Association, 1985), p. 108.]

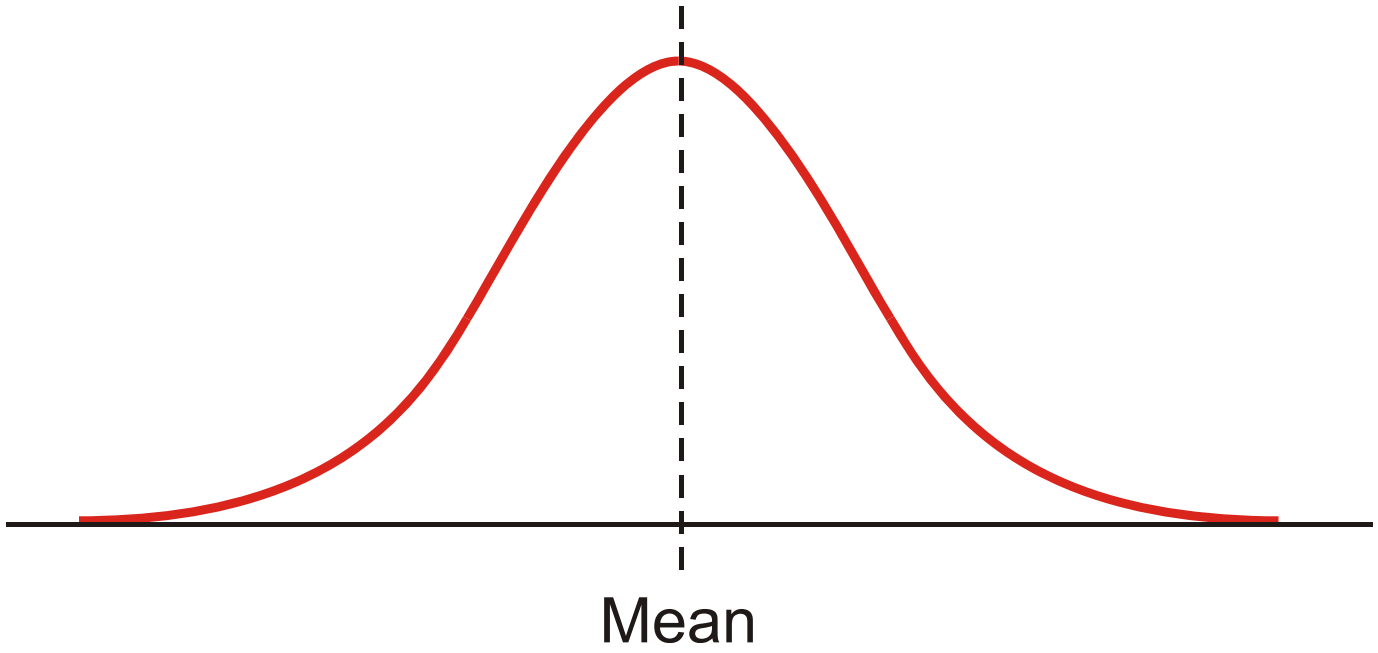
Average age of 3 workers out of 10



Display 1.10 Results of 200 repetitions: the distribution of the average age of the three workers chosen for layoff by chance alone.

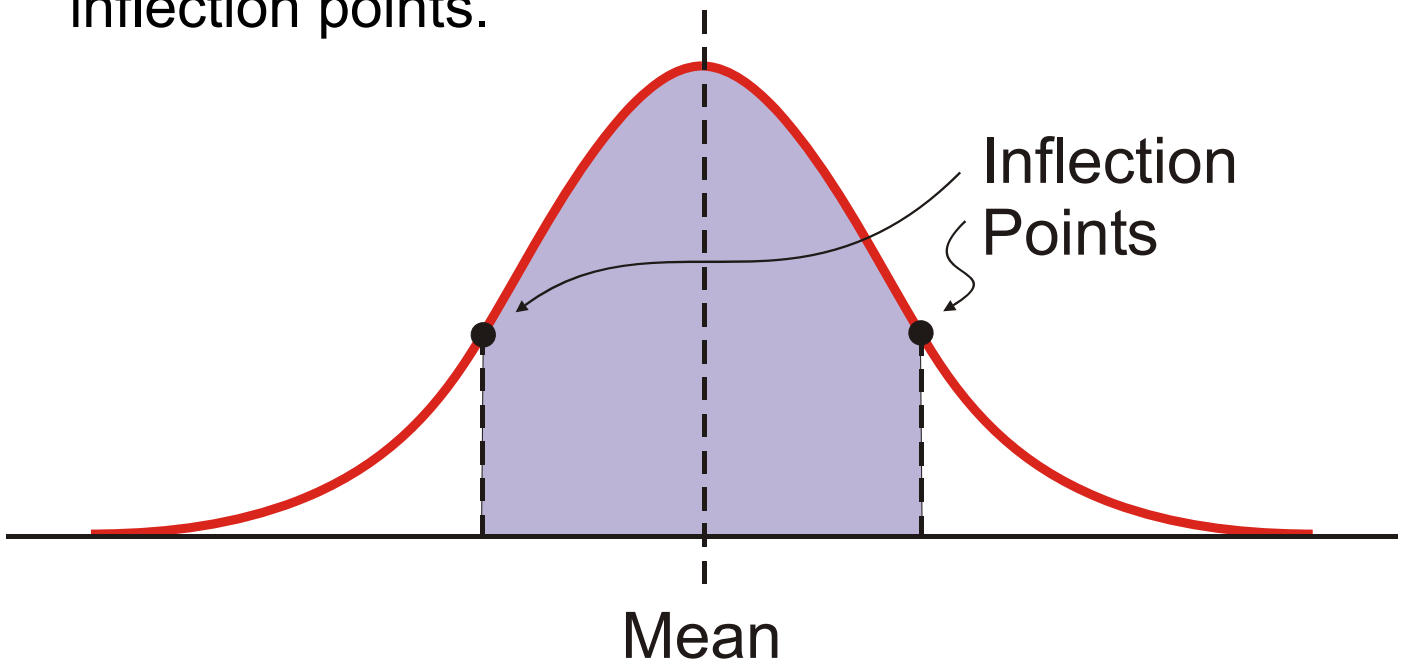
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - Single peak: The x-value of it is called the **mean**.
  - The mean tells us where is the **center** of the distribution.
  - The distribution is symmetric with respect to the mean.



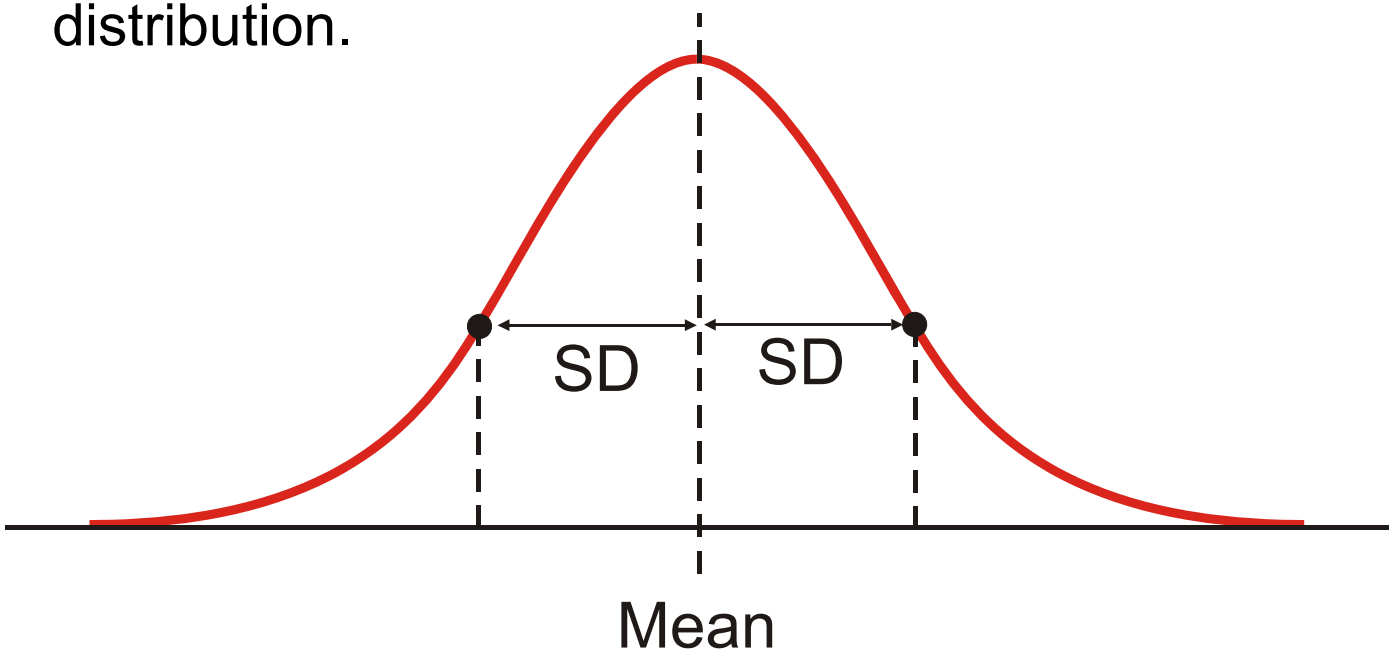
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - Inflection points: Where concavity changes.
  - Roughly 2/3 of the area below the curve is between the inflection points.



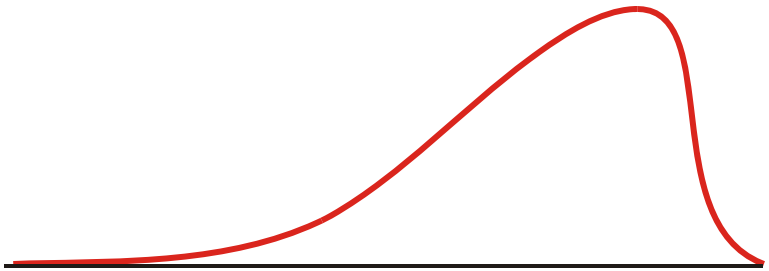
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - The distance between the mean and either of the inflection points is called the **standard deviation** (SD)
  - The standard deviation measures how **spread** is the distribution.

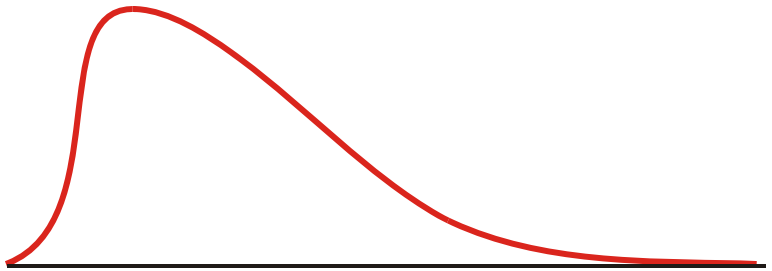


Skewed Distributions

- These are similar to the normal distributions but they are not symmetric. They have values bunching on one end and a long tail stretching in the other direction
- The tail tells you whether the distribution is **skewed left** or **skewed right**.



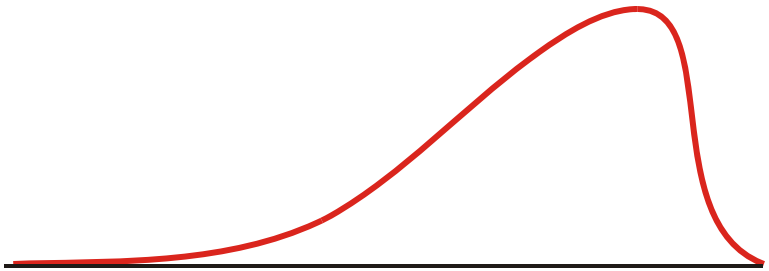
Skewed Left



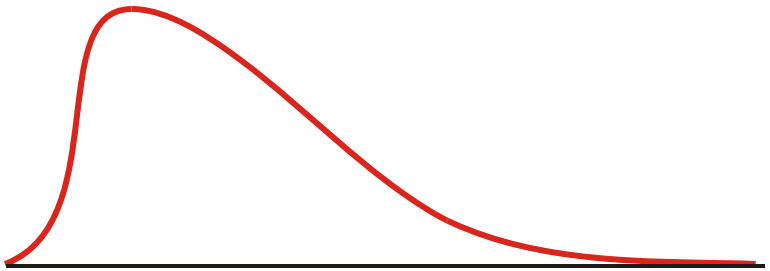
Skewed Right

Skewed Distributions

- Skewed distributions often occur because of a “wall”, that is, values that you cannot go below or above. Like zero for positive measurements, or 100 for percentages.
- To find out about **center** and **spread** it is useful to look at **quartiles**.

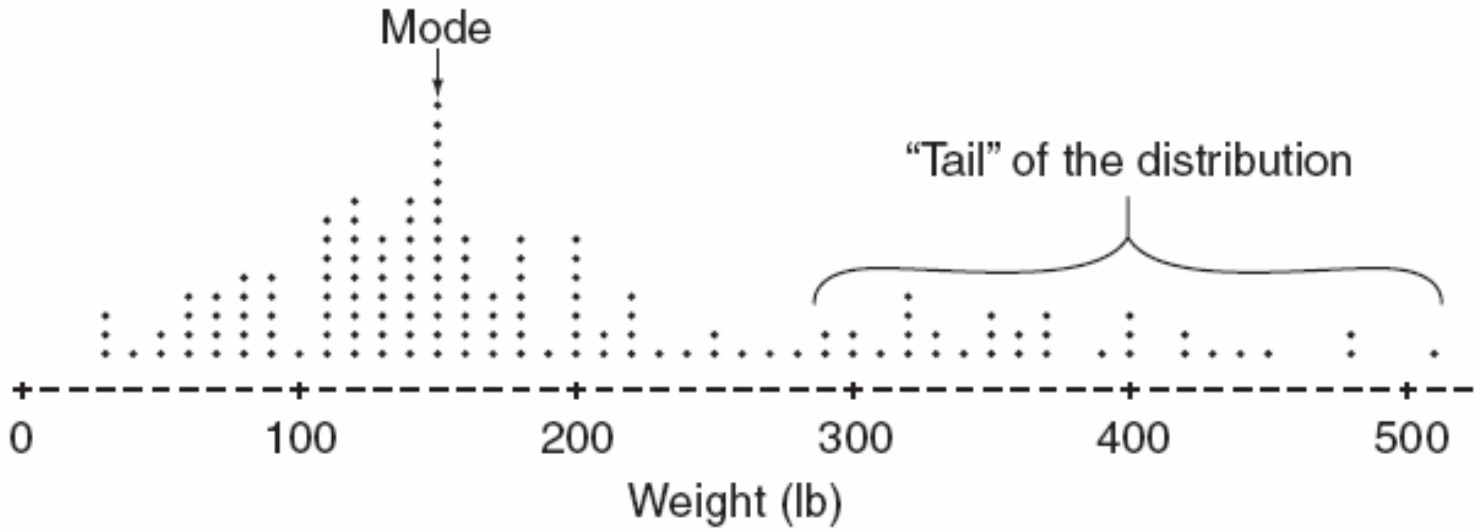


Skewed Left



Skewed Right

Example of a skewed right distribution

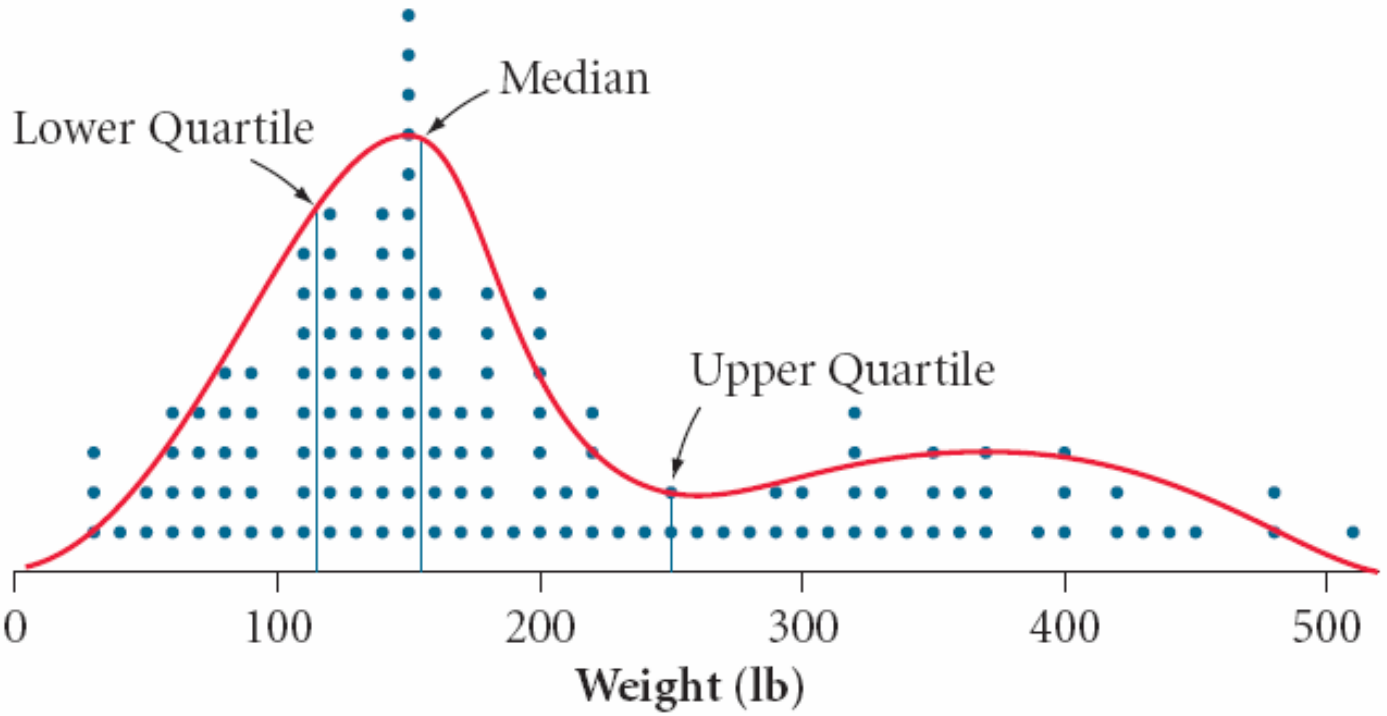


Display 2.6 Weights of bears in pounds. [*Source:* MINITAB data set from *MINITAB Handbook*, 3rd ed.]

Median and Quartiles

- Median: the value of the line dividing the number of values in equal halves. (Or the area under the curve in equal halves.)
- Repeat this process in each of the two halves to find the lower quartile (Q1) and the upper quartile (Q3).
- Q1, the median, and Q3 divide the number of values in **quarters**. The quartiles Q1 and Q3 enclose 50% of the values.

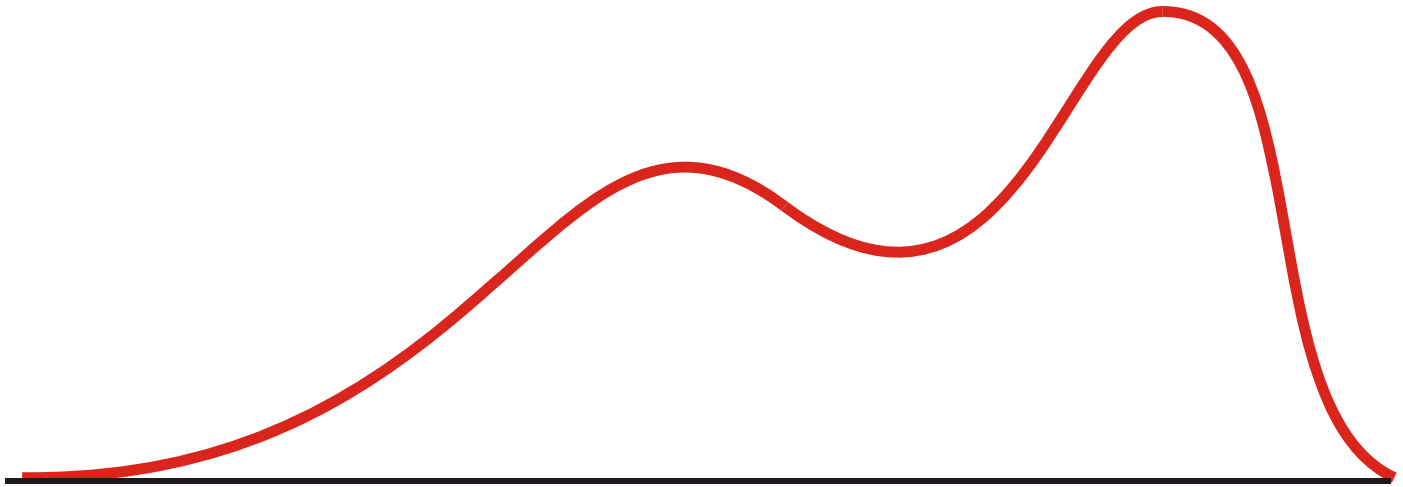
Visualizing Median and Quartiles



Display 2.8 Estimating center and spread for the weights of bears.

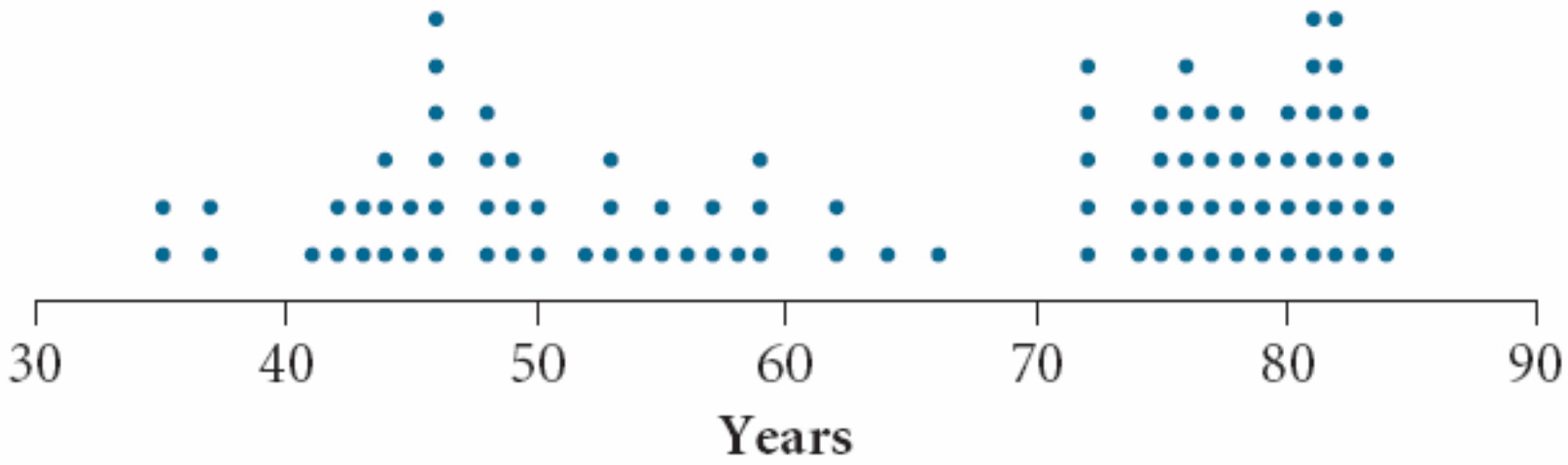
Bimodal Distributions.

- Previous distributions have had only one peak (**unimodal**) but some have two (**bimodal**) or even more (**multimodal**).



Bimodal Distribution

Example of a bimodal distribution



Display 2.9 Life expectancy of females by country on two continents. [*Source:* Population Reference Bureau, *World Population Data Sheet*, 2005.]

Using the calculator (TI-83)

- For more information go to www.keymath.com/x7065.xml and look for the *Calculator Notes* for Chapters 0, 1, and 2.
- You should know how to
  - Generate a list of *n* random integer numbers between *min* and *max*.

  Example: To generate a list of 7 integer numbers between 2 and 10 (inclusive) type

  MATH | PRB | 5.randInt(| Enter | 2, 10, 7) | Enter

Using the calculator (TI-83)

- How to generate a list of *n* random numbers between 0 and 1 (exclusive).

Example: Generate 5 random numbers between 0 and 1.

MATH | PRB | 1.randInt(| Enter | 5) | Enter

- How to store a list of numbers.

Example: Store the previous list of 5 random numbers between 0 and 1 on L_1.

2^{nd} | ANS | → | 2^{nd} | L_1

Using the calculator (TI-83)

Example: Store the list 1,2,3,4,5 to L_1.

STAT 1.Edit Enter

Move to the first row of column L_1 using the arrows.

Type each of the numbers on the list followed by ENTER.

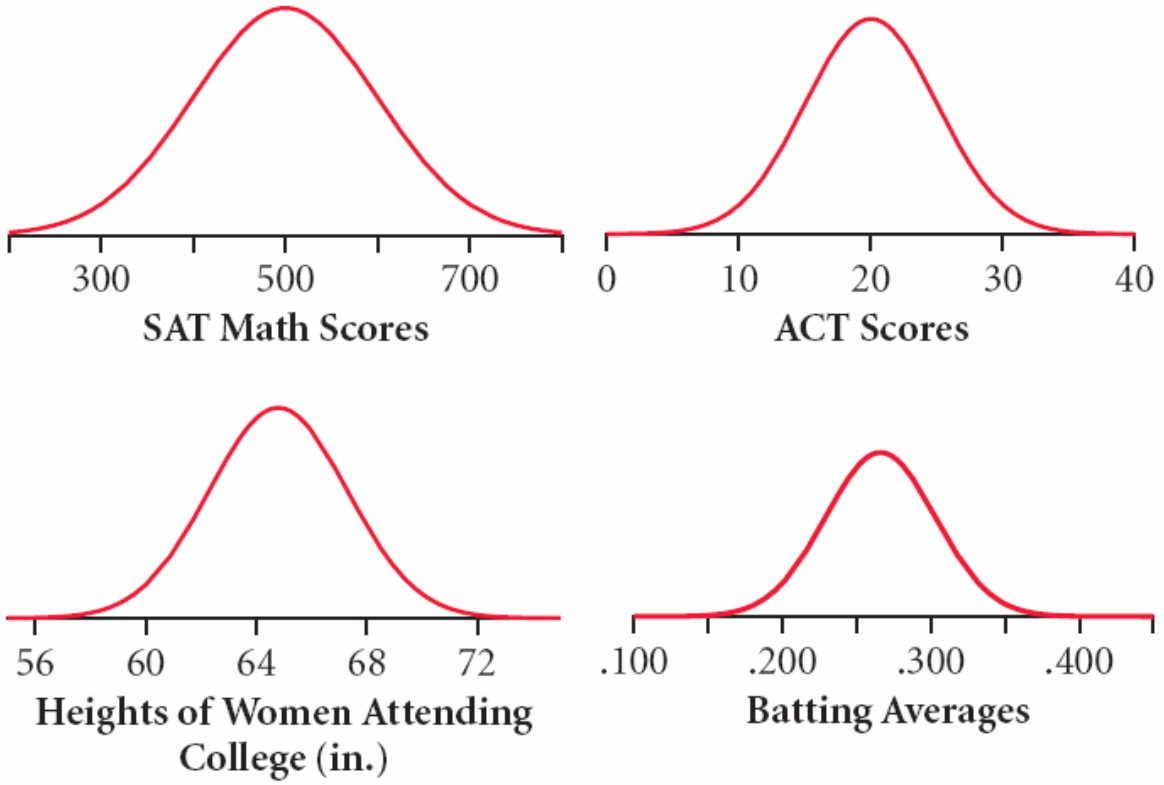
- Compute binomial coefficients.

Example: Compute 10 choose 3.

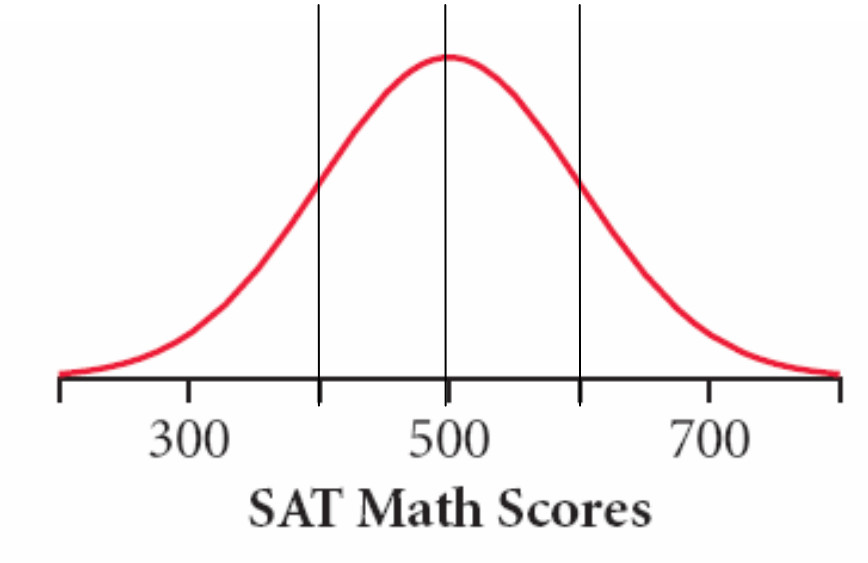
10 MATH PRB nCr Enter 3

Practice

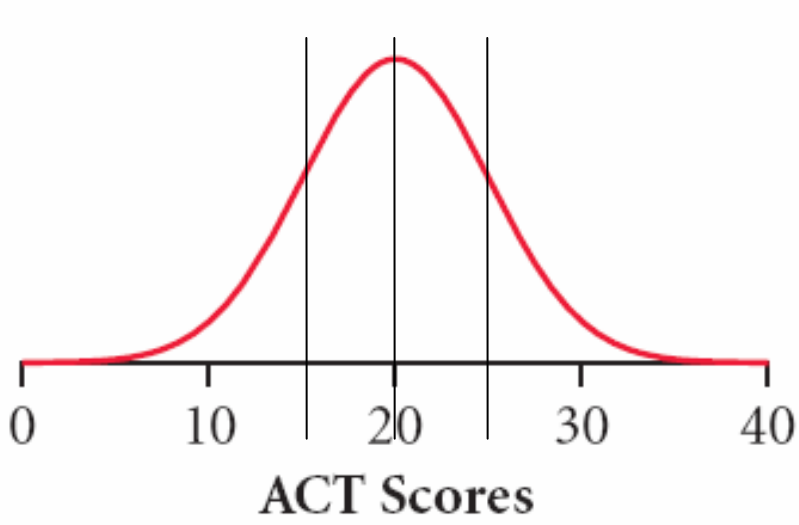
- P3. For each of the normal distributions in below, estimate the mean and standard deviation visually, and use your estimates to write a verbal summary of the form “A typical SAT score is roughly (mean), give or take (*SD*) or so.”



Display 2.13 Four distributions that are approximately normal.



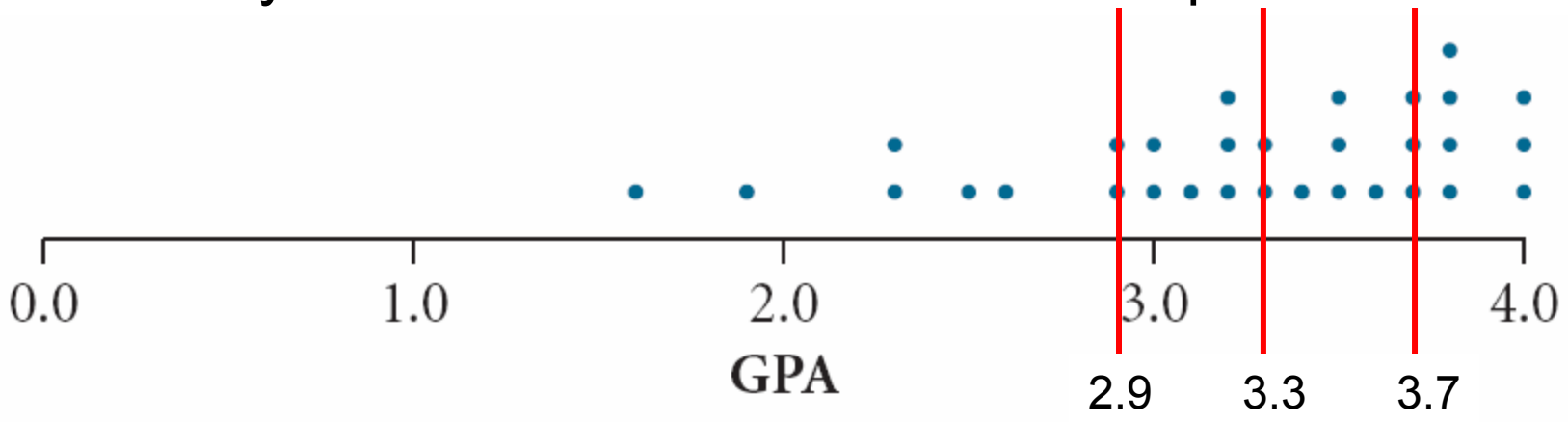
- Mean ~ 500
- SD ~ 100
- A typical SAT score is roughly 500, give or take 100 or so.



- Mean ~ 20
- SD ~ 5
- A typical ACT score is roughly 20, give or take 5 or so.

Practice

- P4. Estimate the median and quartiles for the distribution of GPAs in Display 2.7 on page 34. Then write a verbal summary of the same form as in the example.



Display 2.7 Grade-point averages of 62 statistics students. Each dot represents two points.

Lower quartile ~ 2.9
Median ~ 3.3
Upper quartile ~ 3.7

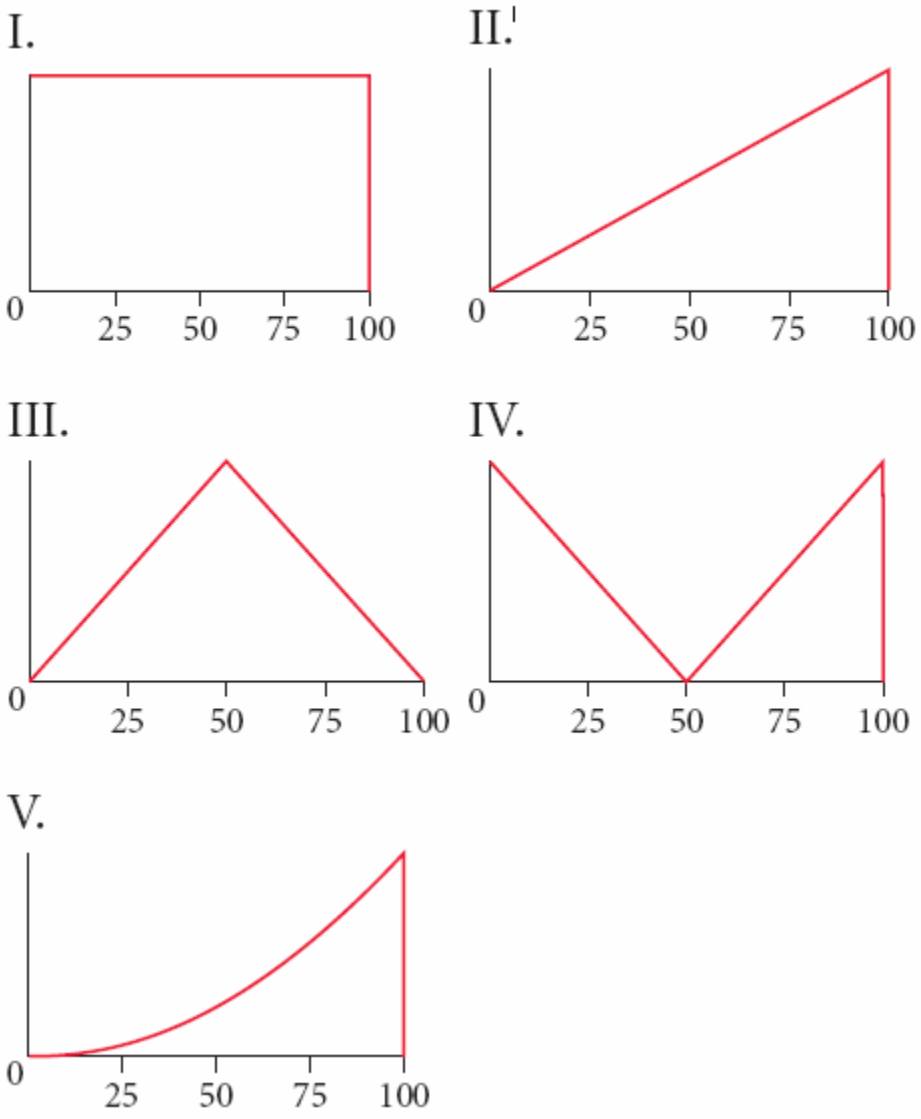
The middle 50% of the GPAs of statistic students were between 2.9 and 3.7, with half above 3.3 and half below.

Practice

P5. Match each plot in Display 2.14 with its median and quartiles (the set of values that divide the area under the curve into fourths).

- a. 15, 50, 85
- b. 50, 71, 87
- c. 63, 79, 91
- d. 35, 50, 65
- e. 25, 50, 75

IV
II
V
III
I



Quantitative vs. Categorical Data

- **Quantitative**: Data about the cases in the form of numbers that can be compared and that can take a large number of values.

- **Categorical**: Data where a case either belongs to a category or not.

Example (D6)

Mammal	Gestation Period (days)	Average Longevity (years)	Maximum Longevity (years)	Speed (mi/h)	Wild (1 = yes; 0 = no)	Predator (1 = yes; 0 = no)
Baboon	187	20	45	*	1	0
Bear, grizzly	225	25	50	30	1	1
Beaver	105	5	50	*	1	0
Bison	285	15	40	*	1	0
Camel	406	12	50	*	1	0
Cat	63	12	28	30	0	1
Cheetah	*	*	14	70	1	1
Chimpanzee	230	20	53	*	1	0
Chipmunk	31	6	8	*	1	0
Cow	284	15	30	*	0	0
Deer	201	8	20	30	1	0
Dog	61	12	20	39	0	1

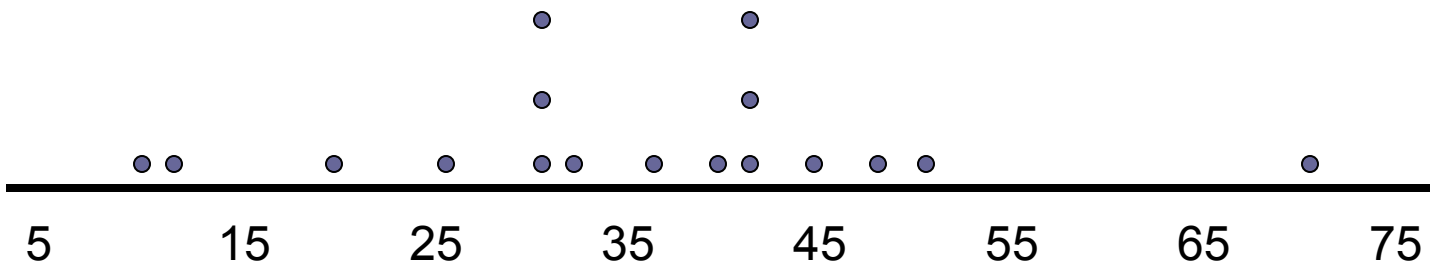
- Quantitative variables: Gestation period, average longevity, maximum longevity, speed.
- Categorical variables: Wild, predator.

Different ways to visualize data

- Quantitative Variables
  - Dot Plots
  - Histograms
  - Stemplots
- Categorical Variables
  - Bar Graphs

Dot Plots

- Each dot represents the value associated to a case.
  - Dots may have different symbols or colors.
  - Dots may represent more than one case.

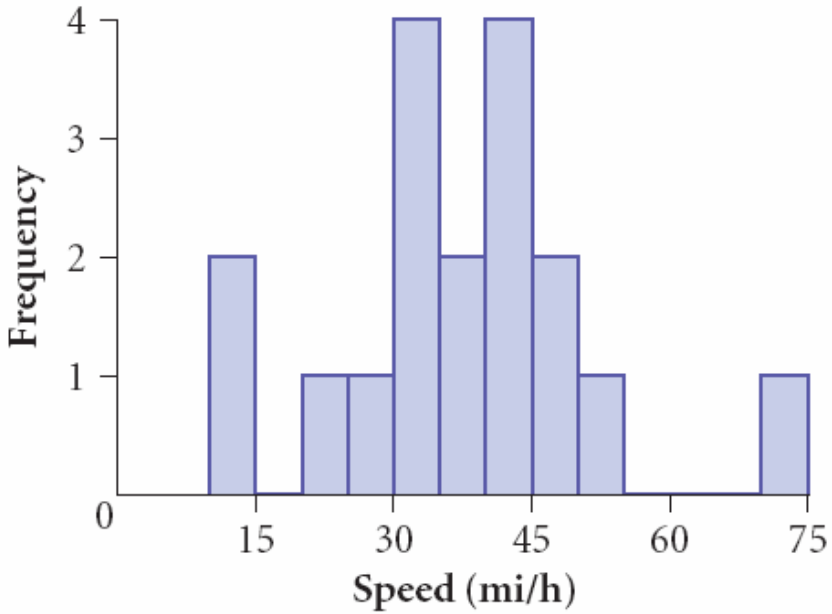


Dot Plots

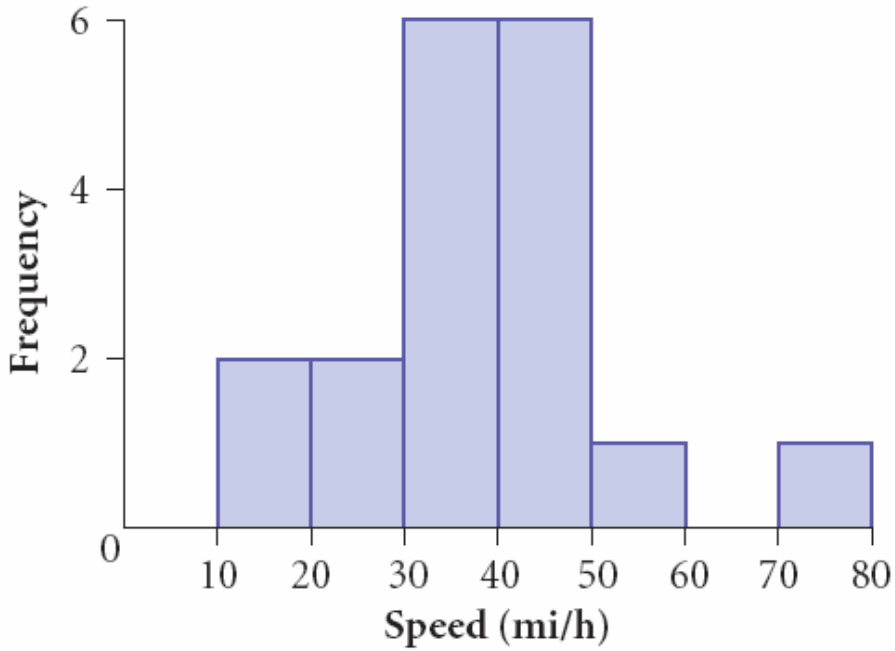
- Dot Plots work best when
  - Relatively small number of values to plot
  - Want to keep track of individuals
  - Want to see the shape of the distribution
  - Have one group or a small number of groups that we want to compare

Histograms

- Groups of cases represented as rectangles or bars
- The vertical axis gives the number of cases (called **frequency** or **count**) for a given group of values.
- By convention borderline values go to the bar on the right.
- There is no prescribed number for the width of the bars.

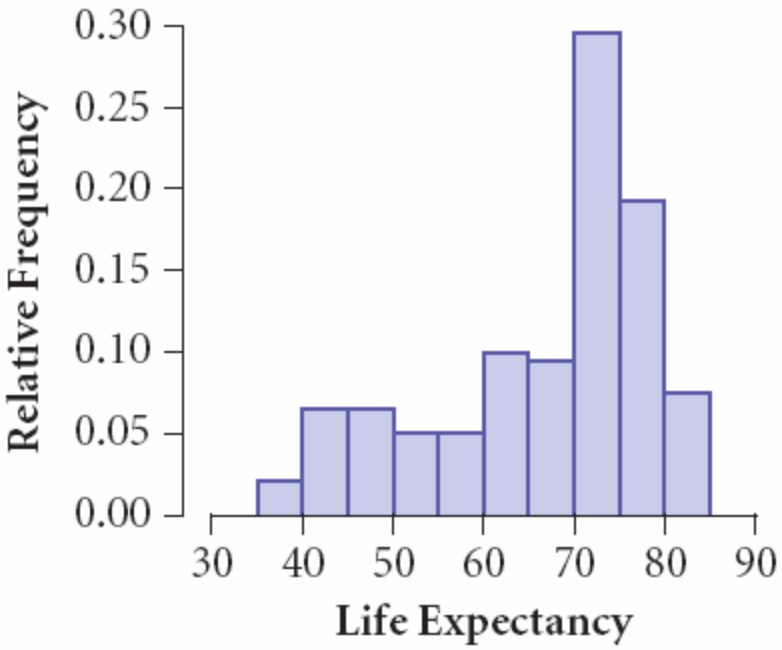


Display 2.26 Histogram of mammal speeds.



Relative Frequency Histograms

- The height of each bar is the proportion of values in that range. (always a number between 0 and 1)
- The sum of the heights of all the bars equals 1.
- To change a regular histogram to a relative frequency histogram just divide the frequency of each bar by the total number of



This histogram shows the relative frequency distribution of life expectancies for 203 countries around the world.

How many countries have a life expectancy of at least 70 but less than 75 years?
.30 x 203 = 60.9

What proportion of the countries have a life expectancy of 70 years or more?
.30+.19+.07 = .56 = 56 %

Histograms (Relative Frequency)

- Histograms work best when
  - Large number of values to plot
  - Don’t need to see individual values
  - Want to see the general shape of the distribution
  - Have one or a small number of distributions we want to compare
  - We can use a calculator or computer to draw the plots

Stemplots

- Also called **stem-and-leaf plots**.
- Numbers on the left are called **stems** (the first digits of the data value)
- Numbers on the right are the **leaves**. (the last digit of the data value)

Mammal speeds:

- 11,12,20,25,30,30,30,32,35, 39,40,40,40,42,45,48,50,70.

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

3 | 9 represents 39 miles per hour.

Stemplots (split)

- Each original stem becomes two stems.
- The unit digits 0,1,2,3,4 are associated with the first stem and they are placed on the first line.
- The unit digits 5,6,7,8,9 are associated with the second stem and they are placed on the second line from that stem.

1	1 2
-	
2	0
-	5
3	0 0 0 2
-	5 9
4	0 0 0 2
-	5 8
5	0
-	
6	
-	
7	0

3 | 9 represents 39 miles per hour.

Stemplot vs split stemplot

Mammal speeds:

- 11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

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

3 | 9 represents 39 miles per hour.

1	1 2
-	
2	0
-	5
3	0 0 0 2
-	5 9
4	0 0 0 2
-	5 8
5	0
-	
6	
-	
7	0

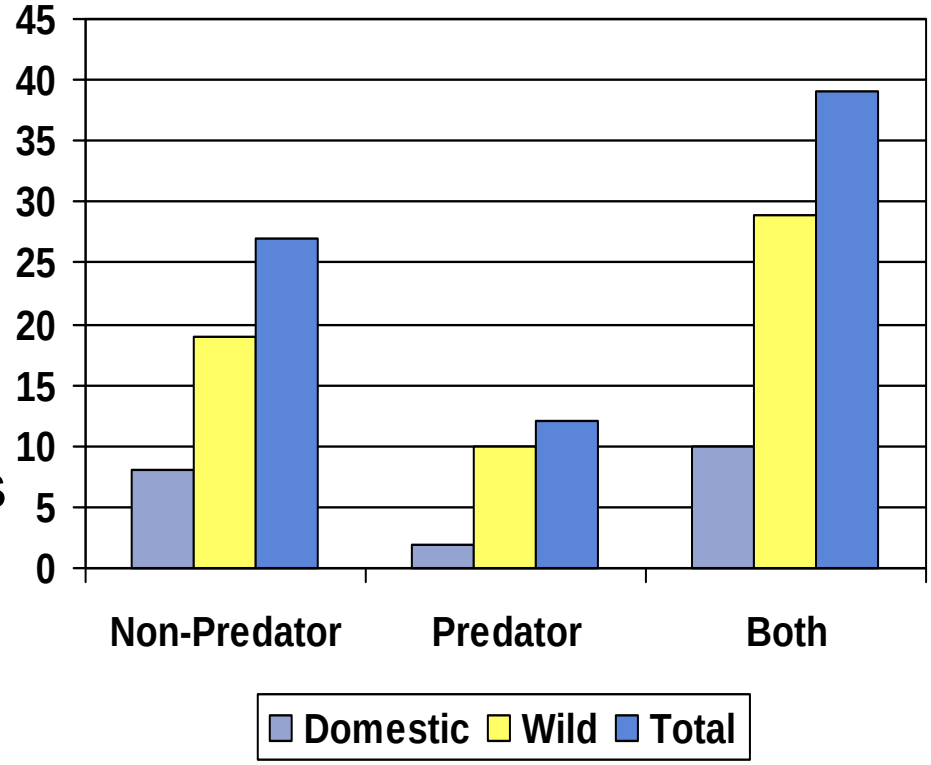
3 | 9 represents 39 miles per hour.

Stemplots

- Stemplots work best when
  - Plotting a single quantitative variable
  - Small number of values to plot
  - Want to keep track of individual values (at least approximately)
  - Have two or more groups that we want to compare

Bar Graphs

- One bar for each category.
- The height of the bar tells the frequency.
- Bar graphs have categories in the horizontal axis, as opposed to histograms which have measurements.



2.3 Measures of Center and Spread

- Before we used visual methods (estimations) to find out center (e.g. mean) and spread (e.g. SD). Now we will learn how to calculate them exactly.

- Measures of Center
  - Mean
  - Median
- Measures of Spread
  - Standard Deviation
  - Inter Quartile Range

Measures of Center

- **Mean**

The average of the data values denoted $\overline{x}$.

- Calculated as:

$$\overline{x} = \frac{\text{sum of values}}{\text{number of values}} = \frac{\sum x}{n}$$

- Example. Data Set: 5,12,34,18,37,11,9,21,30,6

$$\overline{x} = \frac{5+12+34+18+37+11+9+21+30+6}{10} = 18.3$$

Measures of Center

Median

The value that divides the data into equal halves. Denoted $median$ or Q_2.

- Calculated as:
  - List all values in increasing order and find the middle one.
  - If there are n values then the middle one is (n+1)/2
  - If n is even use the fact that the mid-value between a and b is (a+b)/2

Measures of Center

Median (examples)

Data set: 5,12,34,18,37,11,9,21,30,6.

Ordered data set:

5,6,9,11,12,18,21,30,34,37

$$median = \frac{12+18}{2} = 15$$

2. Data set: 6, 5 , 9, 12, 30, 18, 11, 34, 21.

Ordered data set:

5,6,9,11,12,18,21,30,34

Median = 12

Measure of spread around the Mean

- Most useful measure of spread when working with random samples.
- The deviation of a value is how far apart is it from the mean.

  $x - \overline{x}$
- Unfortunately it is easy to see that
- $\sum (x - \overline{x}) = 0$

- Standard Deviation
- There are two kinds σ_n and σ_{n-1}.
- The default is σ_{n-1}.
- They are calculated as:

$$\sigma_n = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

$$\sigma_{n-1} = \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}}$$

Measure of spread around the Mean

- Example. Data: 2,7,8,12,12,19
- $n = 6,\ \bar{x} = (2+7+8+12+12+19)/6 = 10$

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-8	64
7	-3	9
8	-2	4
12	2	4
12	2	4
19	9	81
Sum	60	0	166

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

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

$$\sigma_n = \sqrt{\frac{166}{6}} \approx 5.2599$$

$$\sigma_{n-1} = \sqrt{\frac{166}{5}} \approx 5.7619$$

Measure of spread around the Median

- Q_1 = First Quartile or Lower Quartile.
- Q_3 = Third Quartile or Upper Quartile.

- These are calculated as the medians of each of the two halves determined by the original median.
- In case n is odd then the original median is removed from each of the two halves.

- Inter Quartile Range

IQR = The distance between the Lower Quartile and the Upper Quartile.

$$IQR = Q_3 - Q_1$$

- About 50% of the values are between Q_1 and Q_3.

Five Number Summary

- min = Minimum (value)
- Q_1 = Lower or First Quartile
- Q_2 = Median
- Q_3 = Upper or Third Quartile
- max = Maximum (value)

In addition we also have

- Range = $max - min$
- $IQR = Q_3 - Q_1$

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

- min = 11
- Q_1 = 30
- Median = (35+39)/2 = 37
- Q_3 = 42
- max = 70.

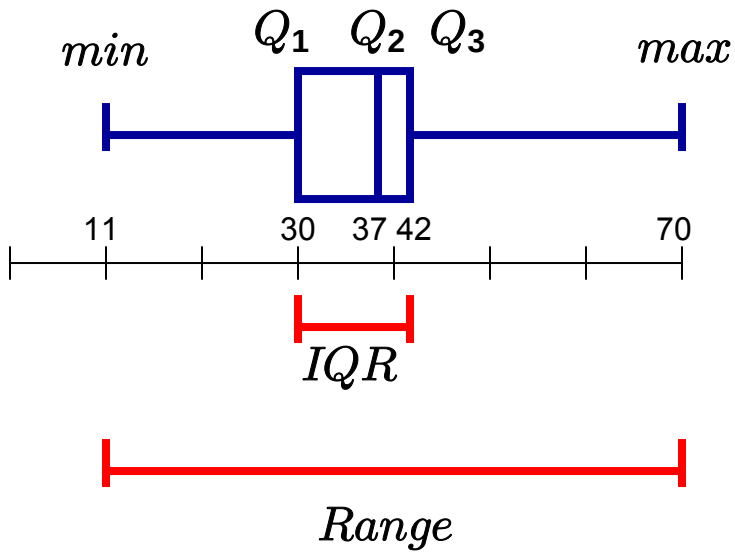
- Range = 70 – 11 = 59
- IQR = 42 – 30 = 12

Box Plots

- A **Box Plot** is a $graphical\ display$ of a five-point summary.

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

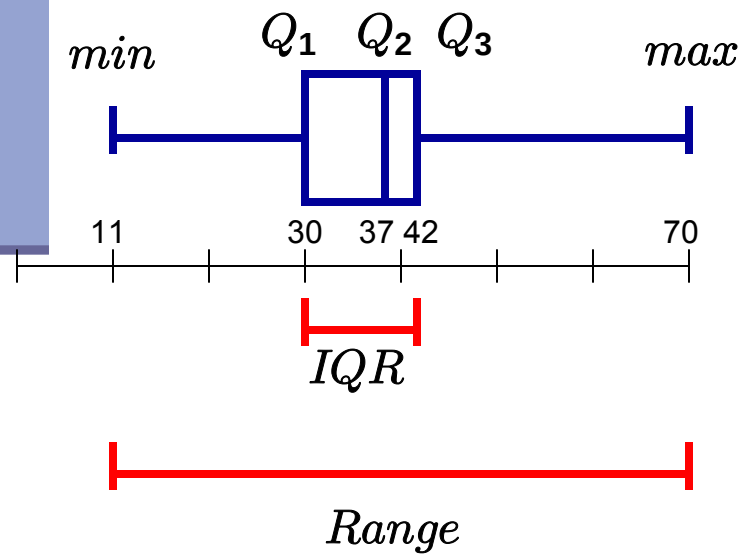
- min = 11
- Q_1 = 30
- Median = (35+39)/2 = 37
- Q_3 = 42
- max = 70.
- Range = 70 – 11 = 59
- IQR = 42 – 30 = 12



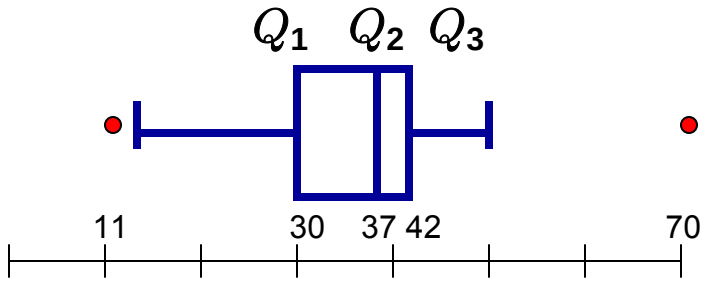
Modified Box Plots

- A **Modified Box Plot** also takes into account the **outliers**.
- An **outlier** is a value which is more than 1.5 times the IQR from the nearest quartile.

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.



- IQR = 42 – 30 = 12
- (1.5)IQR = (1.5)12 = 18
- 30 – 18 = 12 > 11, so 11 is an outlier.
- 42 + 18 = 60 <70, so 70 is an outlier.



Box Plots (Modified)

- Box Plots and Modified Box Plots are useful when plotting a single quantitative variable and:
  - We want to compare shape, center, and spread of two or more distributions.
  - The distribution has a large number of values
  - Individual values do not need to be identified.
  - (Modified) We want to identify outliers.

Four different tables

Cumulative distributions reflect the total value *accumulated* from top to bottom (left to right on a plot) on the corresponding table. (More on this p. 78.)

Frequency table*

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
Total	100

Cumulative frequency table

Weight	Frequency
2.99	1
3.01	5
3.03	9
3.05	13
3.07	20
3.09	37
3.11	61
3.13	78
3.15	91
3.17	97
3.19	99
3.21	100

Relative frequency table

Weight	Frequency
2.99	1/100=0.01
3.01	4/100=0.04
3.03	4/100=0.04
3.05	4/100=0.04
3.07	7/100=0.07
3.09	17/100=0.17
3.11	24/100=0.24
3.13	17/100=0.17
3.15	13/100=0.13
3.17	6/100=0.06
3.19	2/100=0.02
3.21	1/100=0.01
Total	100/100=1

Cumulative relative frequency table

Weight	Frequency
2.99	0.01
3.01	0.05
3.03	0.09
3.05	0.13
3.07	0.20
3.09	0.37
3.11	0.61
3.13	0.78
3.15	0.91
3.17	0.97
3.19	0.99
3.21	1

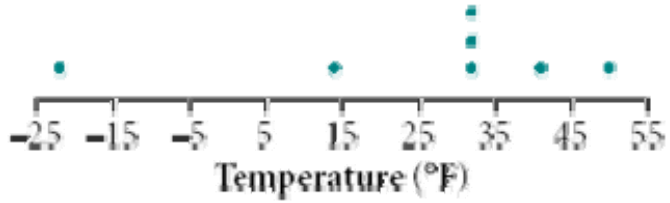
* This table shows the weights of the pennies in Display 2.3 on page 31.

Section 2.4 Recentering and Rescaling

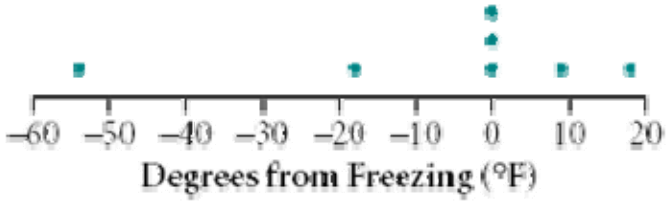
- **Recentering** a data set (adding the same number *c* to all the values in the set)
  - Shape or spread do not change.
  - It slides the entire distribution by the amount *c,* adding *c* to the median and the mean.
- **Rescaling** a data set (multiplying all the values in the set by the same positive number *d)*
  - Basic shape doesn't change.
  - It stretches or shrinks the distribution, multiplying the spread (*IQR* or standard deviation) by *d* and multiplying the center (median or mean) by *d*.

Example

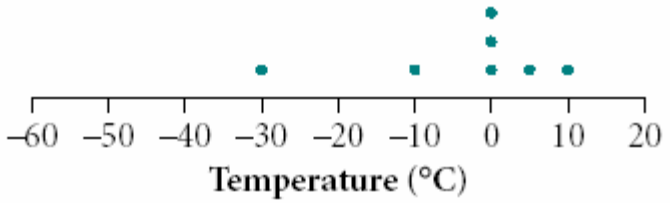
City	Country	Temperature (°F)
Addis Ababa	Ethiopia	32
Algiers	Algeria	32
Bangkok	Thailand	50
Madrid	Spain	14
Nairobi	Kenya	41
Brazilia	Brazil	32
Warsaw	Poland	-22



Display 2.63 Dot plot for record low temperatures in degrees Fahrenheit for seven capitals.



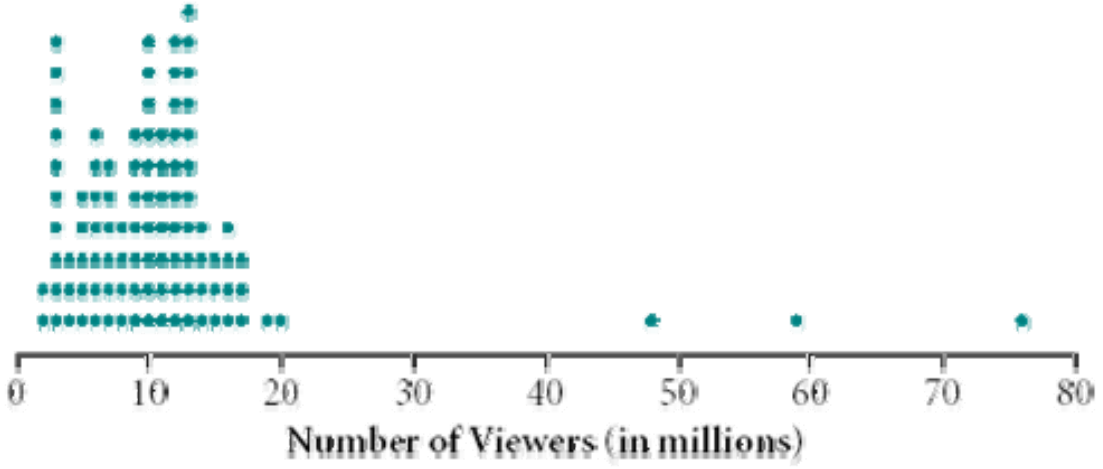
Display 2.64 Dot plot of the number of degrees Fahrenheit above or below freezing for record low temperatures for the seven capitals.



Display 2.65 Dot plot for record low temperatures in degrees Celsius for the seven capitals.

The Influence of Outliers

- A summary statistic is
  - **resistant to outliers** if it does not change very much when an outlier is removed.
  - **sensitive to outliers** if the summary statistic is greatly affected by the removal of outliers.



Display 2.66 Number of viewers of prime-time television shows in a particular week.

Example

Variable	N	Mean	Median	StDev
Ratings	101	11.187	10.150	9.896

Variable	Min	Max	Q1	Q3
Ratings	2.320	76.260	6.160	12.855

Display 2.67 Printout of summary statistics for number of viewers.

Variable	N	Mean	Median	StDev
No Outs	98	9.666	10.145	4.250

Variable	Min	Max	Q1	Q3
No Outs	2.320	20.470	6.065	12.698

Display 2.68 Summary statistics for number of viewers without outliers.

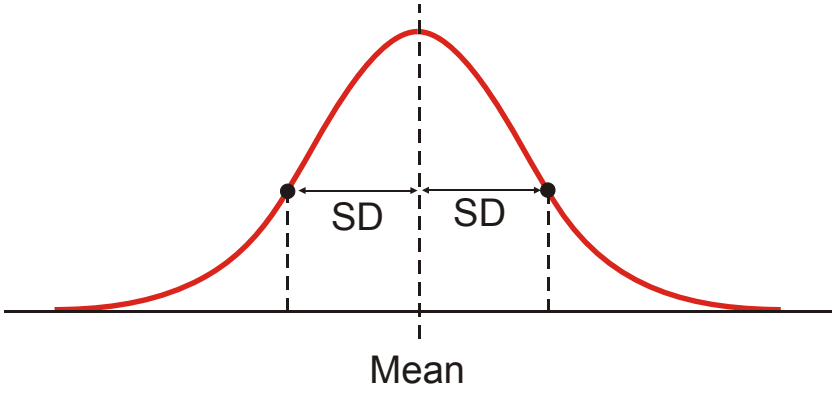
Display 2.66 Number of viewers of prime-time television shows in a particular week.

Percentiles and CRF plots

- You are responsible to read through this and understand the concepts of **percentile**, and **cumulative relative frequency plot**.

2.5 The Normal Distribution

- Shape



- Center: Mean

$$\bar{x} = \frac{\text{sum of values}}{\text{number of values}} = \frac{\sum x}{n}$$

- Spread: Standard Deviation

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

Applications of the Normal Distribution

- The normal distribution tells us how:
  - Variability in measures behaves.
  - Variability in population behaves.
  - Averages and some other summary statistics behave when you repeat a random process.

- Nice property: A normal distribution is determined by its mean and standard deviation!

  (If you know mean and SD you know everything)

The Two Main Problems.

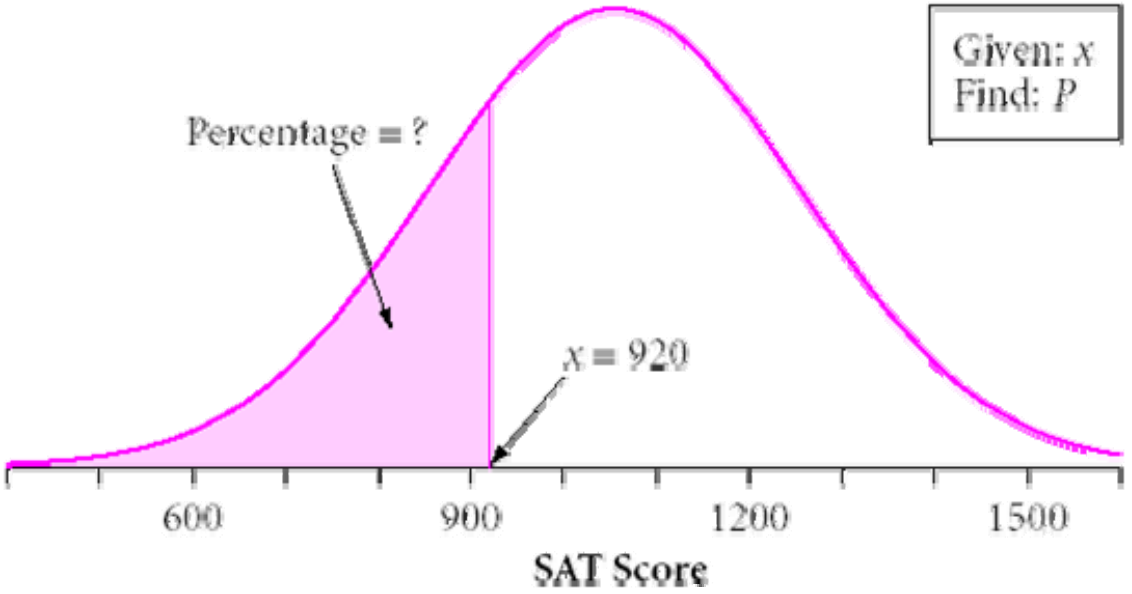
- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

1. What percentage of scores were 920 or below?
   (Unknown percentage problem)
2. What SAT score separates the lowest 25% of the SAT scores from the rest?
   (Unknown value problem)

Unknown percentage problem.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

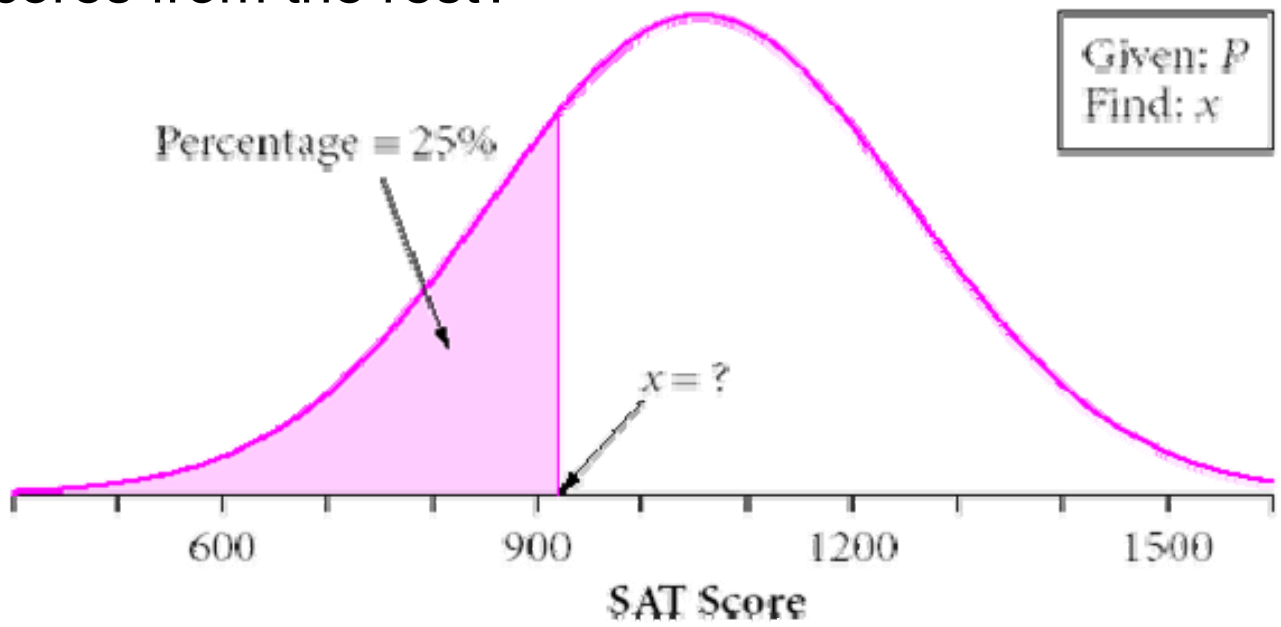
1. What percentage of scores were 920 or below?



Unknown value problem.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

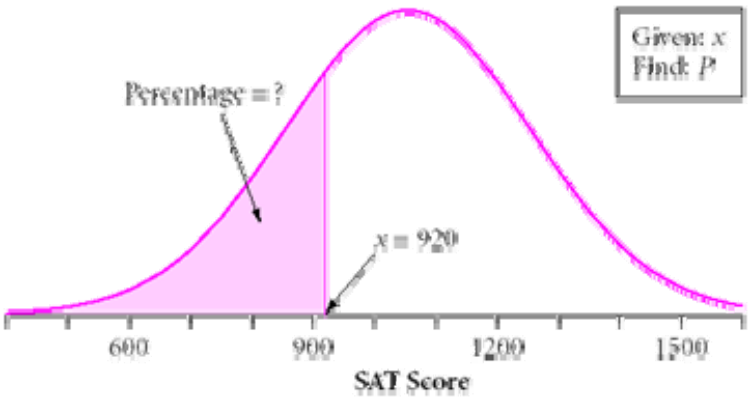
2. What SAT score separates the lowest 25% of the SAT scores from the rest?



Which one is it?

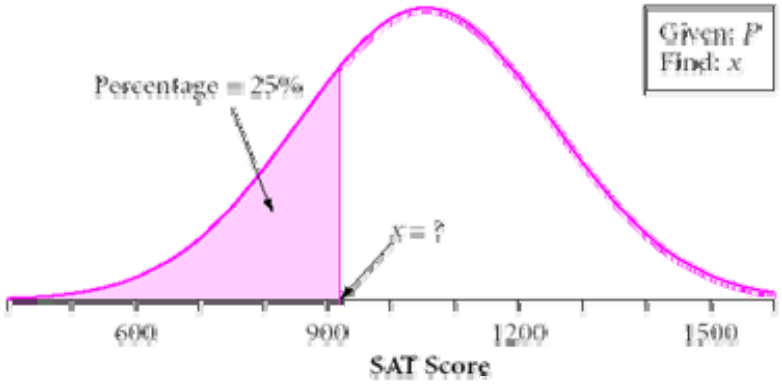
1. Unknown percentage problem.

   Given x, Find P.



2. Unknown value problem.

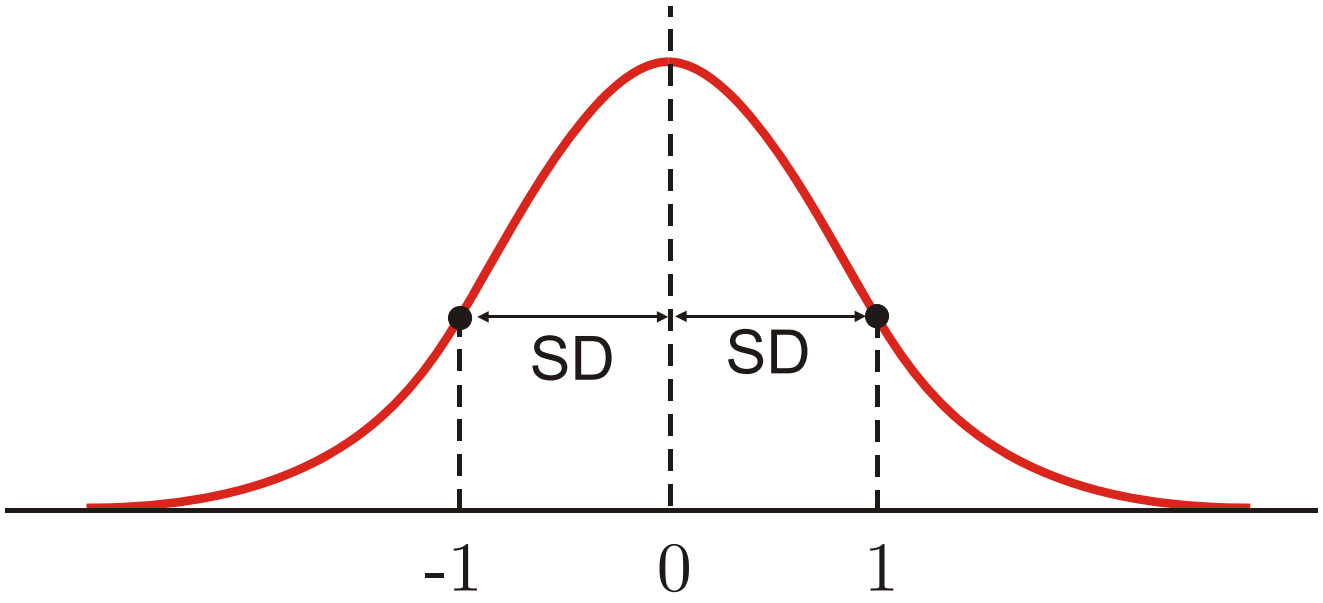
   Given P, Find x.



The Standard Normal Distribution.

- It is the normal distribution with **Mean = 0**, and **standard deviation = 1**.

  The area under the curve equals 1 (or 100%)



The Standard Normal Distribution.

- It is the normal distribution with Mean = 0, and standard deviation = 1.

  The area under the curve equals 1 (or 100%)
- The Standard Normal Distribution is important because any normal distribution can be **recentered** and/or **rescaled** to the standard normal distribution. This process is called **standarizing** or **converting to standard units**.
- Also, the two main problems can be easily solved in the Standard Normal Distribution with the help of **tables** or a **calculator**.

The Two Main Problems in the Standard Normal Distribution.

Unknown Percentage. (Given z, find P)

- With Table A (end of the textbook)
  - Use the units and the first decimal to locate the row and the closest hundredths digits to locate the column. The number found is the percentage of the number of values below z.
- With Calculator
  - Enter `normalcdf`(-99999, z) to get the percentage of the number of values below z.

Example (given z find P)

Find the percentage, P, of values below $z = 1.23$.

Calculator

P = normalcdf(-99999,1.23)

= .8906513833

~ 89.07%

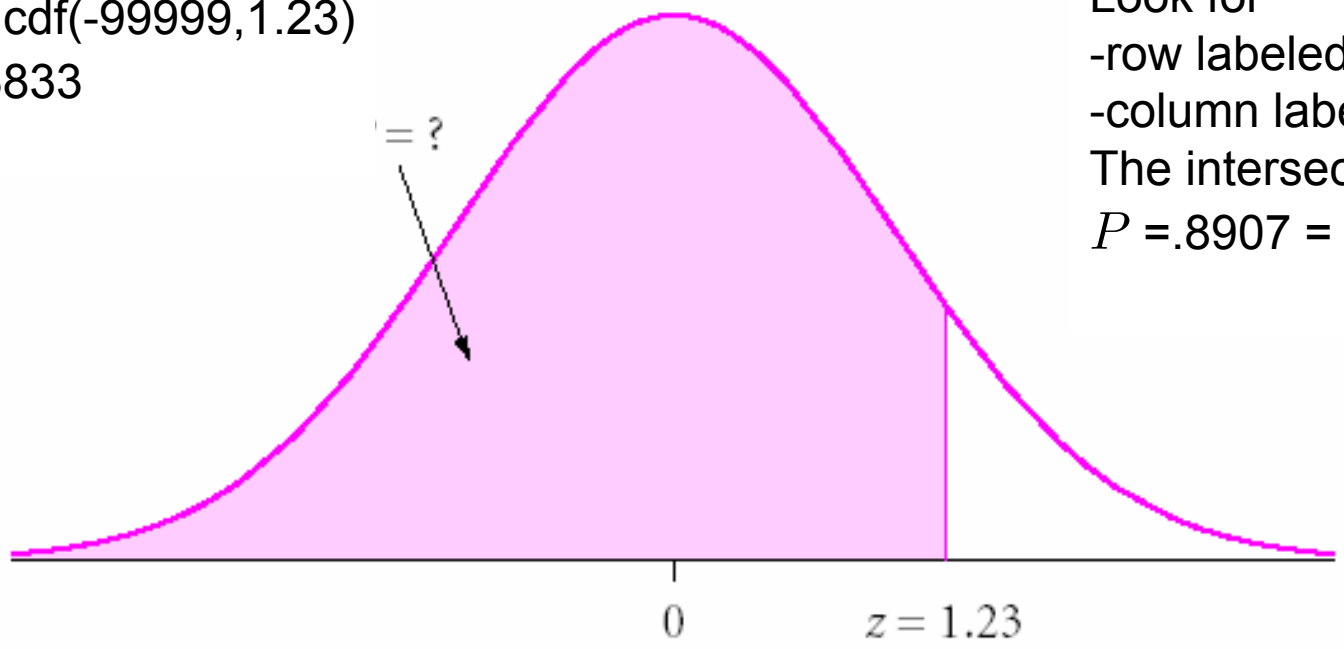


Table A

Look for

-row labeled 1.2

-column labeled .03

The intersection shows

P =.8907 = **89.07%**

Display 2.77 The percentage of values below $z = 1.23$.

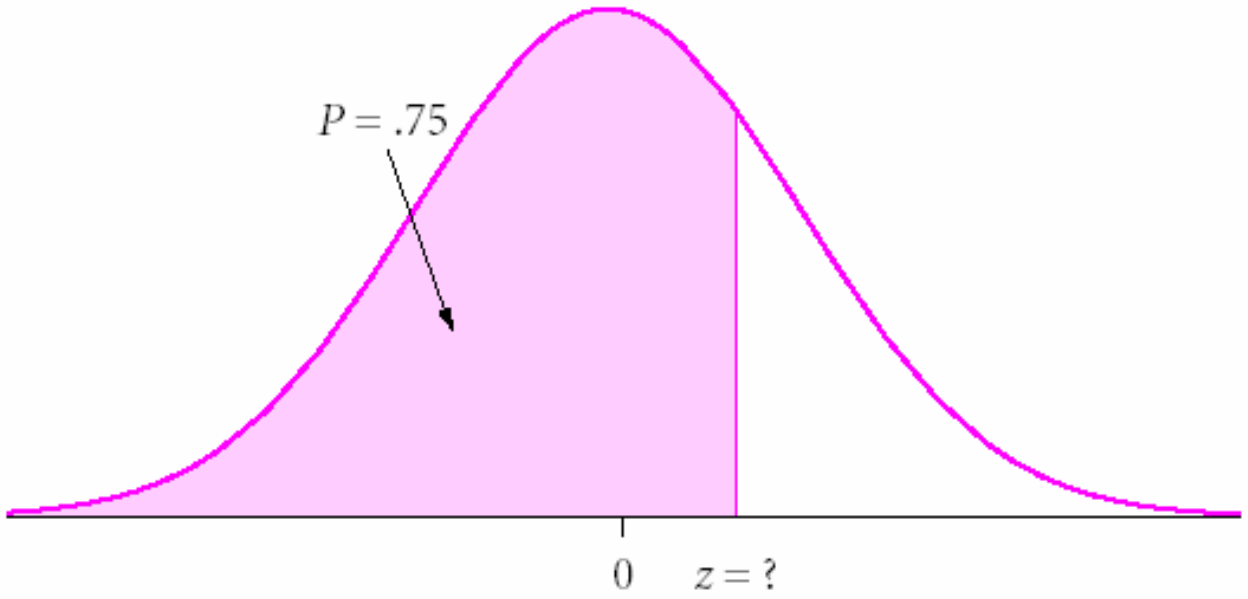
The Two Main Problems in the Standard Normal Distribution.

Unknown Value Problem. (Given P, find z)

- With Table A
  - Look for P in the body of the table. (or the number closest to it). Read back the row and column for that number. Use the row as the units and tenths of z, and the column as the hundredths digits of z. Note that P must be a percentage (written as a proportion, that is, a number between 0 and 1) of the number of values below a certain value z.
- With Calculator
  - Enter `invNorm`(P) to get the value z such that P equals the percentage of the number of values below z.

Example (given P find z-score)

Find the z-score that falls at the 75th percentile of the standard normal distribution; that is, the z-score that divides the bottom 75% of the values from the rest.



Calculator

z=invNorm(.75)

= .6744897495

~ .67

Table A

The value closest to .75 in the body of table A is .7486, which is in row .6 and column .07.

Then the z-score is **.67**

Standarizing

- When we standarize a value x it becomes z. We call z the z -score.

- Standard units = number of standard deviations that a given x value lies above or below the mean.

Standarizing

- As we said before, to standarize we just need to (re)center and (re)scale.
  - Step1. Centering (This makes mean = 0)

  Q: How far and which way to the mean?

  A: $x - \overline{x}$

  Subtract the mean from all values.

  - Step 2. Rescaling (this makes SD = 1)

  Q: How many standard deviations is that?

  A: $\dfrac{x - \overline{x}}{SD}$

  Divide all values from Step 1 by the SD.

$$z = \frac{x - \overline{x}}{SD}$$

Unstandarizing (reverse)

- Solve for x in the previous formula to get

$$x = \overline{x} + z \bullet SD = mean + z \bullet SD$$

where z is the z-score.

Value $\leftrightarrow$ z-score $(x \leftrightarrow z)$

- Standardizing (from x to z)

$$z = \frac{x - \overline{x}}{SD}$$

- Unstandardizing (from z to x)

$$x = \overline{x} + z(SD)$$

The two main problems (summary)

Unknown percentage

given x, find P

x to z to P

$$z = \frac{x - \overline{x}}{SD}$$

normalcdf(-99999, z)

Table: row and column

Unknown value

given P, find x

P to z to x

invNorm(P)

$$x = \overline{x} + z(SD)$$

Table: body

Example (p. 88 – given x find P)

Example

For groups of similar individuals, heights are often approximately normal in their distribution. For example, the heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 inches and standard deviation 2.7 inches. What percentage of these males are more than 74 inches tall?

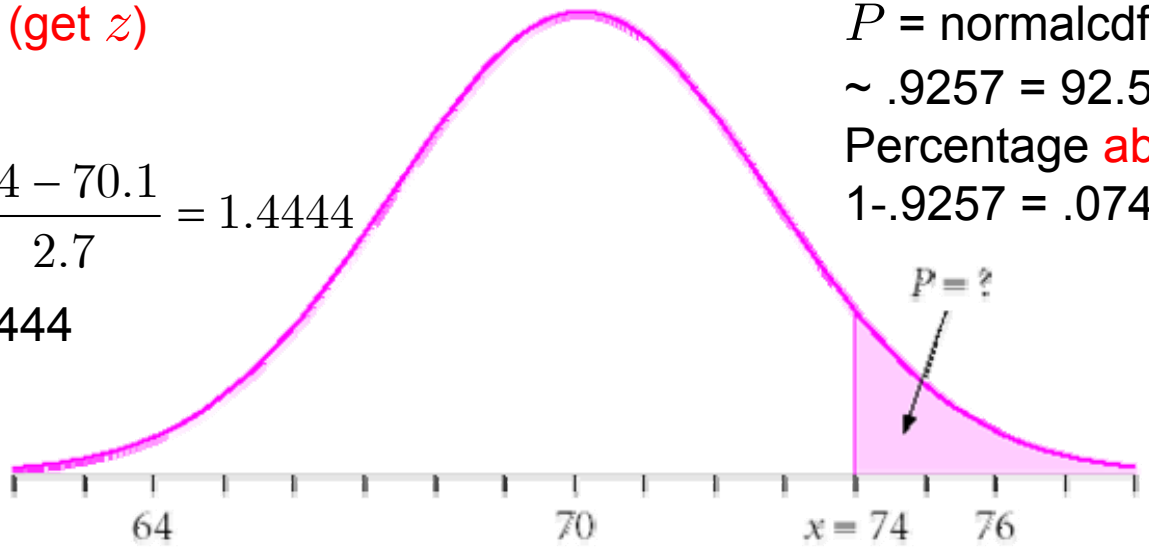
Source: *Statistical Abstract of the U.S. 1991.*

Standardize (get z)

$x = 74$

$$z = \frac{x - \bar{x}}{SD} = \frac{74 - 70.1}{2.7} = 1.4444$$

z-score=1.4444



Percentage below 74 in

P = normalcdf(-99999,1.4444)

~ .9257 = 92.57%

Percentage above 74 in

1-.9257 = .0743 = **7.43%**

or simply

P = normalcdf(74,99999, 70.1,2.7) ~ .0743 = **7.43%**

Example (p. 89 – given P find x)

Example

The heights of females in the United States who are between the ages of 18 and 24 are approximately normally distributed, with mean 64.8 inches and standard deviation 2.5 inches. What height separates the shortest 75% from the tallest 25%?

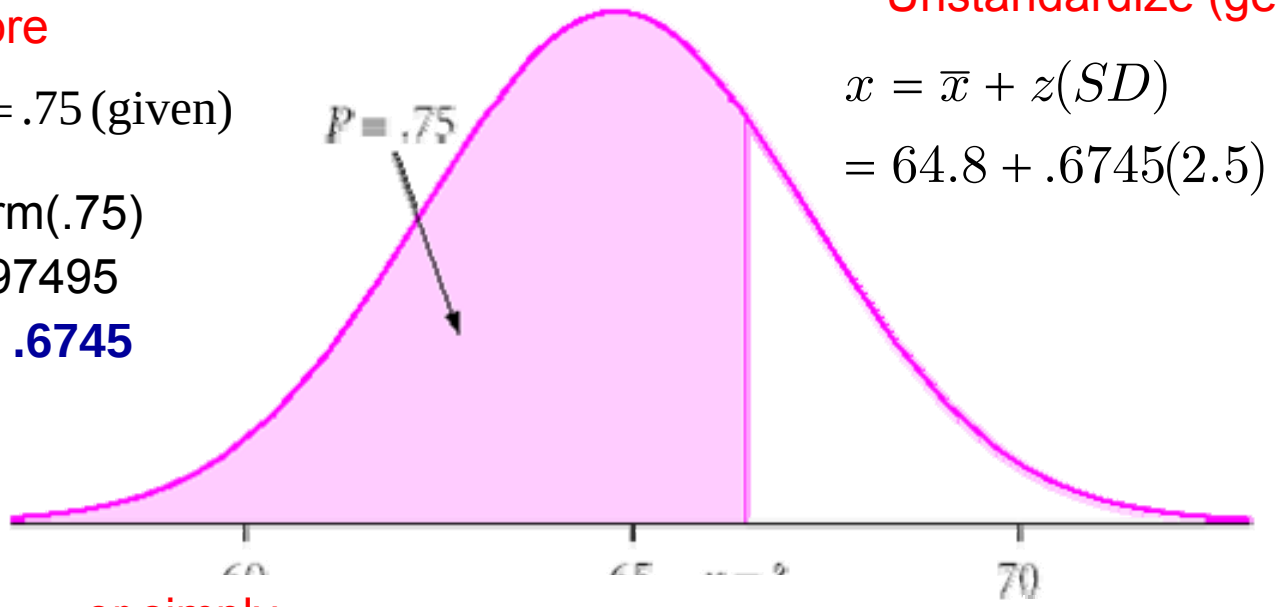
Get z-score

$P = 75\% = .75$ (given)

z =invNorm(.75)

= .6744897495

z-score = **.6745**



Unstandardize (get x)

$x = \overline{x} + z(SD)$

$= 64.8 + .6745(2.5) = 66.486$ in

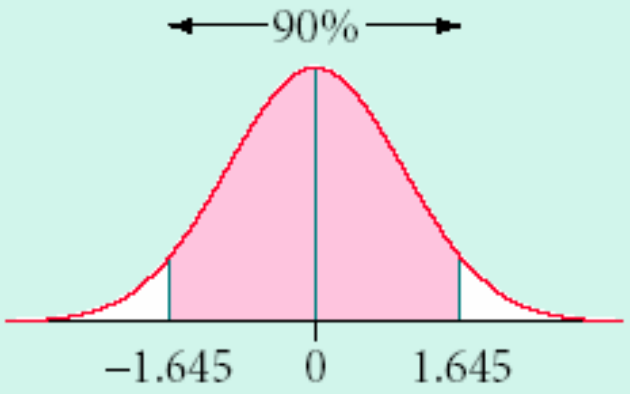
or simply

x = invNorm(.75,64.8,2.5) = **66.486**

Example (p. 91 – given P find x)

- According to the table on page 87, the distribution of death rates from cancer per 100,000 residents by state is approximately normal*, with mean 196 and *SD* 31. The middle 90% of death rates are between what two numbers?

90% of the values lie within 1.645 standard deviations of the mean.



*Provided that Alaska and Utah, which are outliers because of their unusually young populations, are left out.

Example (p. 91 – given P find x) cont.

- According to the table on page 87, the distribution of death rates from cancer per 100,000 residents by state is approximately normal*, with mean 196 and *SD* 31. The middle 90% of death rates are between what two numbers?
- Get z-scores (middle 90% is between 5% and 95%)

  5% =.05 corresponds to $z = \text{-}1.64485$

  95%= .95 corresponds to $z = 1.64485$
- Unstandardize

$$x = \overline{x} + z(SD) = mean + z(SD) = 196 + (-1.64485)(31) = 145.00965$$

$$x = x + z(SD) = mean + z(SD) = 196 + (1.64485)(31) = 246.99035$$

- Or simply x_1 = invNorm(.05,196,31) = **145.0095**

  x_2 = invNorm(.95,196,31) = **246.9905**
- So the middle 90% of states have between 146 and 246 deaths per 100,000 residents.

*Provided that Alaska and Utah, which are outliers because of their unusually young populations, are left out.

Problem 4 – Homework 2

Introduced in 2000, the Honda Insight was the first hybrid car sold in the U.S. The mean gas mileage for the model year 2006 Insight with an automatic transmission is 57.6 miles per gallon on the highway. Suppose the gasoline mileage of this automobile is approximately normally distributed with a standard deviation of 2.8 miles per gallon.

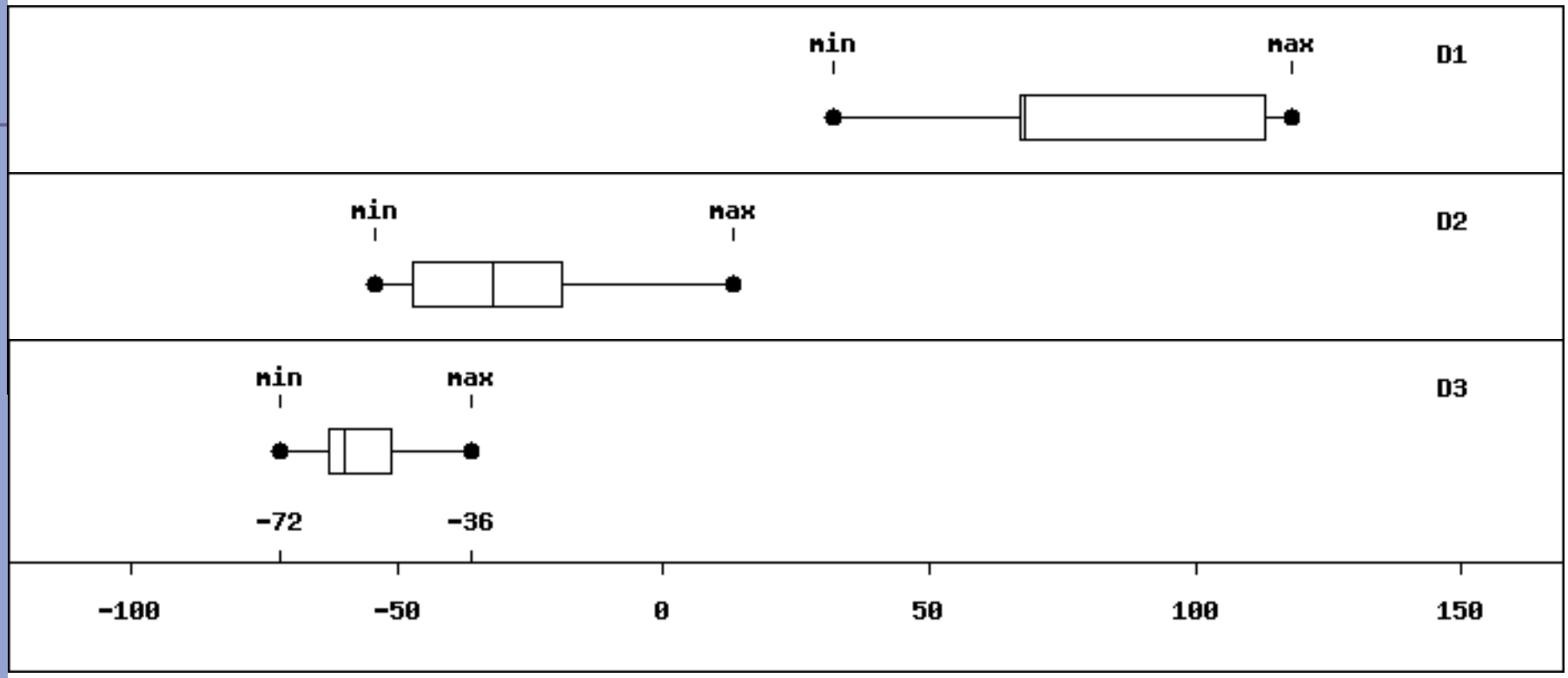
- (a) What proportion of 2006 Honda Insights with automatic transmission gets 60 miles per gallon or less on the highway?
- (b) What proportion of 2006 Honda Insights with automatic transmission gets between 58 and 62 miles per gallon on the highway?

Problem 1 – Homework 2

The scores of students on an exam are normally distributed with a mean of 395 and a standard deviation of 58.

- (a) What is the lower quartile score for this exam?
- (b) What is the upper quartile score for this exam?

Problem 4 – Homework 1



Which of the following are true?

A. At least three quarters of the data values represented in D1 are greater than the median value of D3 .
B. The data represented in D2 is symmetric.
C. The data for D1 has a greater median value than the data for D3 .
D. The data represented in boxplot D3 is skewed to the right.
E. All the data values for boxplot D1 are greater than the median value for D2 .
F. At least one quarter of the data values for D3 are less than the median value for D2

Problem 2 – Homework 2

IQ scores have a mean of 100 and a standard deviation of 15. Greg has an IQ of 118.

- What is the difference between Greg's IQ and the mean?
- Convert Greg's IQ score to a z score:

Problem 3 – Homework 2

Mike took 4 courses last semester: History, Spanish, Calculus, and Biology. The means and standard deviations for the final exams, and Mike's scores are given in the table below. Convert Mike's score into z scores.

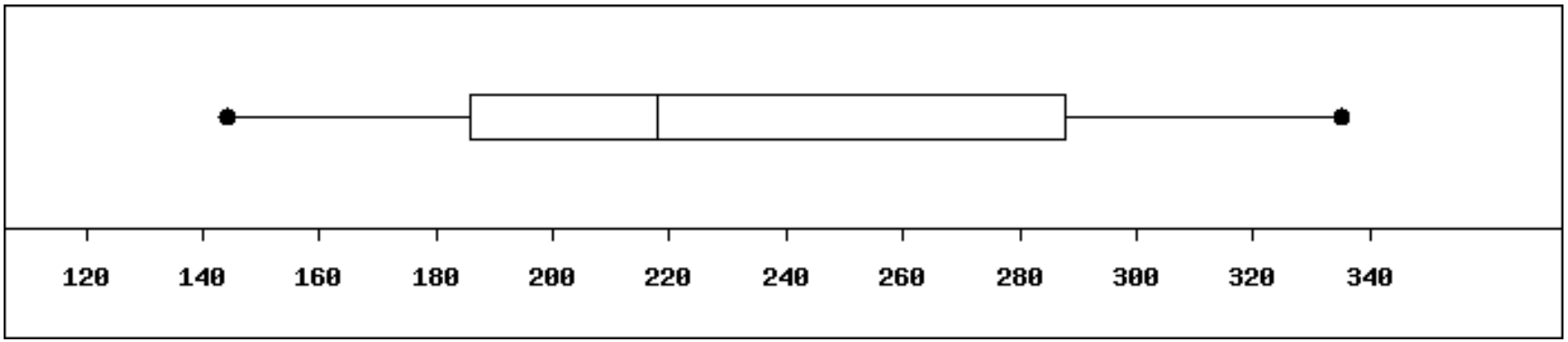
Subject	Mean	Standard deviation	Mike's score	Mike's z-score
History	53	16	49	
Spanish	44	12	38	
Calculus	70	12	88	
Biology	77	10	94.5	

- On what exam did Mike have the highest relative score?

Problem 5 – Homework 1

The boxplot below represents annual salaries of attorneys in thousands of dollars in Los Angeles. About what percentage of the attorneys have salaries between $186,000 and $288,000?

A. 20%
B. 50%
C. 25%
D. None of the Above



Problem 8 – Homework 1

Consider the following data set. Give the five number summary listing values in numerical order:

Data set: 27, 67, 26, 47, 78, 81, 73, 95, 88, 42, 96, 34, 82, 87, 37, 64, 56, 42, 100