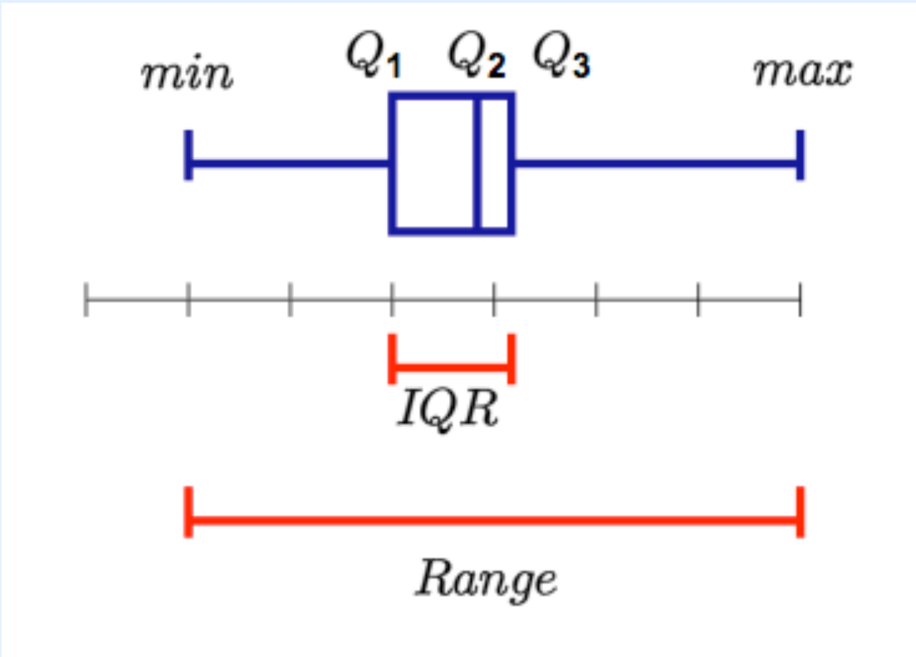
Math 140
Introductory Statistics

First midterm

September 23 2010

Box Plots

Graphical display of 5 number summary
Q1, Q2 (median), Q3, max, min



Outliers

If a value is more than 1.5 times the IQR
from the nearest quartile it may be an outlier

Look at the speeds of the animals.
Is the cheetah an outlier?
Is the pig an outlier?
Is the squirrel an outlier?
Is the lion an outlier?

Which animal is the largest non-outlier?

Outliers

If a value is more than 1.5 times the IQR
from the nearest quartile it may be an outlier

Stem-and-leaf of Speeds N = 18
Leaf Unit = 1.0 N* = 21

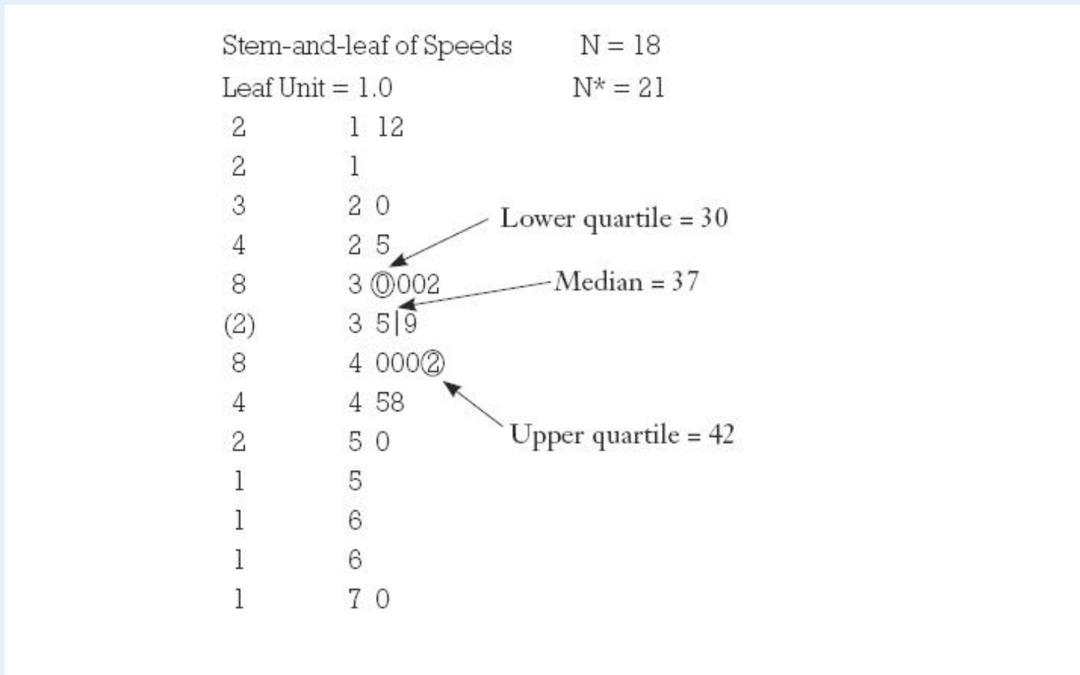
2	1	12
2	1	
3	2	0
4	2	5
8	3	0002
(2)	3	5|9
8	4	0002
4	4	58
2	5	0
1	5	
1	6	
1	6	
1	7	0

Lower quartile = 30
Median = 37
Upper quartile = 42

Q1=30
Q2=37
Q3 =42

Outliers

If a value is more than 1.5 times the IQR
from the nearest quartile it may be an outlier



IQR = 12 1.5*IQR = 18

Q3 + 1.5*IQR = 42+18 = 60 Q1 - 1.5*IQR = 30 - 18 = 12

Outliers

IQR = 12

1.5*IQR = 18

Q3 + 1.5*IQR = 42+18 = 60

Q1 - 1.5*IQR = 30 - 18 = 12

Cheetah = 70

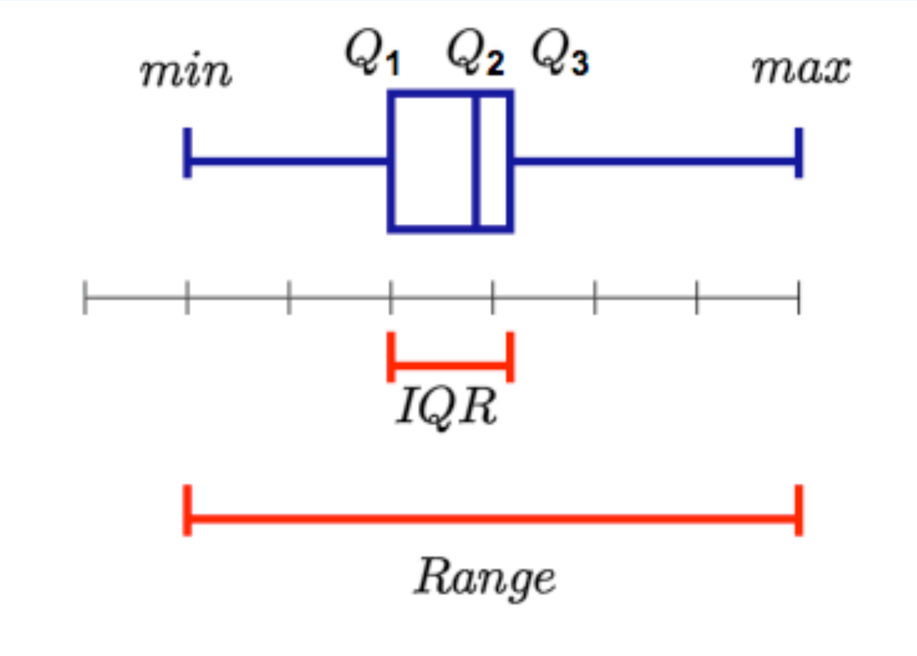
Pig = 11

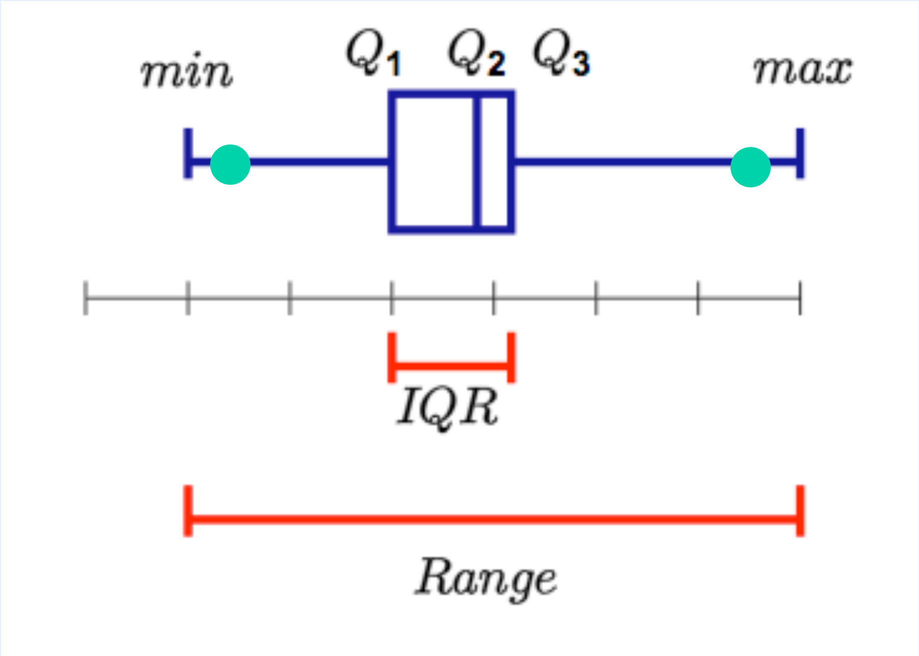
Squirrel = 12

Lion = 50

Modified box plot

Graphical display of 5 number summary
Q1, Q2, Q3, max, min and outliers





Modified box plot

Box plots

Box Plots are useful when

Plotting a single quantitative variable

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.
We may want to identify outliers.

Spread - Deviation

Deviation of a value x is how far it is from the mean

$$x - \bar{x}$$

This value is different for every data point x
and can be negative or positive

Standard deviation

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

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

Standard deviation

Data 2, 7, 8, 12, 12, 19 n=? average $\bar{x}$ = ?

x	$x-\bar{x}$	$(x-\bar{x})^2$
2		
7		
8		
12		
12		
19		
total sum = 60		

Standard deviation

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
60	0	166

Find σ_n and σ_{n-1}

Standard deviation

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
60	0	166

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

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

The square of the standard deviation is the variance

Standard deviation

The standard deviation is considered to be
the typical deviation from the mean

The larger the SD, the more spread out the data is

What if we have a frequency table?

Number of Strikes, x	Frequency, f	$x \cdot f$
0	3	0
1	3	3
3	2	6
5	1	5
6	1	6
7	3	21
8	2	16
9	1	9
14	1	14
15	1	15
Sum	18 years	95 strikes

What if we have a frequency table?

Number of Strikes, x	Frequency, f	$x \cdot f$
0	3	0
1	3	3
3	2	6
5	1	5
6	1	6
7	3	21
8	2	16
9	1	9
14	1	14
15	1	15
Sum	18 years	95 strikes

$$\bar{x} = \frac{\sum x \cdot f}{n} = \frac{95}{18} \approx 5.28$$

To calculate the mean we'd have to sum

0+0+0+1+1+1+3+3+ .. Or use the formula above

What if we have a frequency table?

Number of Strikes, x	Frequency, f	$x \cdot f$
0	3	0
1	3	3
3	2	6
5	1	5
6	1	6
7	3	21
8	2	16
9	1	9
14	1	14
15	1	15
Sum	18 years	95 strikes

$$\bar{x} = \frac{\sum x \cdot f}{n} = \frac{95}{18} \approx 5.28$$

[(0*3) + (1*3) + (3*2) + ….]/95

Recentering and Rescaling

Recentering a data set

Add the same number c to all values

The 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

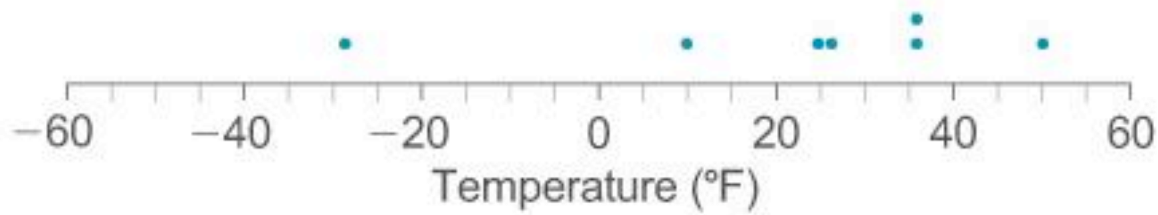
Multiply all values by the same positive number d

The basic shape doesn't change.

It stretches or shrinks the distribution, multiplying the spread (IQR or SD) by d and multiplying the center (median or mean) by d

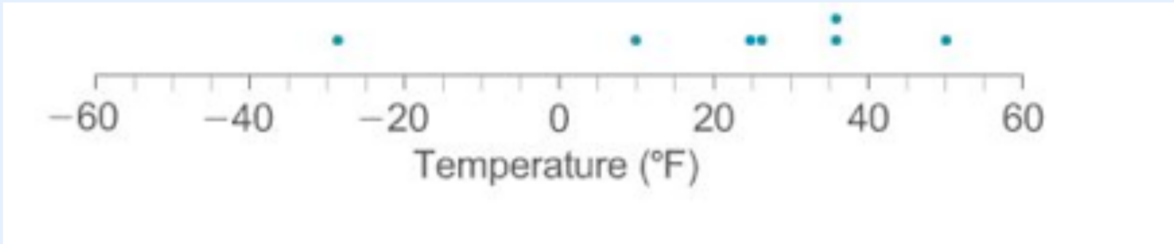
Recentering and Rescaling

City	Country	Temperature (°F)
Addis Ababa	Ethiopia	26
Algiers	Algeria	25
Bangkok	Thailand	50
Madrid	Spain	10
Nairobi	Kenya	36
Brasilia	Brazil	36
Warsaw	Poland	−29

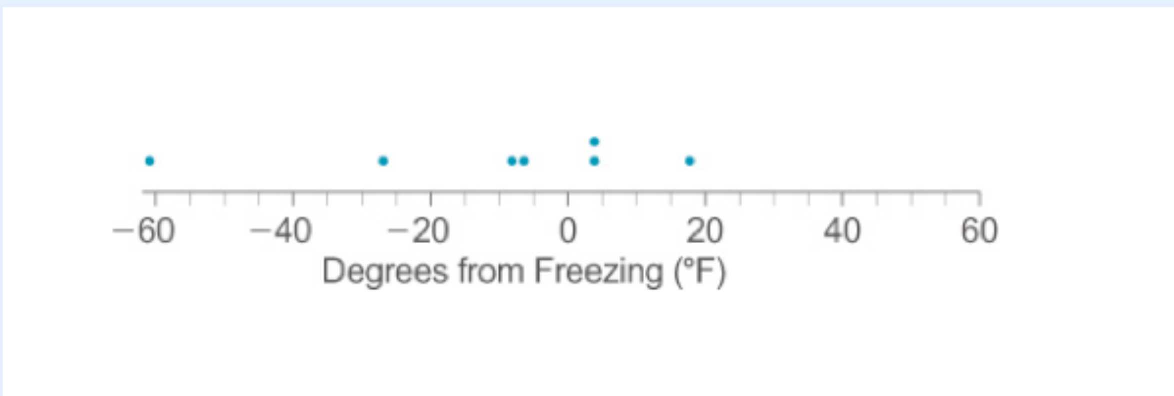


Want to move to Celsius $C = 5/9\,(F-32)$

Recentering

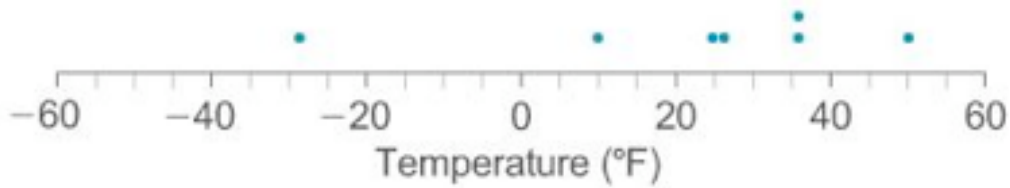


original

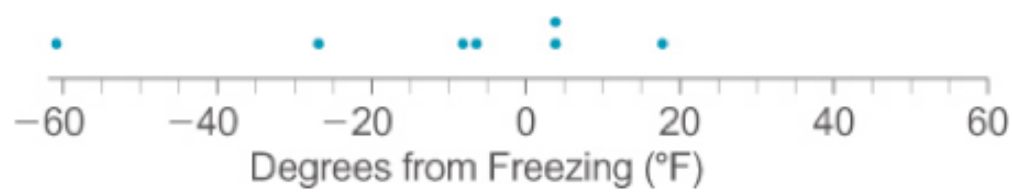


subtract 32

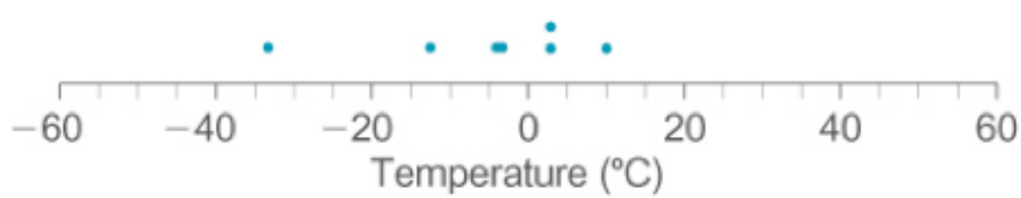
Rescaling



original

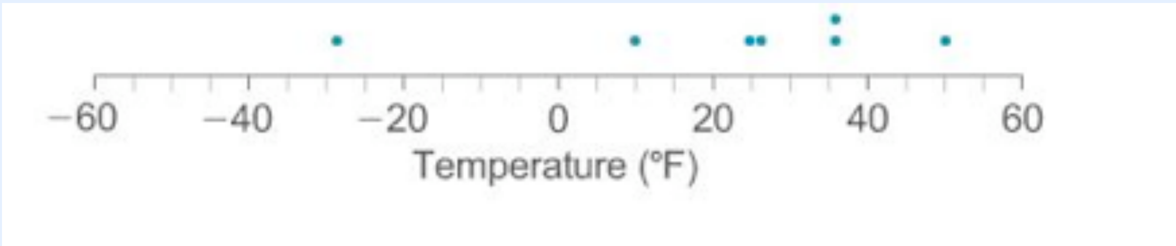


subtract 32

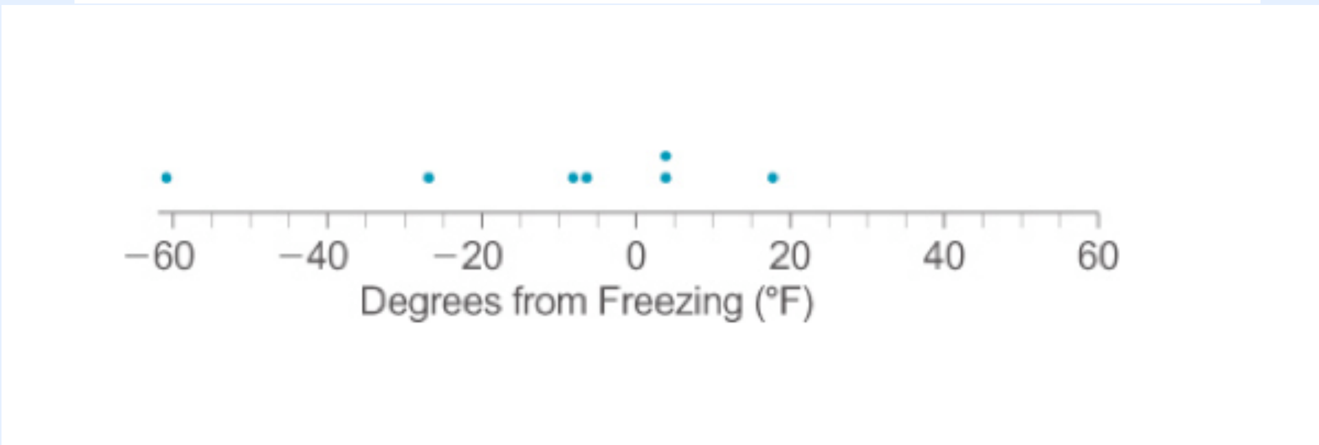


Multiply
by 5/9

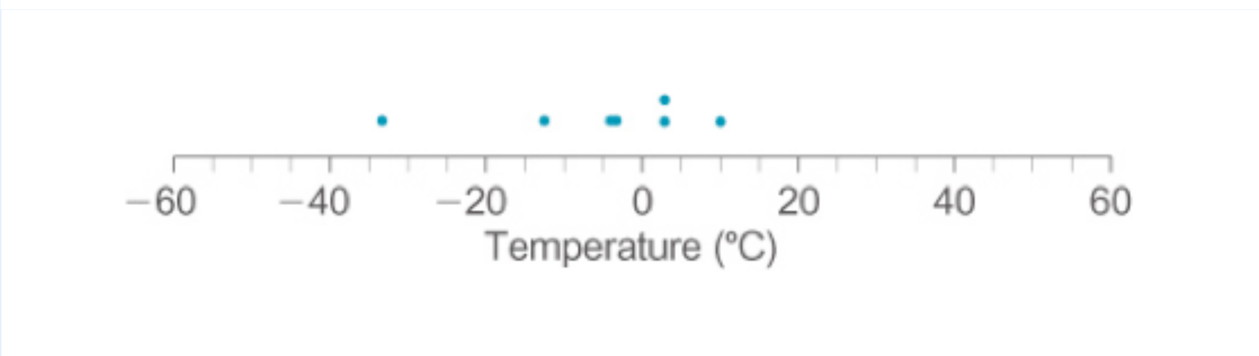
Rescaling



original



subtract 32



Multiply
by 5/9

A problem for you

Suppose a U.S. dollar is worth 14.5 Mexican pesos.

a. A set of prices, in U.S. dollars, has mean \$20 and standard deviation \$5. Find the mean and standard deviation of the prices expressed in pesos.

b. Another set of prices, in Mexican pesos, has a median of 145.0 pesos and quartiles of 72.5 pesos and 29 pesos. Find the median and quartiles of the same prices expressed in U.S. dollars.

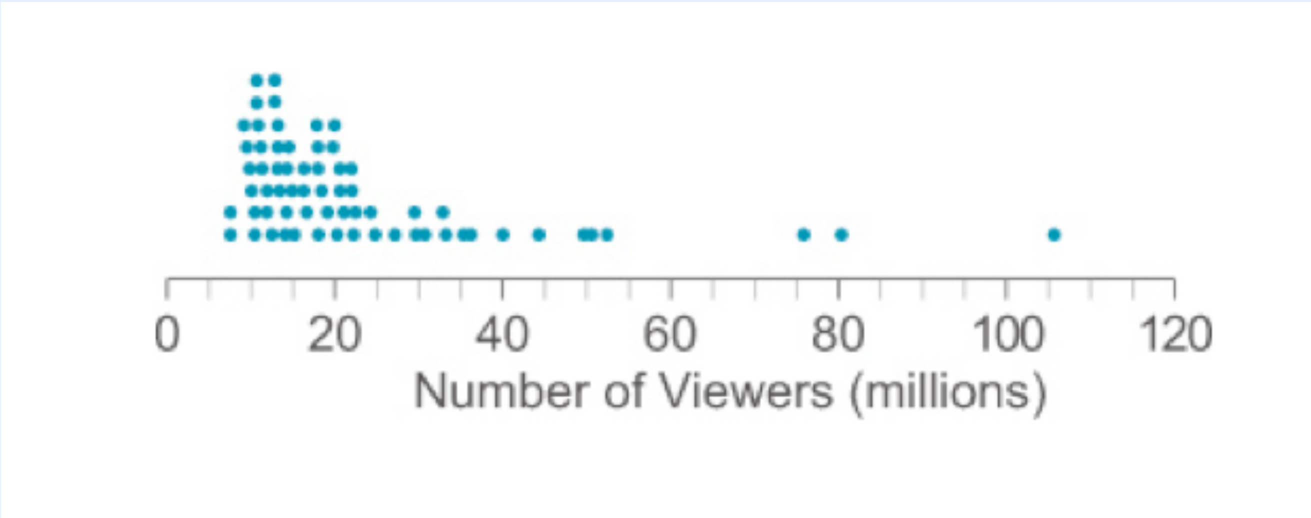
The influence of outliers

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.

The influence of outliers



Viewers for the finale of the most popular TV shows
Who are the outliers?
How do mean and SD change if we remove them?

The influence of outliers

Descriptive Statistics: Viewers

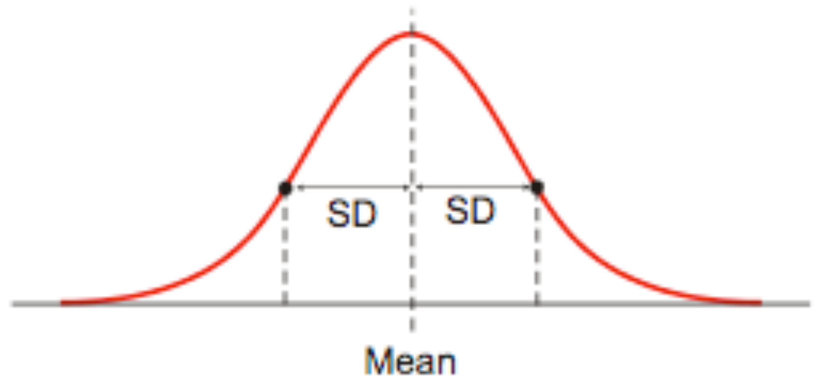
Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Viewers	68	0	23.04	17.63	7.40	12.95	18.00	25.05	105.90

Descriptive Statistics: No Outliers

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
No Outliers	65	0	20.07	10.65	7.40	12.60	17.50	23.40	52.50

Normal distributions

Shape



Center: Mean

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

Spread: Standard Deviation

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

Normal distributions

The normal distribution tells us how

averages and SD 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)

An example

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?

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

An example

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?

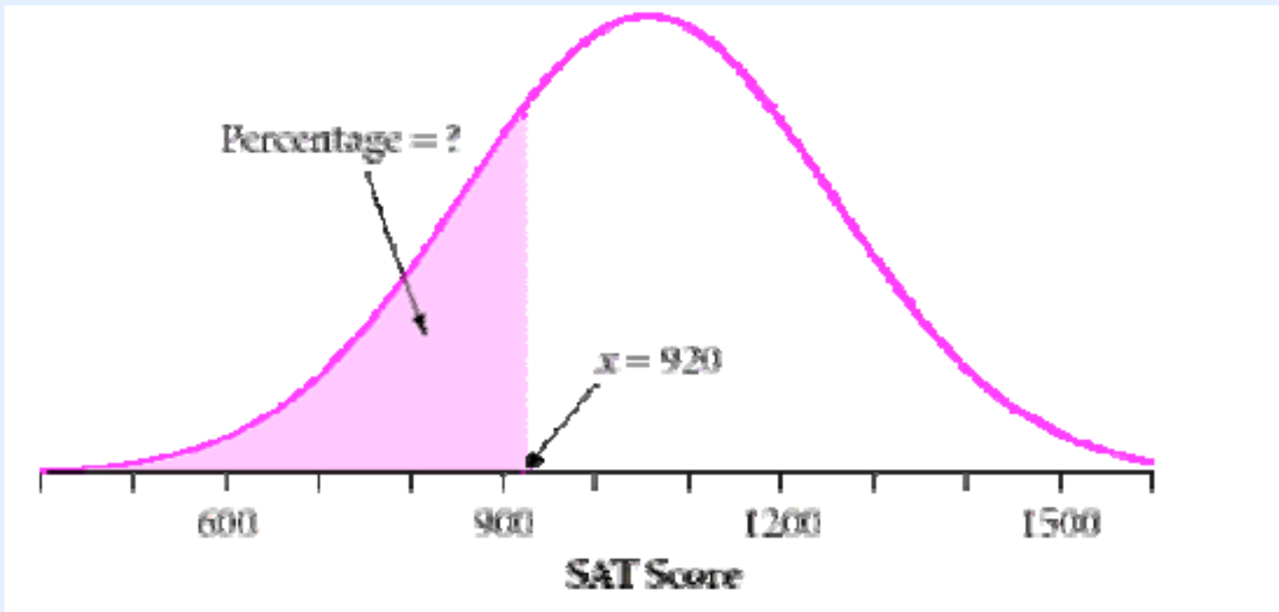
2. What SAT score separates the lowest 25% of the SAT scores from the rest?

We already know that 68% of data is between

855 and 1255

Unknown percentage problem

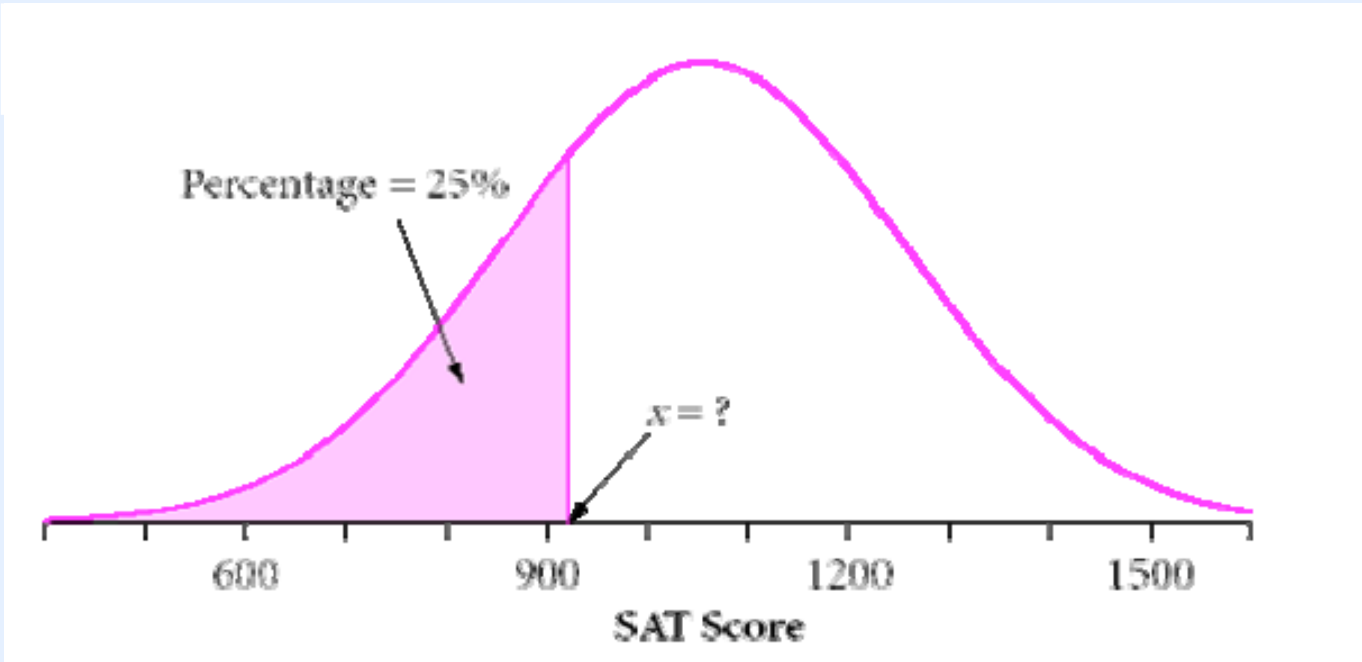
1. What percentage of scores were 920 or below?



Given z (a score), find the percentage

Unknown value problem

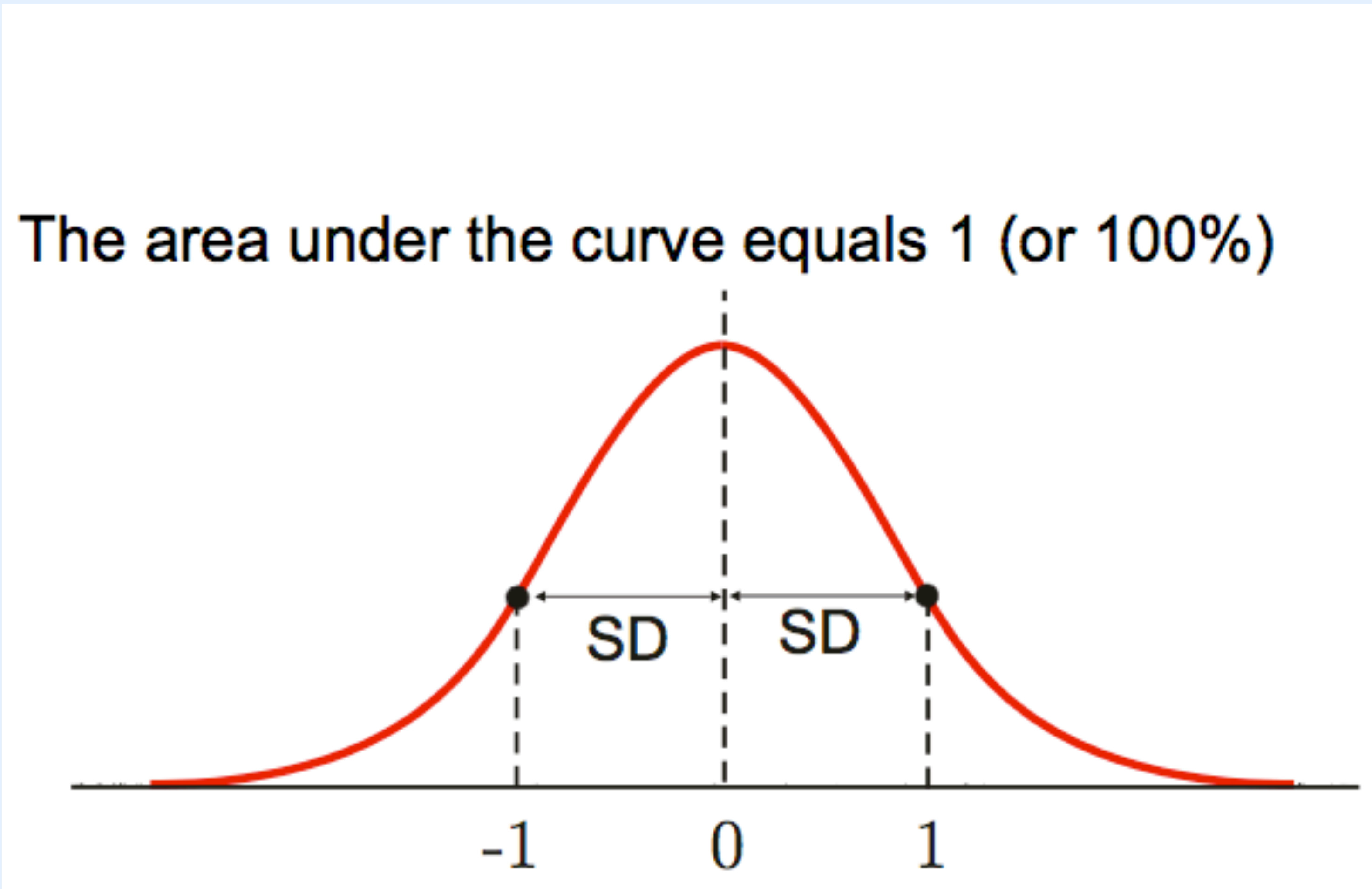
2. What SAT score separates the lowest 25% of the scores from the rest?



Given the percentage P, find the score z

Standard normal distribution

The normal distribution with mean =0 and SD = 1



Standard normal distribution

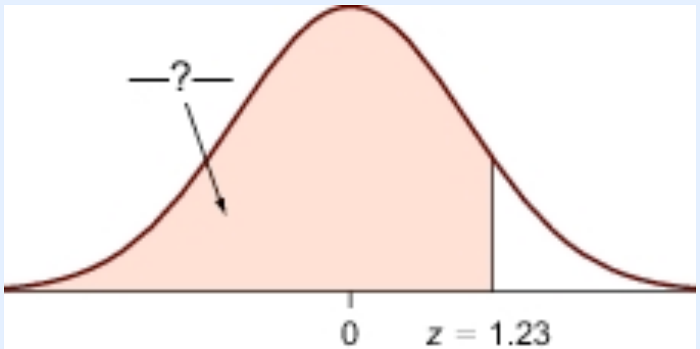
Any normal distribution can be rescaled
or recentered to give you the normal distribution

STANDARDIZING or
CONVERTING TO STANDARD UNITS

Given the score z find P
Unknown percentage

Table A. Page 759
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 value.



Hk

Page 73, E49, E50, E51, E52, E55, E59, E60