CHAPTER 5: EXPLORING DATA DISTRIBUTIONS

5.1 Creating Histograms

Individuals are the objects described by a set of data. These individuals may be people, animals or things.

A ***variable*** is any characteristic of an individual. A variable can take on different values for different individuals. Some variables are numeric and others are not.

Exploratory data analysis is the process of looking at data to describe the main features. Begin by looking at each variable and then the relationships between the variables. Graphs and numerical summaries are useful.

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

A ***histogram*** is a graph of the distribution of outcomes for a single numerical variable. The height of each bar is the number of observations in the class of outcomes covered by the base of the bar. All classes should have the same width and each observation must fall into exactly one class.

EXAMPLE

Display the data below in a histogram:

Value	Count
6	8
7	11
8	9
9	5
10	0
11	2

Steps to creating a histogram 8 to 12 ish

1. Choose the classes: Divide the range of the data into a "reasonable" number of classes of equal width.
2. Count the number of individuals in each class (frequency)
3. Draw the histogram. The vertical axis is the count in each class. The horizontal axis represents the classes.

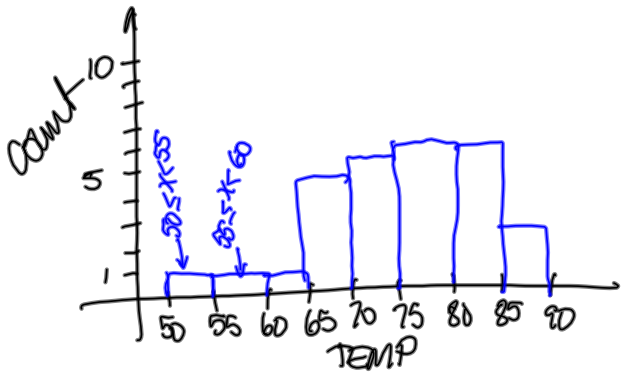
EXAMPLE

The following data is the recorded daily high temperature in College Station for March 2006. Display this data in a histogram.

86, 86, 85, 83, 83, 82, 82, 81, 81, 80,
79, 77, 77, 77, 76, 76, 75, 74, 74, 73,
72, 72, 72, 69, 69, 69, 67, 65, 61, 58, 51

class size 5

Value	Count
50-54	1
55-59	1
60-64	1
65-69	5
70-74	6
75-79	7
80-84	7
85-90	3



How may classes are there when the class size is 3?
(A) Less than 10 (B) 10 (C) 11 (D) 12 (E) More than 12

51-53
54-56
57-59
→ etc to ?

5.2 Interpreting Histograms

In any graph of data, look for patterns and deviations from the patterns.

- Does the graph have one or more peaks?
- Is the graph symmetric or skewed?
- Are there individual values that are far from the rest?
- Where is the center?
- Is most of the data spread out or close together?

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

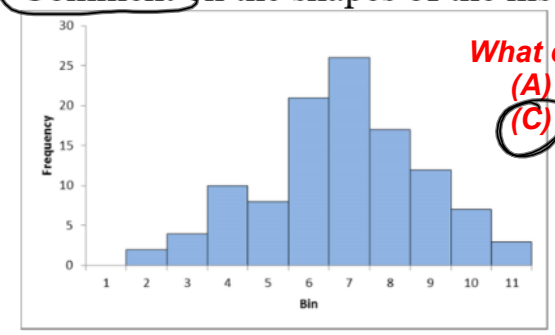
A graph is ***skewed to the right*** if the longer tail is on the right side. This is also called positively skewed

A graph ***is skewed to the left*** if the longer tail is on the left side. This is also called negatively skewed.

An ***outlier*** is an individual data value that falls outside the overall pattern.

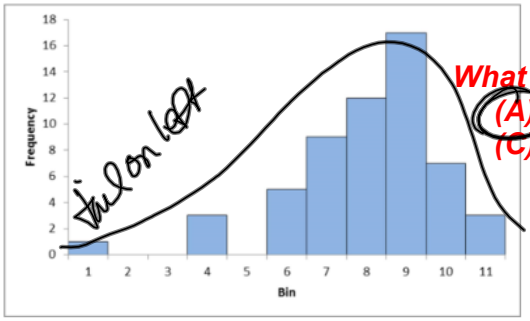
EXAMPLE

Comment on the shapes of the histograms below:



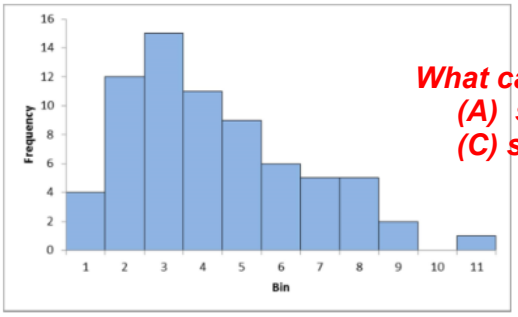
What can you say about the shape of the histogram?
(A) skewed left (B) skewed right
(C) symmetric (D) Not sure

ONE PEAK
No outliers



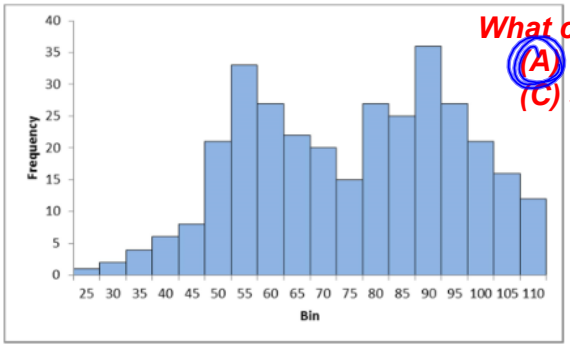
What can you say about the shape of the histogram?
(A) skewed left (B) skewed right
(C) symmetric (D) Not sure

one peak
maybe an outlier



What can you say about the shape of the histogram?
(A) skewed left (B) skewed right
(C) symmetric (D) Not sure

one peak



What can you say about the shape of the histogram?
(A) skewed left (B) skewed right
(C) symmetric (D) Not sure

TWO PEAKS

5.3 Creating Stemplots

A ***stemplot*** is a display of the distribution of a variable that attaches the final digits of the observations as leaves on stems made up of all but the final digit.

To Make A Stemplot

1. Separate each observation into a ***stem*** (consisting of all but the rightmost digit) and a ***leaf*** (the rightmost digit).
2. Write the stems in a vertical column with the smallest at the top. Include all the stem values from largest to smallest, even if some are not used. Draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem in increasing order

EXAMPLE

Display the following data in a stemplot:

02 05 07 09 11 15 16 18 18 23 23 25 25 28 29
29 34 35 37 39 40 43 44 45 45 57 72

0	2579
1	15688
2	33558 99
3	4579
4	03455
5	7
6	
7	2

EXAMPLE

Round the following data to the nearest 10, drop the ending zero and display the result in a stemplot.

118 → 12	210 → 21	301 → 30	321 → 32	349 → 35
122 → 12	216 → 22	302 → 30	328 → 33	393 → 39
160 → 16	247 → 25	304 → 30	333 → 33	403 → 40
161 → 16	250 → 25	313 → 31	334 → 33	411 → 41
203 → 20	266 → 27	316 → 32	335 → 34	605 → 61 (round up)
				1111 → 110

How many leaves are on the 4th stem?

(A) 0 (B) 1 (C) 2 (D) 3 (E) More than 3

1	2266
2	012557
3	0001223334459
4	01
5	
6	1

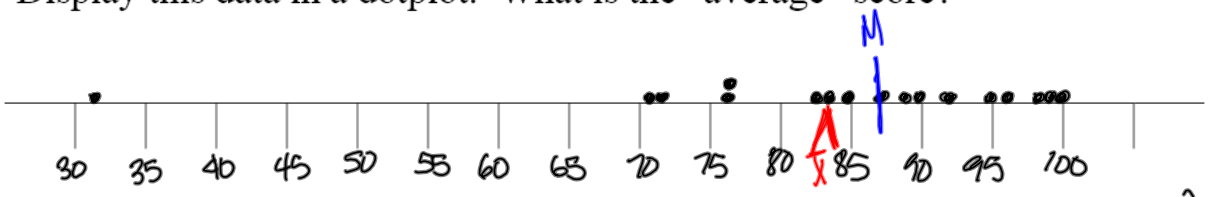
high

11	0

5.4 Describing Center: Mean and Median

EXAMPLE

The following are scores on an honors exam from a class of 17 students:
32, 71, 72, 77, 77, 83, 84, 85, 87, 89, 90, 92, 95, 96, 98, 99, 100
Display this data in a dotplot. What is the "average" score?



$$\text{Mean} = \frac{32+71+72+\ldots+100}{17} = \frac{1427}{17} = 83.9412 = \bar{x} \text{ (x bar)}$$

Median = 87 = M

Mode = 77

To find the ***mean*** of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$ then

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

To find the ***median, M,*** of a set of observations,

1. Arrange all the observations in increasing order
2. If the number of observations is odd, the median is the observation in the center of the ordered list.
3. If the number of observations is even, the median is the mean of the two center observations in the ordered list.

The ***mode*** of a set of observations is the observation that occurs the most frequently.

NO MODE
ONE MODE
MULTIPLE MODES

QUIZ

5.5 Describing Spread: The Quartiles

The ***range*** is a measure of spread of a set of observations. It is obtained by subtracting the smallest observation from the largest.

To calculate the ***quartiles*** Q_1 and Q_3,

1. Use the median to split the data set into two halves – an upper half and a lower half.
2. The ***first quartile*** Q_1 is the median of the lower half.
3. The ***third quartile*** Q_3 is the median of the upper half.

The ***interquartile range*** (or ***IQR***) is $Q_3 - Q_1$

EXAMPLE

The following are scores on an honors exam from a class of 17 students:

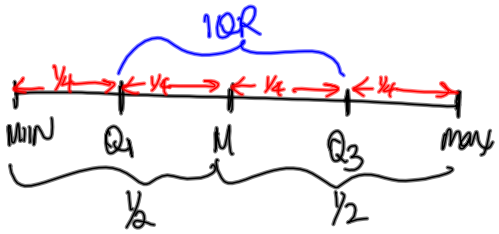
$Q_1 = \frac{77+77}{2} = 77$ M $Q_3 = \frac{96+95}{2} = 95.5$

32, 71, 72, 77, 77, 83, 84, 85, 87, 89, 90, 92, 95, 96, 98, 99, 100

What is the range of scores? What are the scores of the first and third quartiles? What is the IQR?

Range = 100 − 32 = 68

IQR = 95.5 − 77 = 18.5



EXAMPLE

The following data is the recorded daily high temperature in College Station for March 2006.

86, 86, 85, 83, 83, 82, 82, 81, 81, 80,
79, 77, 77, 77, 76, 76, 75, 74, 74, 73,
72, 72, 72, 69, 69, 69, 67, 65, 61, 58, 51

3 modes

Find the mean, the median, the mode, the range, the first and third quartiles, and the IQR. FOR THURSDAY

What is true about the mode for this data

(A) no mode (B) one mode

(C) more than one mode (D) Not sure

Median = 76

Mean $= \frac{86+86+85+\ldots+51}{31} = \frac{2312}{31} = 74.5806$

Clicker Question: Calculator?

(A) Brought or Bought one

(B) forgot it

(C) Borrowed one

Range = 86 − 51 = 35 ★

$Q_3 = 81$ $Q_1 = 69$

IQR = 81 − 69 = 12

5.6 The Five-Number Summary and Boxplots

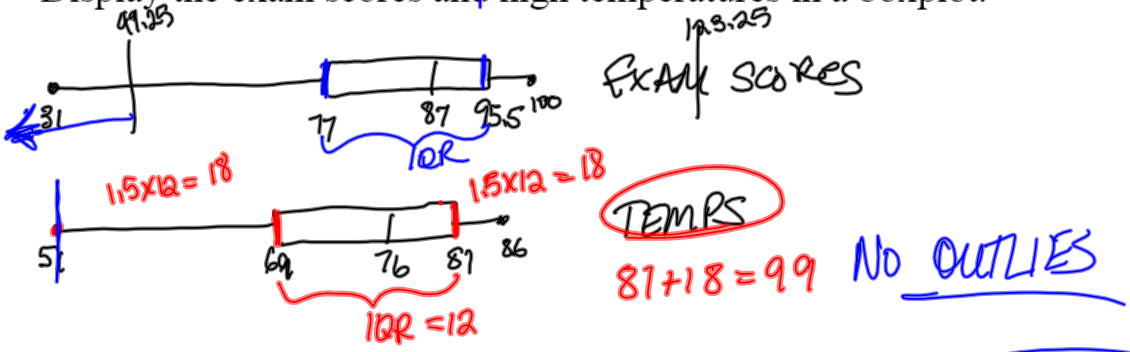
The ***five-number summary*** of a distribution consists of the following:

Minimum Q_1 M Q_3 Maximum

A boxplot is a graph of the five-number summary,

EXAMPLE

Display the exam scores and high temperatures in a boxplot.



One definition of an ***outlier*** is a data value that is less than $Q_1 - 1.5 \times IQR$ or is greater than $Q_3 + 1.5 \times IQR$.

Are there outliers in the exam grades?

IQR = 18,5

77 − 1,5 x 18,5 = 49,25 ⇒ 31 is a low outlier

95,5 + 1,5 x 18,5 = 123,25 ⇒ NO high outliers

EXAMPLE

The CDC reports that the number of new AIDS cases per year from 1990 to 2004 in Iowa are: 15 data points

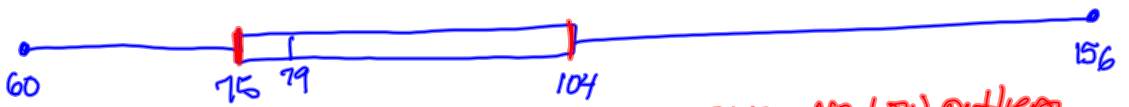
75, 118, 156, 104, 110, 104, 97, 76, 60, 78, 79, 80, 75, 76, 69

Show this data in a boxplot with labels. Are there outliers?

(A) Yes (B) No (D) Not sure

60, 69, 75, 75, 76, 76, 78, 79, 80, 97, 104, 104, 110, 118, 156

Q1 M Q3



IQR = 104 − 75 = 29

75 − 1.5 × 29 = 31.5 No low outliers

104 + 1.5 × 29 = 147.5 One high outlier

EXAMPLE

The CDC reports that the number of new Lyme disease cases per year from 1990 to 2004 in Iowa are:

16, 22, 33, 8, 17, 16, 19, 8, 27, 24, 34, 54, 65, 72, 49, 88

Show this data in a boxplot with labels. Are there any outliers?

(A) Yes (B) No (D) Not sure

8, 8, 16, 16, 17, 19, 22, 24, 27, 33, 34, 49, 54, 65, 72, 88

$Q_1 = \frac{16+17}{2} = 16.5$ $M = \frac{27+24}{2} = 25.5$ $Q_3 = \frac{49+54}{2} = 51.5$

EXAMPLE

Six numbers have min=5, max=19, mode=17 and med=16. Find a possible data set for these results.

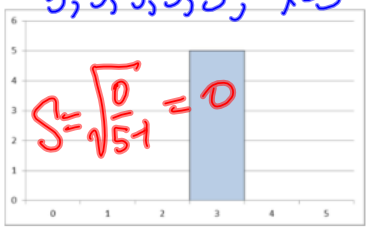
5 4 15 17 17 19

16

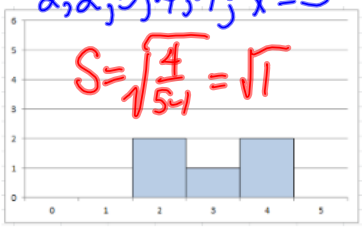
6, 7, 8, ... 14

5.7 Describing Spread: The Standard Deviation

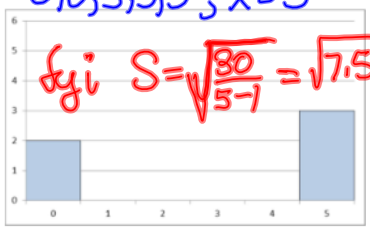
3,3,3,3,3; $\bar{x}=3$ $s=\sqrt{\frac{0}{5-1}}=0$



2,2,3,4,4; $\bar{x}=3$ $s=\sqrt{\frac{4}{5-1}}=\sqrt{1}$



0,0,5,5,5; $\bar{x}=3$ $s=\sqrt{\frac{30}{5-1}}=\sqrt{7.5}$



On "average", how far away are the measurements from the mean?

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	2-3 = -1	1
2	2-3 = -1	1
3	3-3 = 0	0
4	4-3 = 1	1
4	4-3 = 1	1
mean	0	4

$$s=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}=\sqrt{\frac{4}{5-1}}=\sqrt{1}=1$$

$$s=\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\dots+(x_n-\bar{x})^2}{n-1}}$$

EXAMPLE

The following table gives the average monthly temperature in two different cities for four different months. Find the standard deviation for each city and determine which one varies the most.

Month	Jan	Apr	Jul	Oct
San Diego	65	68	76	75
Chicago	29	59	84	64

(A) San Diego varies the most
(B) Chicago varies the most
(C) The are nearly the same
(D) Not sure

SD: $\frac{65+68+76+75}{4}=\bar{x}=71$

x	$x-\bar{x}$	$(x-\bar{x})^2$
65	65-71 = -6	36
68	68-71 = -3	9
76	76-71 = 5	25
75	75-71 = 4	16
	0	86

$s=\sqrt{\frac{86}{4-1}}=\sqrt{\frac{86}{3}}\approx 5.35$

Chicago $\bar{x}=\frac{29+59+84+64}{4}=\frac{236}{4}=59$

x	$x-\bar{x}$	$(x-\bar{x})^2$
29	-30	900
59	0	0
84	25	625
64	5	25
		1550

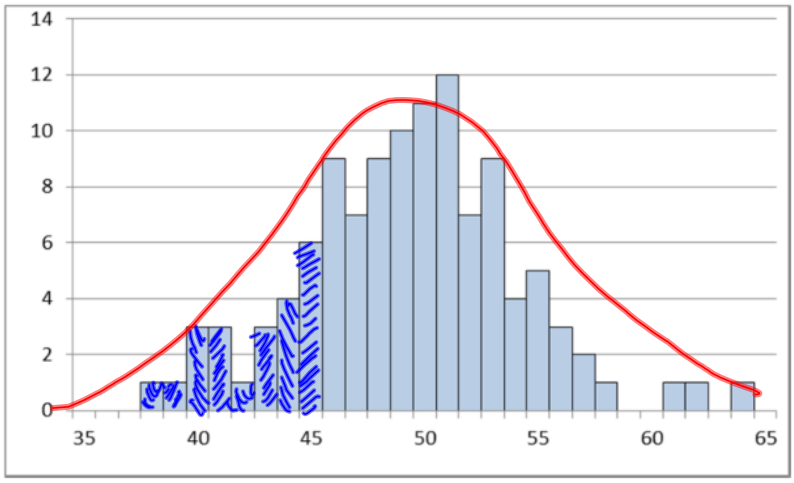
$s=\sqrt{\frac{1550}{4-1}}=\sqrt{\frac{1550}{3}}\approx 22.73$

5.8 Normal Distributions

When you have data for a single numerical variable you should

1. Graph the data as a histogram, stemplot or dotplot
2. Look at the shape, center and spread of the graph
3. Calculate numerical summary numbers such as the 5 number summary or the mean and standard deviation.
4. Is the distribution so regular that it could be described by a smooth curve? If so, is the curve bell shaped?

One hundred fair coins were flipped and the number of heads was counted. This experiment was repeated many times and the results were



114 times

6+ ... = 22

How many times were 45 or fewer heads observed?

(A) less than 8 (B) 8 (C) 16 (D) 22 (E) more than 22

What proportion of the time were 45 or fewer heads observed?

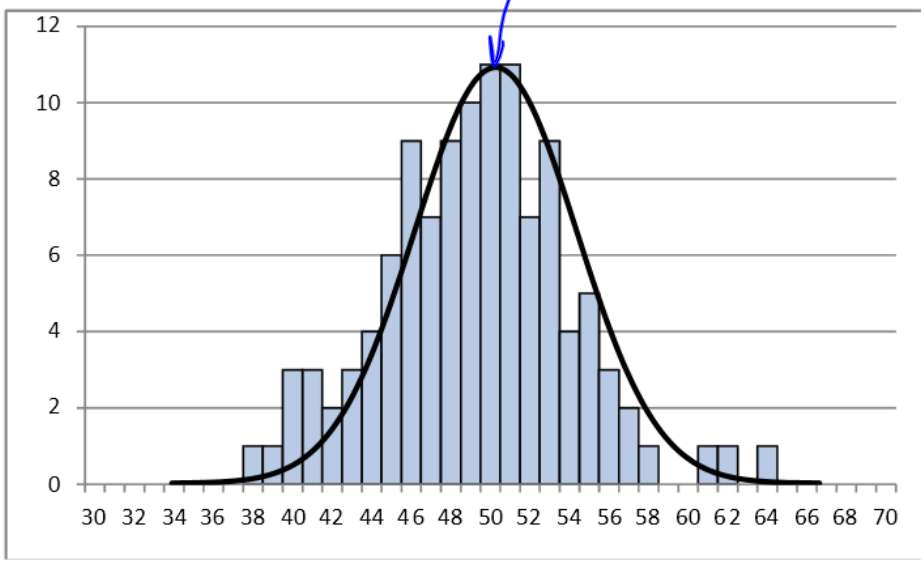
$P(X \leq 45) = \frac{22}{114}$

What proportion of the time were more than 50 heads observed?

$P \approx \frac{46}{114}$

Did you read your book yet?
A) Yes
B) No

Can we approximate this with a bell curve? The data has a mean of 50.28 and a standard deviation of 5.1. This generates a bell curve with the shape as shown.

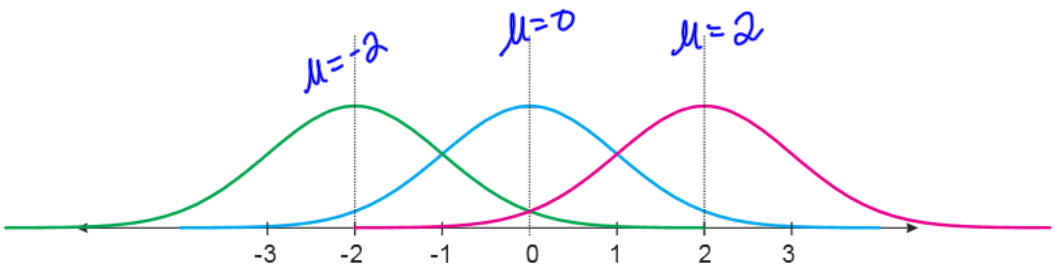


Can we use the smooth curve to find the probabilities? yes

$\bar{x}$ vs μ?

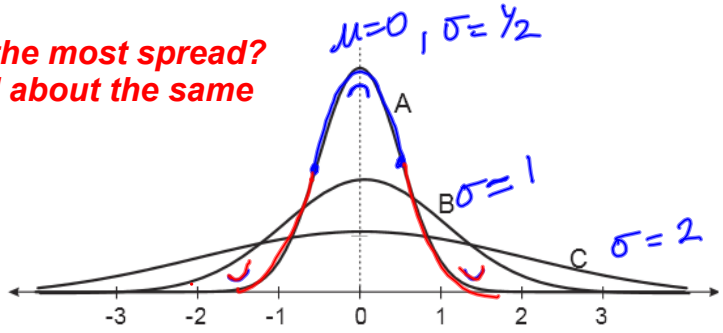
Some general information about bell curves.

1. Referred to as the NORMAL CURVE or NORMAL DISTRIBUTION
2. The location of the peak depends on where the mean is located. Typically use the symbol μ (mu) for the mean.

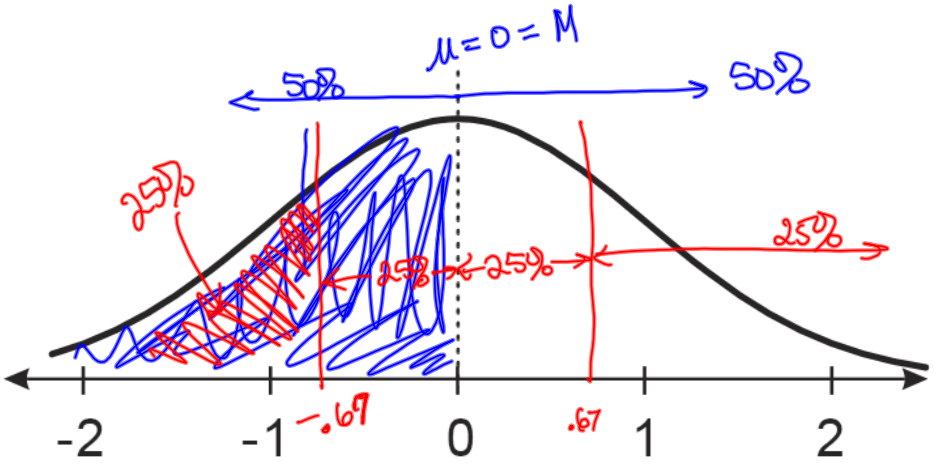


3. The spread is determined by the standard deviation. Typically use the symbol σ (sigma) for the standard deviation.

Which curve has the most spread?
(A) they are all about the same
(B) A
(C) B
(D) C
(E) Not sure



To find when you are one standard deviation away from the mean, look for where the curvature changes.

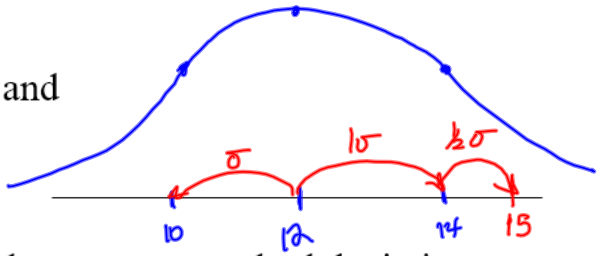


The first quartile is located about 0.67 standard deviations below the mean and the third quartile is located about 0.67 standard deviations above the mean.

EXAMPLE

A certain type of washing machines last an average of 12 years with a standard deviation of 2 years.

(a) Draw a normal curve with the mean and standard deviation located correctly.



(b) If a washing machine lasts 10 years, how many standard deviations below the mean is that washing machine? What if it lasted 15 years?

1 std dev below the mean

1.5 std dev above the mean

The ***standard score*** (or ***z-score***) of a measurement X is how many standard deviations (σ) the measurement is away from the mean (μ).

To calculate the standard score, Z, or to find X with a given Z value, use the formula

$$Z = \frac{X - \mu}{\sigma} \text{ or } X = \mu + \sigma Z$$

EXAMPLE

A normal distribution has a mean of 50 and a standard deviation of 8.

(a) Find the z-scores for the following values of X:

X = 42, Z = $\frac{42-50}{8} = -1$

X = 38, Z = $\frac{38-50}{8} = -1.5$

X = 60, Z = $\frac{60-50}{8} = 1.25 = \frac{5}{4}$

$X = 62$

$= 50 + 1.5 * 8$

$\mu = 50, \sigma = 8$

(b) If the z-score of a measurement is 1.5, what is the value of X?

(A) less than 50 ***(B) between 51 and 60***

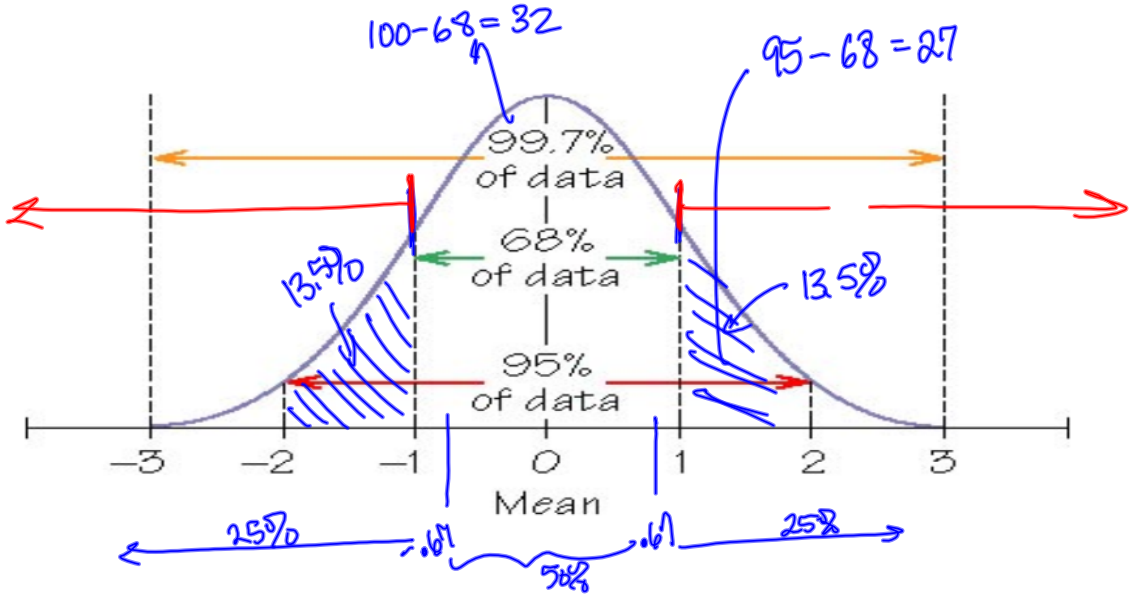
(C) between 61 and 70 ***(D) more than 70***

(E) Not sure

5.9 The 68-95-99.7 Rule

In any normal distribution,

- About 68% of the data is within one standard deviation of the mean
- About 95% of the data is within 2 standard deviations of the mean
- About 99.7% of the data is within 3 standard deviations of the mean.



EXAMPLE

The length of tape on a roll of a certain type of masking tape is normally distributed with a mean of 25 meters and a standard deviation of 50 centimeters.

(a) What is the range of lengths that most (99.7%) of the rolls?

23.5 m to 26.5 m

(b) What percent of the rolls are longer than 26 meters? 2.5%

(c) What lengths bracket the middle 50% of the rolls of tape?

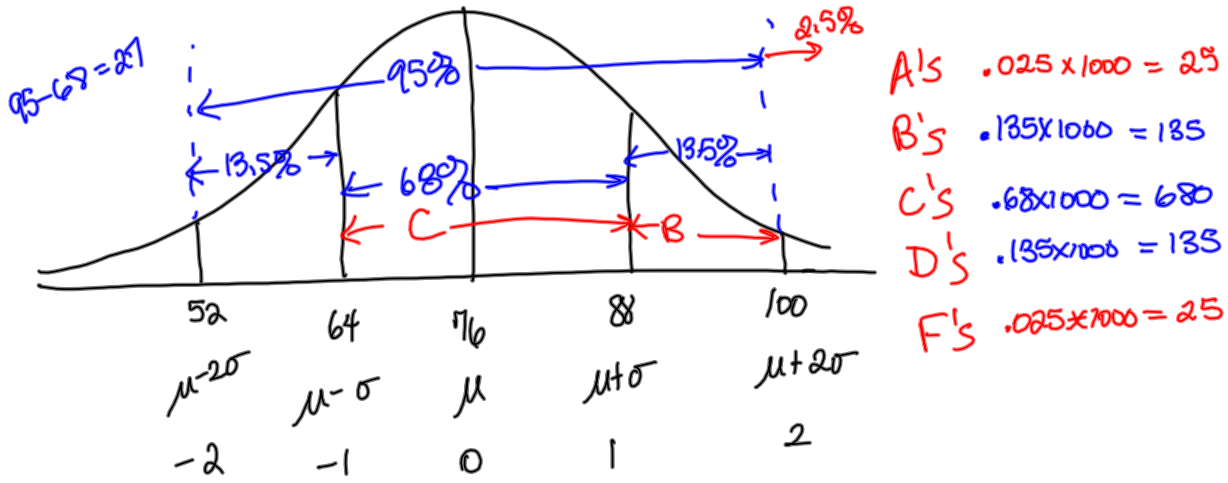
(A) 24.5 m to 25.5 m (B) 24.665 m to 25.335 m
(C) 24.33 m to 25.67m (D) not sure
(E) none of these

EXAMPLE

A class of 1000 students will be graded based on the normal curve with

- A grade of "C" assigned to the students within one standard deviation of the mean.
- A grade of "B" assigned to students who are between 1 and 2 standard deviations above the mean
- A grade of "A" assigned to students who are more than two standard deviations above the mean
- A grade of "D" assigned to students between 1 and 2 standard deviations below the mean
- A grade of "F" assigned to students more than two standard deviations below the mean.

How many students receive each grade?



If the mean grade in the class is 76 with a standard deviation of 12, what are the grade cutoffs?