CHAPTER 5 – EXPLORING DATA DISTRIBUTIONS

"RAW" 11, 10, 7, 11, 10, ...

A sample of jelly bean bags is chosen and the number of blue jelly beans in each bag is counted. The results are shown in the table below:

No. of bags	1	0	6	7	8
No. of blue jelly beans	7	8	9	10	11

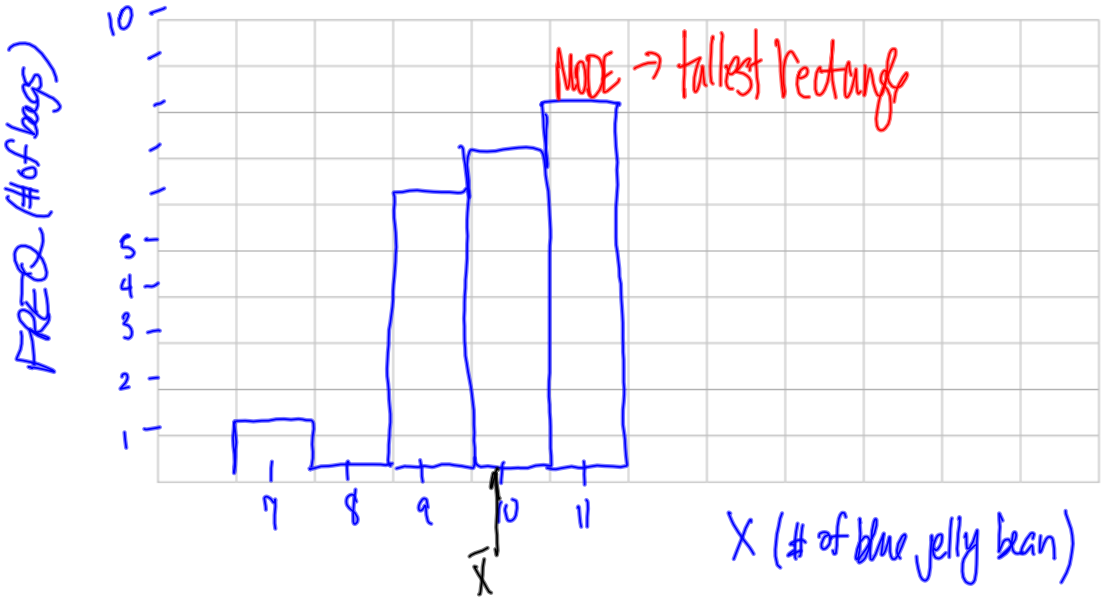
What are the individuals and what is the variable in the experiment above?

22

Individuals are the objects described by a set of data. These individuals may be people, animals or things. Bags of J. bean (freq)

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. # of blue (x)

Display this information in a ***histogram***:



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.

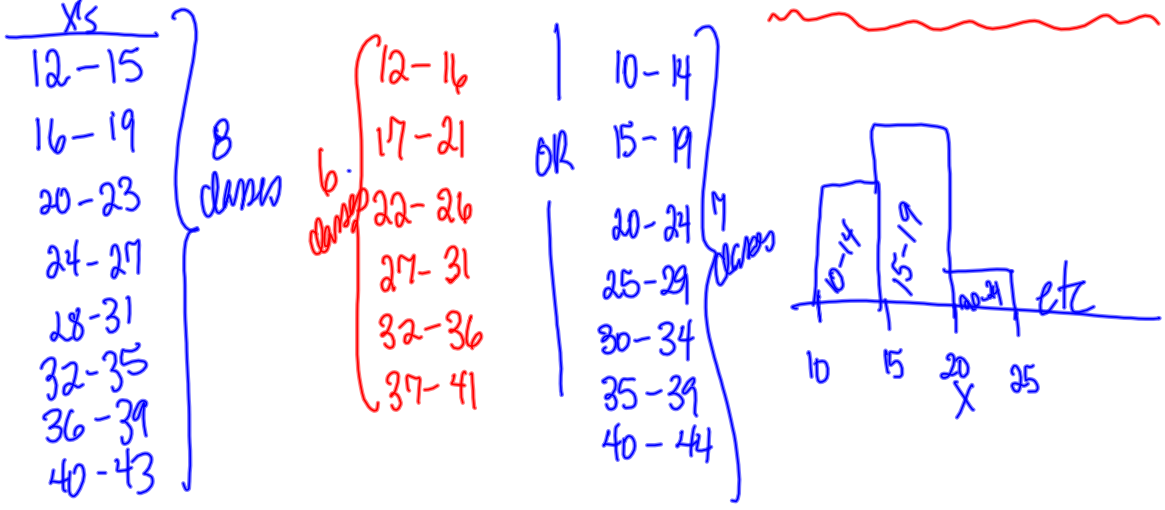
The ***distribution*** of a variable tells us what values the variable takes and how often it takes these values.

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.

Steps to creating a histogram

1. Choose the classes: Divide the range of the data into a reasonable number of classes of equal width. ≈ 8-12 ish
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.

If you have data that ranges from a low value of 12 to a high value of 40, what will the classes be if the class size is 4? What if the class size is 5?

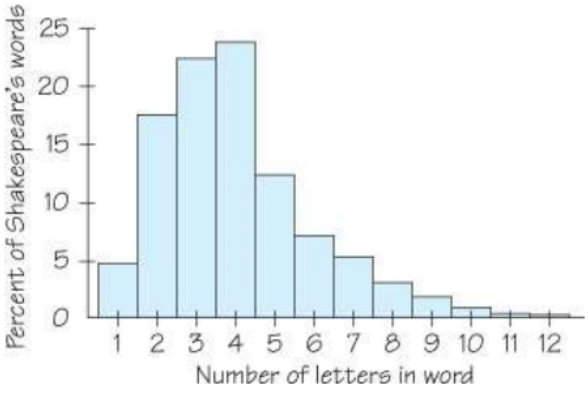


A graph is ***symmetric*** if the right and left sides of the histogram are approximately mirror images of each other.

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

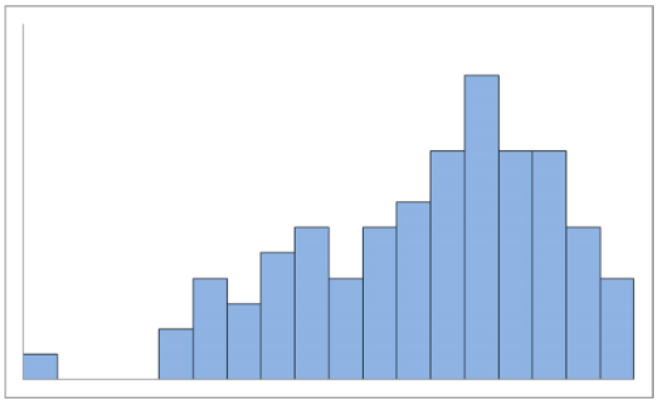
A graph ***is skewed to the left*** if the longer tail is on the left side. This is also called negatively skewed.

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



The histogram shows the lengths of words used in Shakespeare's plays as a percent of all the words in his plays. What is the overall shape of this distribution?

Skewed right, 1 peak, no outliers



Comment on the shape of the histogram on the left

Skewed left

one peak

may be an outlier

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

Arrange the data set below into a stemplot

2, 3, 8, 10, 11, 11, 12, 16, 16, 22, 23, ~~57~~ → 107

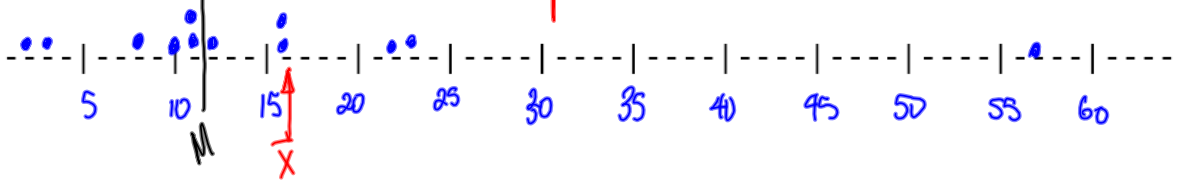
0	2 3 8
1	0 1 1 2 6 6
2	2 3
3	
4	
5	7

0	2 3 8
1	0 1 1 2 6 6
2	2 3

high

10	7

Display the data 2, 3, 8, 10, 11, 11, 12, 16, 16, 22, 23, 57 in a "dotplot."



Where would this plot "balance"? Where is the data cut in half? Which value occurred the most often?

$\bar{x} = \frac{2+3+\cdots+57}{12} = \frac{191}{12} \approx 15.9167$

$\frac{11+12}{2} = 11.5$

MODES

11 and 16

The **mean** of the n numbers $x_1, x_2, \ldots, x_n$ is $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

This is also the "balance" point for our dot plots and histograms.

The **median** of the n numbers $x_1, x_2, \ldots, x_n$ is the number in the middle when the n number are arranged in order of size and there are an odd number of values. When there are an even number of values, the median is the mean of the two middle numbers.

Half of our data is above the median and half is below.

The **mode** of the n numbers $x_1, x_2, \ldots, x_n$ is the number that occurs the most often. If no number occurs more often than any other number, there is no mode. If two numbers both occur the most often, then there are two modes.

What is the mean, median and mode for the number of blue jelly beans? Interpret these values in the context of the histogram.

No. of bags	1	0	6	7	8
No. of blue jelly beans	7	8	9	10	11

$$\text{mean} = \bar{x} = \frac{7 + \underbrace{9 + 9 + \ldots + 9}_{6 \text{ of these}} + \underbrace{10 + 10 + \ldots 10}_{7} + \underbrace{11 + \ldots + 11}_{8}}{22} = 9.9545$$

where the histogram balances

Median 7, 9, 9, ... 9, 10 ... | 10, 10, ... 11 ...

11 values | M | 11 values

$\Rightarrow$ cuts the area in half on the histogram

MODE is 11 (occurred most often) $\Rightarrow$ tallest rectangle (s)

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,

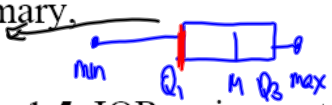
1. Use the median to split the data set into two halves – an upper half and a lower half.
2. The ***first quartile*** Q_1 is the median of the lower half.
3. The ***third quartile*** Q_3 is the median of the upper half.

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

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.



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

Find the range, the quartiles, the IQR and outliers (if any) for the data below. Display the data in a boxplot.

lower M upper

2, 3, 8, | 10, 11, 11, | 12, 16, 16, | 22, 23, 57

range = 57 − 2 = 55

IQR = 19 − 9 = 10

$Q_1 = \frac{10+8}{2} = 9$ $\frac{11+12}{2} = 11.5$ $Q_3 = \frac{22+16}{2} = 19$

low outlier?

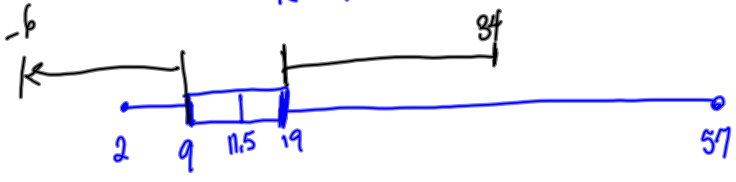
$Q_1 - 1.5 \times IQR$

$= 9 - 1.5 \times 10 = -6$

high outlier?

$Q_3 + 1.5 \times IQR$

$= 19 + 1.5 \times 10 = 34$



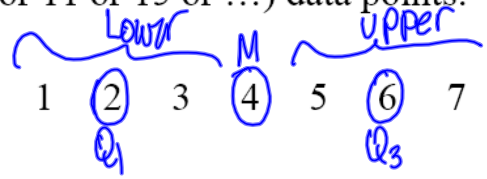
IQR

since 57 > 34

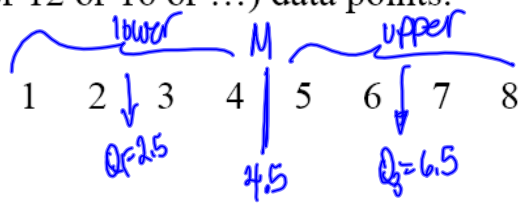
57 is a high outlier

There are 4 different cases when finding the quartiles:

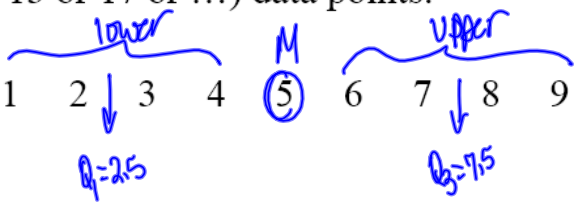
7 (or 11 or 15 or …) data points:



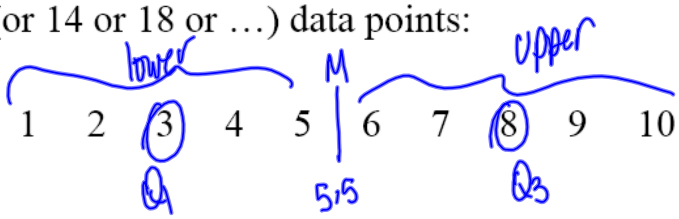
8 (or 12 or 16 or …) data points:



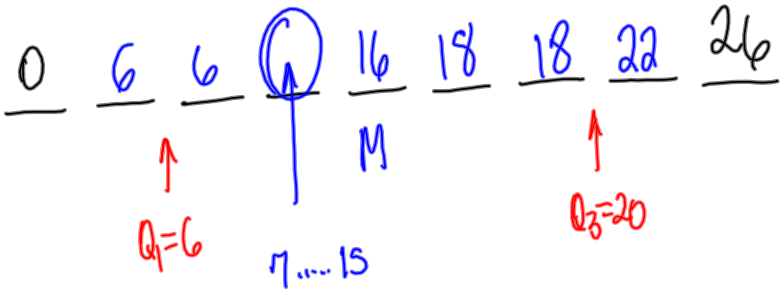
9 (or 13 or 17 or …) data points:



10 (or 14 or 18 or …) data points:



The boxplot below is for the number of grams of whole grain in a serving of 9 different cereals. If there are two modes, 6 and 18, what is a possible data set?



Another way to measure spread is the standard deviation, S. This will determine, on "average", how far away are the measurements from the mean. To find the standard deviation in a set of n numbers,

1. Find the mean of the numbers, $\bar{x}$.
2. Subtract the mean from each of the values, $x-\bar{x}$. Note that the sum of these values must be zero.
3. Square the result of the subtractions, $\left(x-\bar{x}\right)^2$
4. Add the squares up, $\sum\left(x-\bar{x}\right)^2$
5. Divide by $n-1$ and then take the square root, $S=\sqrt{\dfrac{\sum\left(x-\bar{x}\right)^2}{n-1}}$

Find the standard deviation in the golf scores for Amy, Bart, and Caleb. Whose scores vary the least?

Amy	67	71	66		
Bart	70	72	80	66	
Caleb	79	78	80	82	81

$\bar{x}=(67+71+66)/3=68$

$\bar{x}=(70+72+80+66)/4=72$

vanes least

Caleb : $\sqrt{2.5}\approx 1.5811$

Amy:

x	$x-\bar{x}$	$(x-\bar{x})^2$
67	$67-68=-1$	$(-1)^2=1$
71	$71-68=3$	$3^2=9$
66	$66-68=-2$	$(-2)^2=4$
sum	0	14

$S=\sqrt{\frac{14}{3-1}}=\sqrt{\frac{14}{2}}=\sqrt{7}$

≈ 2.6458

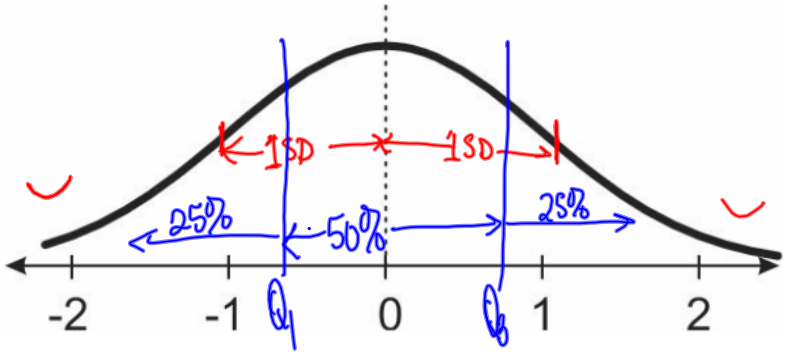
Bart

x	$x-\bar{x}$	$(x-\bar{x})^2$
70	-2	4
72	0	0
80	8	64
66	-6	36
sum	0	104

$S=\sqrt{\frac{104}{4-1}}=\sqrt{\frac{104}{3}}$

≈ 5.8878

Many natural and social phenomena produce a continuous distribution with a bell-shaped curve.



Every bell-shaped (NORMAL) curve has the following properties:

- Its peak occurs directly above the mean
- The curve is symmetric about a vertical line through the mean. The curve never touches the x-axis. It extends indefinitely in both directions.
- The area between the curve and the x-axis is always 100% as that is the total probability that something is observed.
- The shape of the curve is completely determined by the mean and the standard deviation.
- The place where the curve changes from concave down [☹]to concave up [☺] is one standard deviation away from the mean.
- The first and third quartiles are located 0.67 standard deviations away from the mean.
- The probability that a data value will fall between two values a and b is given by the area under the curve between a and b.
- The standard normal curve has a mean of 0, standard deviation of 1 and uses the letter Z for the variable.

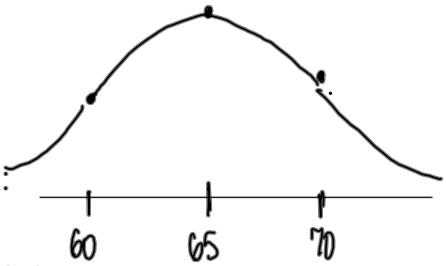
To find the z-score (how many standard deviations from the mean) for a value X, use the formula

$$Z = \frac{X - \mu}{\sigma}$$

If you are given the z-score, the X value is $X = \mu + \sigma Z$

The weight of an egg is normally distributed with a mean weight of 65 grams with a standard deviation of 5 grams.

(a) Sketch th normal curve for this distribution:



(b) Determine the z-score for the following weights:

70 grams $Z = \frac{70-65}{5} = 1$

79 grams $Z = \frac{79-65}{5} = 2.8$

63 grams $Z = \frac{63-65}{5} = -.4$

fyi μ is mean
σ is std dev

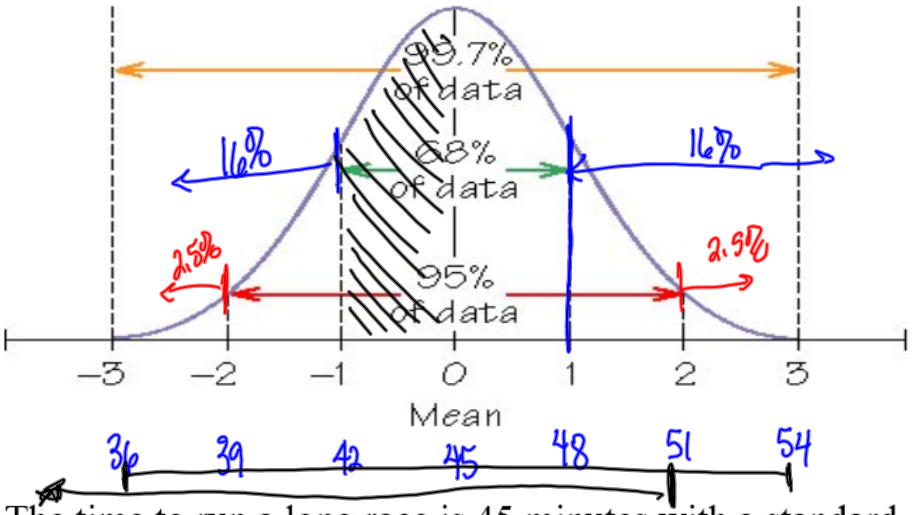
(c) Where are the first and third quartiles located?

$Q_1 = \mu - .67\sigma = 65 - .67(5) = 61.65$ GRAMS

$Q_3 = \mu + .67\sigma = 65 + .67(5) = 68.35$ GRAMS

In any normal distribution,

- About 68% of the data is within 1 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.



The time to run a long race is 45 minutes with a standard deviation of 3 minutes.

(a) What is the probability that a racer takes longer than 48 minutes?

100 − 68 = 32% not between 42 & 48 | $\frac{32}{2}$ = 16%

16%

(b) What is the probability that a racer takes less than 51 minutes?

100 − 2.5 = 97.5% or 95% + 2.5%

(c) What is the probability that a racer takes between 42 and 45 minutes?

34%

(d) If there are 1000 runners, how many will finish between 36 and 54 minutes? .997 * 1000 = 997

SAMPLE EXAM QUESTIONS FROM CHAPTER 5

A bag of chocolate chip cookies is opened and the number of chocolate chips in each cookie is counted. The results are

Q_1 Q_3

17 18 19 19 20 20 | 21 22 22 23 25 25 (29) 30 30 30 31 31 31 | 32 33 36 37 38 39

(a) Find the following quantities:

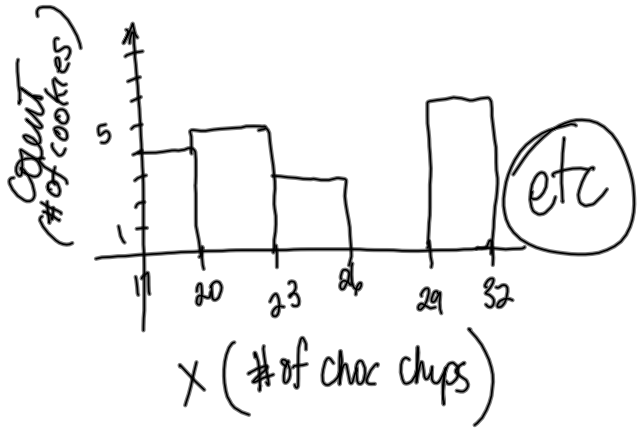
Mean = $\frac{17+18+\dots+39}{25} = 27.12$ Mode = 30 and 31

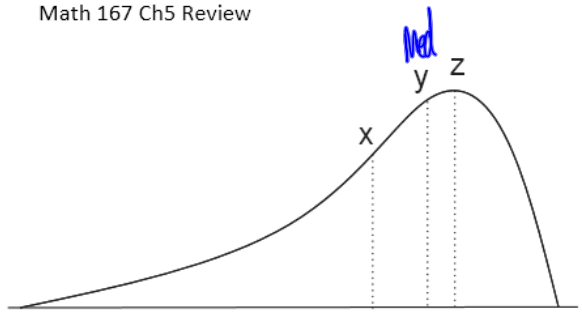
Median = 29 Range = 39 − 17 = 22

Q_1 = 20.5 Q_3 = 31.5

(b) Display the information in a histogram with class size of 3

class	count
17-19	4
20-22	5
23-25	3
26-28	0
29-31	7





The distribution above is

(A) skewed right (B) symmetric (C) skewed left

If y is the median of this distribution, which value is the mean?

(A) x (B) z (C) more information is needed

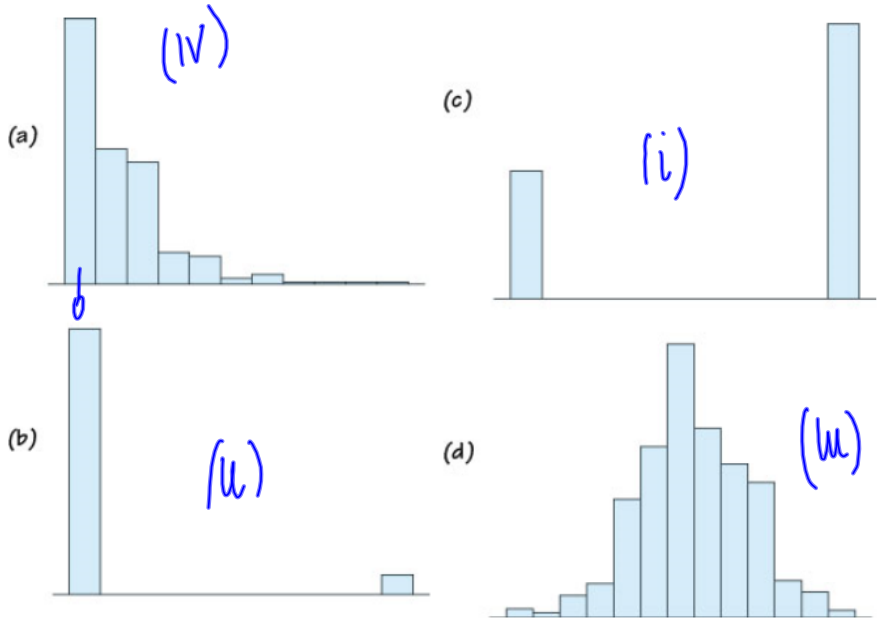
Match the distributions to the survey questions

(i) Are you male (0) or female (1)?

(ii) Are you right (0) or left (1) handed?

(iii) What is your height to the nearest inch?

(iv) How many minutes do you study on Friday evening?



The data below is the number of cases of xyz disease reported in a county each month recently. Arrange this in a stemplot.

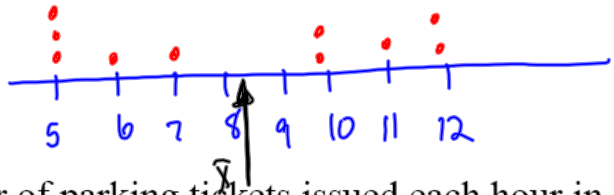
91 115 115 125 127 129 130 131 140

9	1
10	
11	5 5
12	5 7 9
13	0 1
14	0

Display the data below in a dot plot and indicate where the balance point is located.

6 12 11 7 5 5 5 10 12 10

$\bar{x} = \frac{6+12+\cdots+10}{10} = \frac{83}{10} = 8.3$



The number of parking tickets issued each hour in a parking garage is given below. What is the standard deviation in the number of parking tickets issued?

5 4 7 10 4

$\bar{x} = \frac{5+4+7+10+4}{5} = 6$

x	$x-\bar{x}$	$(x-\bar{x})^2$
5	-1	1
4	-2	4
7	1	1
10	4	16
4	-2	4
Sum	0 ✓	26

$s = \sqrt{\frac{26}{5-1}} = \sqrt{\frac{26}{4}} = \sqrt{6.5}$

≈ 2.5495

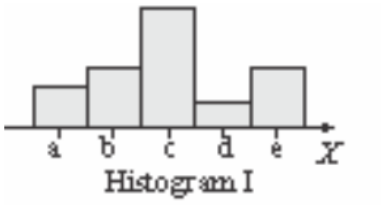
Which histogram has the larger standard deviation?

(A) Histogram I

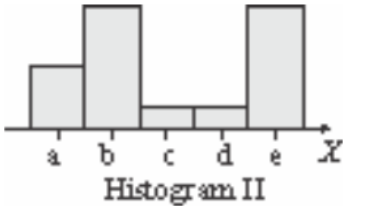
(B) Histogram II

(C) They are nearly the same

(D) There is not enough information to decide



Histogram I



Histogram II

IQR = 49 − 33 = 16

A data distribution has a minimum value of 7, $Q_1 = 33$, $M = 42$, $Q_3 = 49$, and the maximum value is 66. Which statements below (if any) are true about this distribution?

(A) There are no outliers

(B) There is at least one low outlier

(C) There is at least one high outlier

(D) There is not enough information to determine if there are outliers.

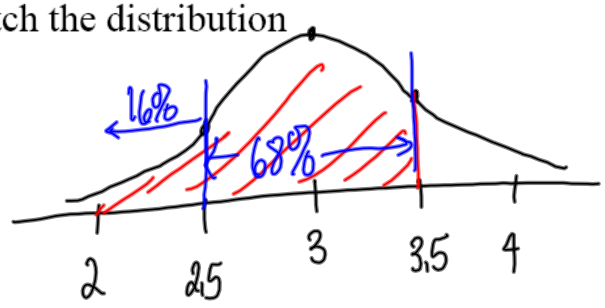
(E) None of these

Low one? $Q_1 - 1.5 * IQR = 33 - 1.5 * 16 = 9$ ⇒ 7 is less than 9, so a low one

high one? $Q_3 + 1.5 * IQR = 49 + 1.5 * 16 = 73$ ⇒ since $66 < 73$ no high ones

The length of crayons in a box of used crayons is normally distributed with a mean of 3 inches and a standard deviation of 0.5 inches.

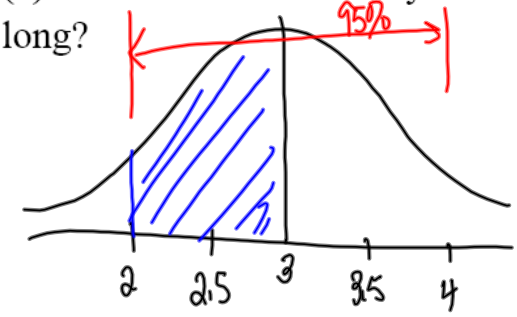
(a) Sketch the distribution



(b) What is the probability that a crayon is shorter than 3.5 inches.

68% + 16% = 84%

(c) If the box has 200 crayons in it, how many are between 2 and 3 inches long?



$\frac{95\%}{2} = 47.5\%$

A normal curve has a mean of 80 and a standard deviation of 10. Where is the 1st quartile located?

(A) need more information to determine. (B) none of these

(C) 77.5 (D) 73.3 (E) 70.0

$Q_1 = \mu - .67\sigma = 80 - .67(10) = 73.3$