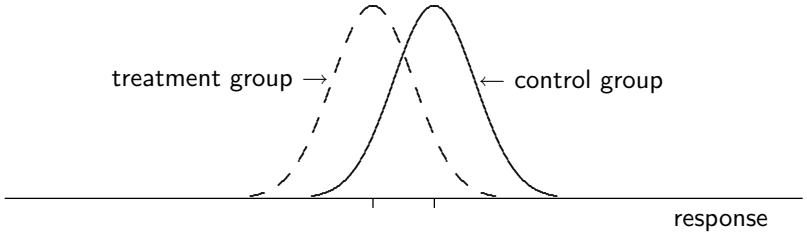
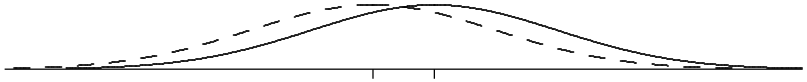
MAKING COMPARISONS

• Suppose some study turns out this way:



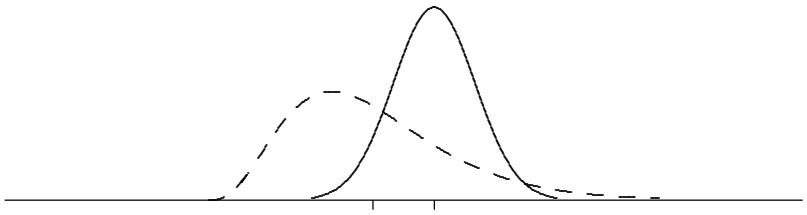
It would natural to measure the effect of the treatment by the change in _____.

• How are the results of this next study similar to those of the previous one? How are they different?



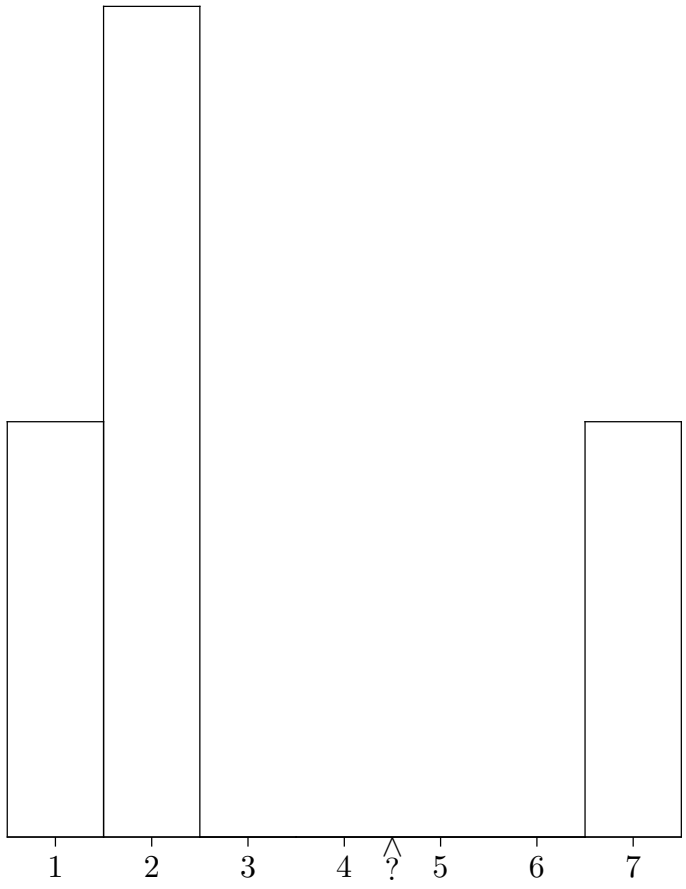
The greater the _____ of the distributions, the less important is a given change in location.

• Measures of location and spread are important summaries of a distribution. But watch out for situations where they don't tell the whole story:



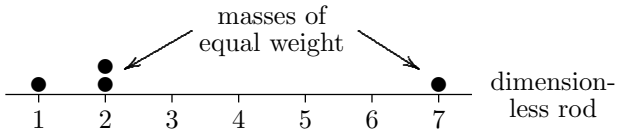
A BALANCING ACT

• At what value on the horizontal axis would you pivot this histogram in order to get it to balance?



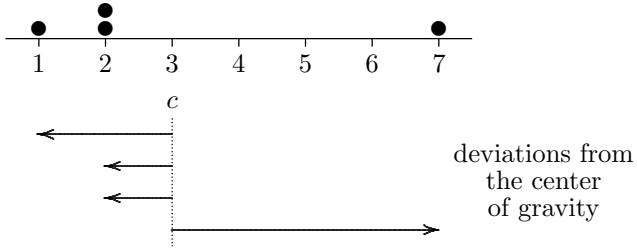
A PROBLEM FROM PHYSICS

• Consider the following physical system:



What is the center of gravity of this system, i.e., the point c at which the rod would balance?

- By experiment: _____
- By theory: _____

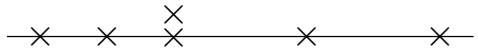


$$\begin{aligned} 0 &= \text{sum of deviations} \\ &= (1-c)+(2-c)+(2-c)+(7-c) \\ &= (1+2+2+7)-4c \end{aligned}$$

so $c =$ — =_____.

MEASURING LOCATION: THE AVERAGE

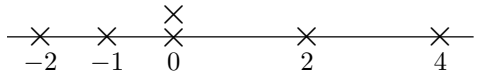
• In general, if a list of numbers is marked off on a dimensionalless line



then the center of gravity c of the $\times$'s is

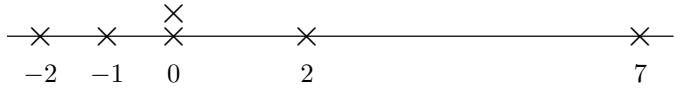
$$c = \frac{\text{Sum of the values}}{\text{How many there are}} = \begin{matrix}\text{Average of}\\ \text{the list}\end{matrix} = \text{The Mean.}$$

• What is the average of the list -2, -1, 0, 0, 2, 4?



- By eye: _____
- By calculator: $\dfrac{-2-1+0+0+2+4}{6} = \dfrac{3}{6} = \dfrac{1}{2}$

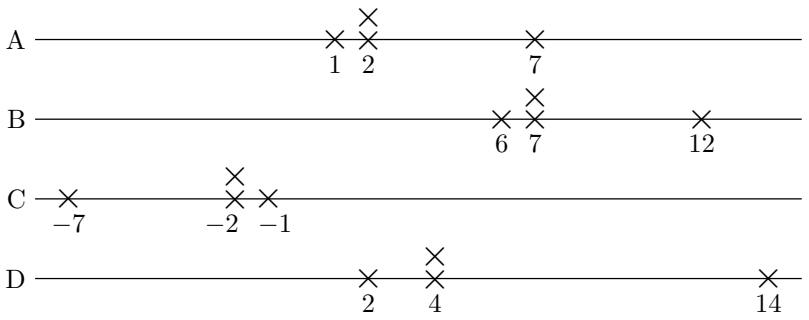
• What is the average of the list -2, -1, 0, 0, 2, 7?



• If you need the exact value of an average, use a calculator. If a rough approximation will suffice, guess the center of gravity by eye.

HOW IS THE AVERAGE EFFECTED BY CHANGES OF LOCATION AND SCALE?

- Consider these four lists:



  - How is list B related to list A? How is the average of list B related to the average of list A?

    The numbers shifted by 5, and so did their average.

  - Ditto, for C?

    The numbers changed sign, and so did their average.

  - Ditto, for D?

    The numbers doubled, and so did their average.

- What is the general rule?

  - If each element of a list is shifted by some constant, the average will be _____ by that constant.

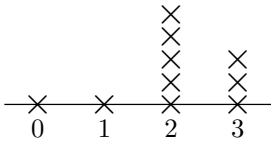
  - If each element of a list is multiplied by some constant, the average will be _____ by that constant.

FROM LISTS TO HISTOGRAMS

- An instructor gives a quiz with 3 questions, each worth 1 point.

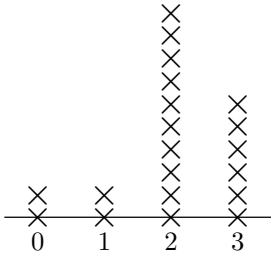
30%	of the class scores	3
50%	"	2
10%	"	1
10%	"	0

  - If there are 10 people in the class, what is the average score?



Average =

  - Ditto, for 20 people?



Average =

  - Can you figure the average score without being told the number of people in the class?

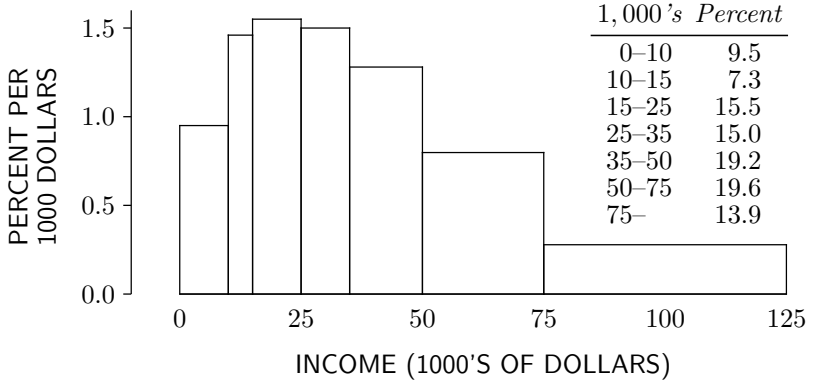
    Yes. All you need are the percentages.

- Can you figure the average of a list if the only information available is a histogram of the numbers in that list?

  Yes. The histogram gives you the percentages.

THE AVERAGE AND THE HISTOGRAM

• The following histogram shows the distribution of family income in the U.S. in 1992:†



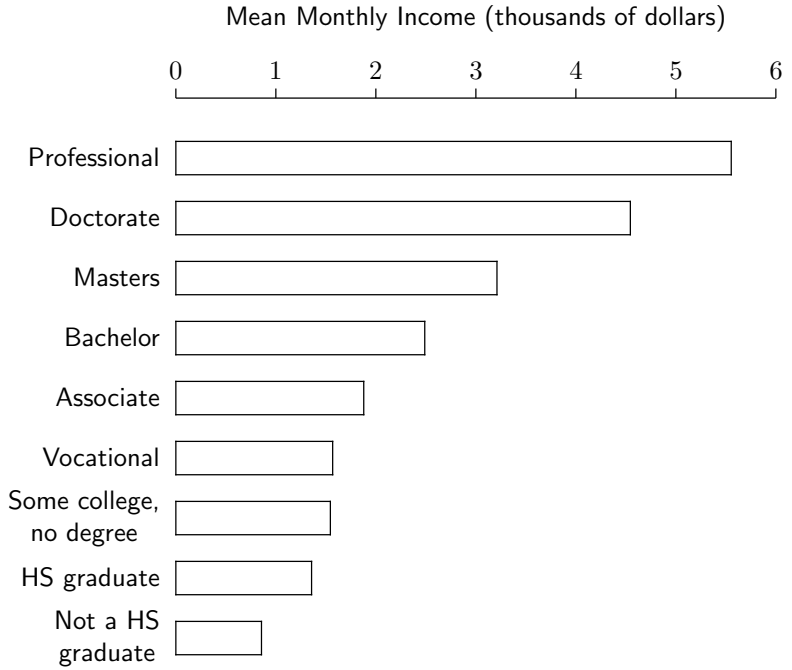
1,000's	Percent
0–10	9.5
10–15	7.3
15–25	15.5
25–35	15.0
35–50	19.2
50–75	19.6
75–	13.9

- What was the average family income in 1992?
  - By eye, the center of gravity is _____
  - By calculator: $43,300
- What was the median (50^{th}-percentile) for family income in 1992?
  - By eye: _____
  - By calculator: $37,100
- Why is the average bigger than the median?
  The long right tail pulls the average (the center of gravity) to the right.
- Which number, the average or the median, do you prefer as a summary of the distribution of family income?

† Source: Statistical Abstract of the United States; `http://www.medaccess.com/census/94s0714.htm`

HOW MUCH MONEY CAN YOU EXPECT TO MAKE?

• The following chart shows Mean Monthly Income by Highest Degree Earned, for people aged 18 and over in the US in 1990:†



- What would you like to know besides the averages?

† Source: Statistical Abstract of the United States 1994, (`http://www.medaccess.com/census/94s0235.htm`); based on the Survey of Income and Programs Participation.

MEASURING SPREAD: THE SD

• What are the steps in computing the SD (standard deviation) of the list 1, 2, 2, 7?

1 Find the average of the list.
2 Find the deviations from average.
3 Square the deviations.
4 Average the squared deviations.
5 Take the square root.

• Do it:

1 Average of the list = 3

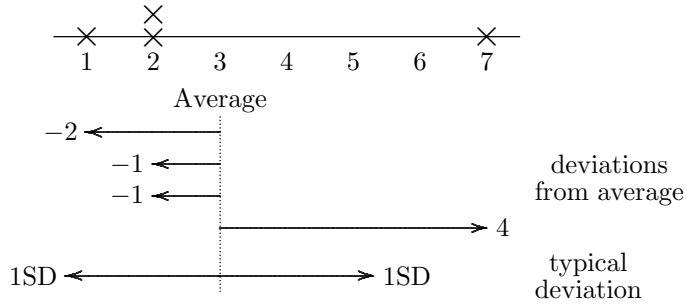
2 Deviations from the average = -2, -1, -1, 4

3 Squared deviations = 4, 1, 1, 16

4 Average square = $\dfrac{4+1+1+16}{4} = \dfrac{22}{4}$

5 SD = $\sqrt{22/4} \approx 2.3$

• Picture:



THE SD, CONTINUED

• For any list, the

$$\text{SD} = \begin{array}{c}\text{Root Mean Square (RMS) of the amounts} \\ \text{off from the average of the list}\end{array}$$

is a common measure of the spread of the numbers on the list.

• The SD says how far away the numbers on a list are from their average. Most entries on the list will be somewhere around 1 SD away from the average. Very few will be 2 or 3 SDs away.

• Each of the following lists has an average of 50. Which list has the smallest SD? The largest?

A: 0, 20, 40, 50, 60, 80, 100

B: 0, 48, 49, 50, 51, 52, 100

C: 0, 1, 2, 50, 98, 99, 100

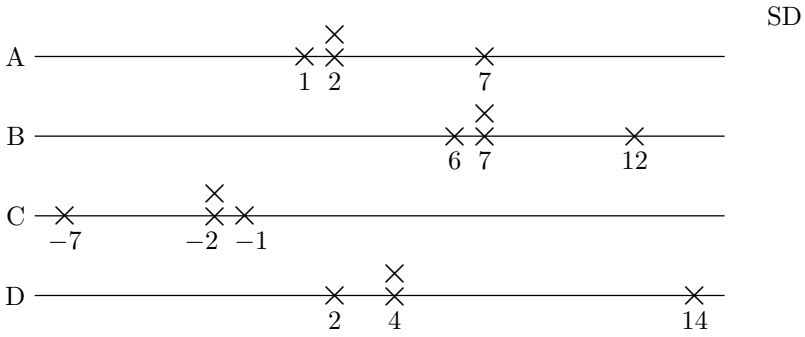
• What are the SDs?

	A	B	C
By eye			
By calculator	32	27	45

List B shows that the SD pays considerable attention to large deviations.

HOW IS THE SD EFFECTED BY CHANGES OF LOCATION AND SCALE?

- Consider these four lists:



  - How is the SD list B related to that of list A?

    The two lists have the same spread, and the same SD.

  - Ditto, for C?

    The two lists have the same spread, and the same SD.

  - Ditto, for D?

    The spread doubled, and so did the SD.

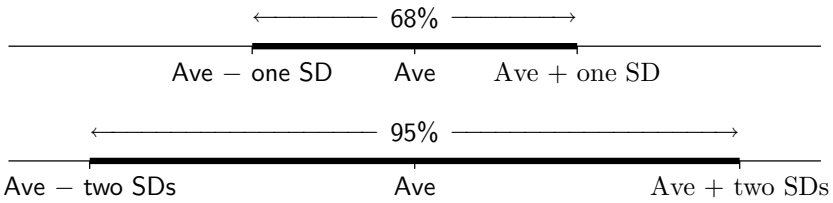
- What is the general rule?

  - If each element of a list is shifted by some constant, the SD is _____.

  - If each element of a list is multiplied by some constant, the SD gets _____ by the _____ value of that constant.

THE 68%- AND 95%-RULES

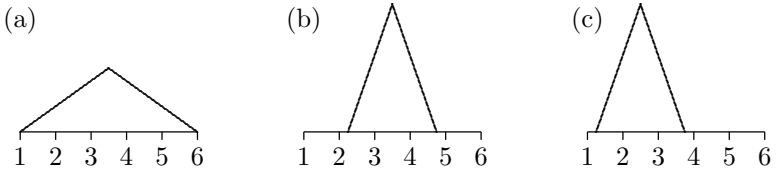
- Roughly 68% of the entries on a list (two in three) are with one SD of the average, the other 32% are further away. Roughly 95% (19 in 20) of the entries on a list are within two SDs of the average; the other 5% are further away. This is so for many lists, but not all.



  - These rules work especially well for lists whose histograms are bell-shaped.

- Below are sketches of histograms for three lists. Match the sketch with the description. Some descriptions will be left over. Give your reasoning in each case.

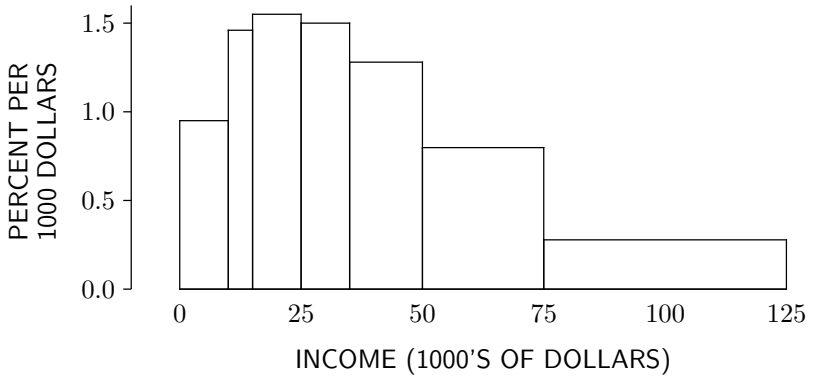
(i) ave ≈ 3.5, SD ≈ 1
(ii) ave ≈ 3.5, SD ≈ 0.5
(iii) ave ≈ 3.5, SD ≈ 2
(iv) ave ≈ 2.5, SD ≈ 1
(v) ave ≈ 2.5, SD ≈ 0.5
(vi) ave ≈ 4.5, SD ≈ 0.5



(a): The average is 3.5. About 95% of the entries lie between 1.5 and 5.5, a range of 4 units. By the 95% rule, this is a range of about 4 SDs. So the SD is about 1. (The SD of the histogram is actually 1.02.)

FAMILY INCOME, REVISITED

- Recall the distribution of family income in the U.S. in 1992:



- The SD of family income is around

  $15,000 $30,000 $45,000

Pick one option, and explain.

The range from two SDs above the average to two SDs below the average covers about 95% of incomes, or about $120,000. So the SD is about $120,000/4 = $30,000. (The SD of the histogram is actually $29,750.)

MATHEMATICAL NOTATION

- In mathematical notation, a generic list can represented as

$$x_1,\, x_2,\, \ldots,\, x_n.$$

Here x_1 stands for the first element on the list, x_2 for the second, $\ldots$, and x_n for the last; n stands for the number of elements on the list.

- The average of the generic list is

$$\frac{\text{sum of the values}}{\text{how many there are}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Here $\sum_{i=1}^{n} x_i$ stands for the sum of x_i as the index i ranges from 1 to n. $\frac{1}{n}\sum_{i=1}^{n} x_i$ is commonly denoted by $\bar{x}$, read "x bar".

- The standard deviation of the generic list is the RMS deviation from average:

$$\begin{aligned}\text{SD} &= \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n}} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\,.\end{aligned}$$

Most statistical calculators and computer programs compute the standard deviation as the slightly larger number

$$\text{SD}^{+} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\,,$$

which is commonly denoted by s. Thus

$$\text{SD} = \sqrt{\frac{n-1}{n}}\,s = \sqrt{\frac{\text{number of entries} - \text{one}}{\text{number of entries}}}\,\text{SD}^{+}.$$