
Guess at the **densities** (the height of each bar), then multiply by the base to find the area = percent of values contained in that bar.

<table>
<thead>
<tr>
<th>Limits</th>
<th>Length of base</th>
<th>Density</th>
<th>Percent in bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>90-100</td>
<td>10</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td>100-105</td>
<td>5</td>
<td>1.6</td>
<td>8</td>
</tr>
<tr>
<td>105-110</td>
<td>5</td>
<td>2.0</td>
<td>10</td>
</tr>
<tr>
<td>110-115</td>
<td>5</td>
<td>3.0</td>
<td>15</td>
</tr>
<tr>
<td>115-120</td>
<td>5</td>
<td>3.2</td>
<td>16</td>
</tr>
<tr>
<td>120-125</td>
<td>5</td>
<td>2.2</td>
<td>11</td>
</tr>
<tr>
<td>125-130</td>
<td>5</td>
<td>2.1</td>
<td>10.5</td>
</tr>
<tr>
<td>130-135</td>
<td>5</td>
<td>1.8</td>
<td>9</td>
</tr>
<tr>
<td>135-140</td>
<td>5</td>
<td>1.0</td>
<td>5</td>
</tr>
<tr>
<td>140-150</td>
<td>10</td>
<td>0.8</td>
<td>8</td>
</tr>
<tr>
<td>150-160</td>
<td>10</td>
<td>0.25</td>
<td>2.5</td>
</tr>
</tbody>
</table>

a. Percent of women with blood pressure over 130 = 9 + 5 + 8 + 2.5 = 24.5 (area of last 4 bars)
b. One hundred percent of data is between 90 and 160.
c. The **bar** for the interval 135-40 is higher than for 140-150, but the **area** of the interval 140-150 is greater, and the **areas show the percentage**.
d. The heights of the bars show the density, and **density shows how crowded each unit interval is**.
e. Base of interval 125-130 mm = 130 - 125 = 5; density = 2.1 % per mm.; hence percent in bar = 5 * 2.1 = 10.5 %
f. The interval 97-98 mm has about 0.5 % of the population (density = population per unit of the x-axis, here in mm); the interval 102-103 mm has about 1.6 % of the population.
g. The most crowded millimeter of all will lie between 115 and 120 mm. We cannot tell from our data whether that is 115-116 or 116-117 or 117-118 or 118-119 or 119-120; all are shown as equally dense or crowded.
Problem 5. One block of a family income histogram

The question is simple if you understand density:
Since the vertical axis shows the density, in this case PERCENT PER $ 1000 interval of the X-axis, we know that 50 percent of the population make between $ 50,000 and $ 100,000 per year:

\[ \text{Height} \times \text{base} = \frac{1\ \text{percent}}{\text{thousand}} \times (100\ \text{thousand} - 50\ \text{thousand}) \]

\[ = 50\ \text{percent} \]

Note that the "thousands" will cancel out: percent / thousand * thousand = percent.

We can also see that the distribution is uniform, so the same percentage must be in any equal-sized slice. Since the slice we are interested in is 90,000 to 100,000, this means that 10 percent of the total population of the wealthy suburb is in this slice.
Problem 8. Income distribution "histogram"? NO.

Begin with part (c). The graph as shown in the text is NOT a histogram, because:
-- heights do not represent densities, but percentages.
-- areas do not represent anything meaningful.

Drawing the histogram properly will allow one to see that:

(a) is TRUE. Families that earn between $10 K and $35 K are ARE spread fairly evenly over that range. The bar representing 10 - 15 K has a base of 5 and a height which SHOULD be \( \frac{7.3}{5} = 1.46\% \) per $1000. The bar representing 15 - 25 K has a base of 10 and a height which SHOULD be \( \frac{15.6}{10} = 1.56\% \) per $1000. The bar representing 25 - 35 K has a base of 10 and a height which SHOULD be \( \frac{15.0}{10} = 1.50\% \) per $1000. Since the densities are very close, the percent of families earning $10-11 K is very close to the percentage of families earning 34 - 35 K, and so on for any $1,000 interval in between. The tallest three bars on the accurately drawn histogram below represent these three families.

(b) is FALSE. Although there are similar percentages in total in families making between $35 - $50 K, and between $50 - 75 K, the DENSITIES differ greatly: the bar representing $35 - 50 K has a base of 15 and a height which SHOULD be \( \frac{19.2}{15} = 1.28\% \) per $1,000. the bar representing $50 - 75 K has a base of 25 and a height which SHOULD be \( \frac{19.6}{25} = 0.78\% \) per $1,000.

Code to draw histogram: 

```lisp
> (bind bounds (list 0 5 10 15 25 35 50 75 100 200))
> (bind pcts (list 3.7 5.8 7.3 15.6 15.0 19.2 19.6 7.7 6.2))
Since (sum pcts) = 100.1, we will have to let the computer adjust the data.
> (hist pcts bounds table), followed by (hist newpcts bounds table)
```

```
Income Distribution
U.S., 1992
```

![Histogram of Income Distribution](image)
Chapter 4, FPP, p. 74-76
Review Exercises
Average and Standard Deviation

Problem 2. Which list has the smaller SD? Explain.
Information: Both lists have a mean of 50.

Part a. \[ X = (50 \ 40 \ 60 \ 30 \ 70 \ 25 \ 75) \]
\[ Y = (50 \ 40 \ 60 \ 30 \ 70 \ 25 \ 75 \ 50 \ 50 \ 50) \]

Let the sum of squared deviations of \( X \) be SSD\( x \).

We know that SSD\( y = SSD\) \( x \) since the three added values to the Y list are all equal to the mean of Y, and \[ 3 \times (square (50 - 50)) = 0. \]

Now the VARIANCE of \( X = \frac{SSD\( x \)}{7} \), and the VARIANCE of \( Y = \frac{SSD\( y \)}{10} = \frac{SSD\( x \)}{10}. \)
Since the numerator is the same for both, and the denominator bigger for the variance of \( Y \), the variance of \( Y \) must be smaller.

If you want the exercise, \( SSD\( x = 2250 \), and \( SD\( x = 17.9284; SSD\( y = 2250 \), and \( SD\( y = 15. \) \)

Part b. \[ X = (50 \ 40 \ 60 \ 30 \ 70 \ 25 \ 75) \] [the same as list \( X \) in part a]
\[ Z = (50 \ 40 \ 60 \ 30 \ 70 \ 25 \ 75 \ 99 \ 1) \]
[two extreme values, which will add \[ (square (99 - 50)) + (square (1 - 50)) = 2 \times (square 49) = 4802 \] to the SSD\( z \).

SSD\( z \) is more than twice SSD\( x \), and the divisor used in computing the variance is only slightly bigger than for \( Z \). SSD\( z \) will definitely be bigger. [Values: SSD\( z = 7052 \), Var. \( Z = 783.56 \), SD\( z = 27.99 \) ]

Problem 3. Guessing the mean and the SD of the list \( X = \)

\[
\begin{align*}
0.7 & \quad 1.6 & \quad 9.8 & \quad 3.2 & \quad 5.4 & \quad 0.8 & \quad 7.7 & \quad 6.3 & \quad 2.2 & \quad 4.1 \\
8.1 & \quad 6.5 & \quad 3.7 & \quad 0.6 & \quad 6.9 & \quad 9.9 & \quad 8.8 & \quad 3.1 & \quad 5.7 & \quad 9.1
\end{align*}
\]

Part a. Guess whether the average is closer to 1, 5, or 10.

Easy -- values range from close to zero or one \((0.7, 1.6, 0.8, 0.6)\) to close to 9 or 10 \((9.8, 9.9, 8.8, 9.1)\), so the average cannot be at zero or one or 9 or 10. Must be close to 5.

Part b. Guess whether the standard deviation is closer to 1, 3, or 6.

Harder, but note that in Part a, we found 8 numbers out of 20 (40 percent of the data) close to one extreme or the other. An interval from the mean to plus or minus 1 SD should cover SOME of these values if it covers 68 percent of the data. Hence, we want 5 +/- 1 SD to go below 2 and above 8. This would point to a SD of about 3.

You could also note that an interval of 5 +/- 6 would include ALL the data, which is more than a 1 SD interval normally does; an interval of 5 +/- 1 would contain only the points from 4 to 6 -- only 3.2 and 4.1 in the first row, and no points at all in the second row -- only 20 percent of the data.
Problem 5. Unusual blood pressure?

Given: Normal systolic blood pressure (for males 18-24) is 124 mm with SD of 14 mm.
Readings are not unusual if in the range 124 +/- 14 mm or 110 to 138; somewhat unusual if outside this range, but in a 2 SD range of 124 +/- 28 or 96 to 152; and very unusual if beyond this.

Hence the values of 80 and 210 mm are very unusual; those of 115 and 135 mm are not unusual from a statistical viewpoint (though your doctor might consider 135 borderline hypertension; the language used in the problem is statistical and not medical).

We could standardize these scores by subtracting the mean then dividing by the SD:

(80 mm - 124 mm) / 14 mm = -44 mm / 14 mm = -3.1429  
(note: the "mm" cancels out, leaving a standardized score a pure number).

Three standard deviations below the mean is very rare: we would expect only around half a percent of the population to have pressures that low or lower.

Problem 6. Sketches of histograms:

(i) Left skewed, with highest point at 75
(ii) Symmetric, with highest point at 50
(iii) Right skewed, with highest point at 25

Note that the median will be closer than the mean to the mode or highest point for skewed distributions. The left skewed distribution will have the lowest average, and the right skewed distribution the highest,
  a. Averages of 40, 50 and 60 are really not in scrambled order.
  b. Median < average for (iii); median = average for (ii) and median < average for (i).
  c. SD for histogram (ii) [I know the text says (iii), but SD of (ii) should be easier to judge first] is definitely less than 25 -- shade the area between the 25 and 75 and you have most of the area shaded in; you should have only about 68 percent of the area shaded in for a "mean +/- one SD" area.
Likewise, shading the area between 45 and 55 would give you a narrow central strip, certainly less than 68 %.
This leaves 15 as the most likely SD here.
  d. SD for histogram (iii) will be bigger than for (ii) -- since the mean is off to one side of the highest point, there will be a lot of points to the left of the mean, and this will run up the sum of squared deviations. But note that the SD cannot be as high as 50, for any of the distributions, because this would cover ALL the data.

Problem 8. Average heights of boys and girls.

Given: Boys at age 9: 136 cm. at age 11: 146 cm.
Average heights of mixed random sample of boys and girls at age 11: 147 cm.

Part a. Are boys taller than girls at age 11? NO, since average height for the group is higher than the average height of boys alone.

Part b. Estimate the average height of boys at age 10. Answer: the average of the average heights at 9 and 11 is 141 cm. Assuming growth is linear, this will be our best guess.

Part c. (not in text). Suppose that the sample of 11 year olds had 600 girls and 400 boys. What would the average height of girls at 11 be?

We know that (600 * x + 400 * 146) / 1000 = 147 or
0.6 x + 0.4 * 146 = 147
Hence 0.6 x = 147 - 0.4 * 146 = 147 - 58.4 = 88.6
and x = 88.6 / 0.6 = 147.67
Problem 9. Mean, median and outliers.

Computer file with 1000 households has incomes in the range $5,800 to $98,600.
    By accident, the highest income gets an extra zero, and is recorded as $986,000.

Part a. Is the average affected? Yes. Suppose the average with the true value is $X_{\text{bar}}$
    Then $1000 \times X_{\text{bar}}$ is the sum of all incomes.
    With the mistaken value, the sum of all incomes is $1000 \times X_{\text{bar}} + (986,000 - 98,600)$
    or the computed average will go up by $(986,000 - 98,600) / 1000 = \$887.40$

Part b. Is the median affected? NO, not at all. The identity of the middle household does not change.

Problem 10. Law school scores.

Incoming students have mean LSAT = 163 and SD = 8. Pick a student at random and guess their score.
    For each point you are off, you will be penalized (absolute value of actual score - your guess)

Part a. What should you guess the score will be? Answer: 163, since the mean will minimize your likely loss.

Part b. You have about 1 chance in 3 of being more than 1 SD off. If you are more than 8 points off, you will lose more than $\$8$. 
Problem 4. SAT scores -- Math

In 1994, male SAT scores were distributed with mean of 500 and SD of 120.
women's SAT scores had mean 460 and SD of 120.

Find the percentage of (a) men and (b) women scoring above 660.


Standardized score = \( Z = \frac{600 - 500}{120} = 0.8333 \) Round to 0.85 to use the table.
Table lookup: Area = 60.47
Two tail area = 100 - 60.47 = 39.53
One tail area = 39.53 / 2 = 19.77. Nearly 20 percent of men scored above 600.

Part b. Women

Standardized score = \( Z = \frac{600 - 460}{120} = 1.1667 \) Round to 1.15 to use the table.
Table lookup: Area = 74.99
Two tail area = 100 - 74.99 = 25.01
One tail area = 25.01 / 2 = 12.51. Just over 12.5 percent of women scored above 600.

Note: with the computer, more accurate results can be found:
(anorm 1.166667) = 0.7567 or 75.67 percent, so two tail area is 24.33 percent and one tail area 12.17 percent.
(p-value 1.1667) = .1217 is a shorter way to get the upper tail area of a normal distribution.
(pnorm -1.16667) = .1217 gets the right tail area (the cumulative distribution function) of the normal distribution.
(normal-area 1.16667 6) shades the right tail area (only to 6, but the area to the right of 6 is close to 0).


Problem 7. Finding percentiles.
Assume the math SAT for applicants to a school had a mean of 500 and SD of 100 and followed a normal curve.

Part a. A score of 350 was at the **seventh** percentile of the distribution. Begin by standardizing the score of 350.  
\[ Z = \frac{350 - 500}{100} = -1.5 \]
Find the area to the LEFT of -1.5, that is, one of the tails lying outside the central area given by the text table.
Area between - 1.5 and + 1.5: 86.64 percent (from the normal table on A - 105).
Area outside the central area (two-tail area) 100 - 86.64 = 13.36 hence one tail area = 13.36 / 2 = 6.68 %

Part b. To be at the 75th percentile of the distribution a student would need a score of **570**.
This means that 25 percent lies in the right-hand tail defined by the area in the tables, so 25 percent is also in the left-hand tail (normal distribution is symmetric). The central area is therefore 50 percent. Look up 50 percent in the AREA column of the table: 51.61 is the closest, and corresponds to a Z-score of 0.7. This Z-score would be computed as  
\[ 0.7 = \frac{X - 500}{100} \]
which can be solved for the answer: **the needed score is 570**
By computer: Part a.  
\( \text{pnorm} -1.5 \) = 0.0668 or 6.68 percent.
Part b.  
\( \text{qnorm} 0.75 \) = 0.6745 is the needed Z-score to start the algebra.
(pnorm x) gives the cumulative density function (area UP TO x); (qnorm x) the quantile function (point on the axis to the left of which x percent of the data is located). Mind your Ps and Qs!
REVIEW EXERCISES (page 104)

Problem 3. Bias and chance error.
(a) Bias is NOT a form of CHANCE error, but is a SYSTEMATIC and PREDICTABLE error.
(b) Chance error is NOT a kind of bias, but will exist in ANY measurement.
(c) Measurements are always affected by chance error, but not always (and not usually if done
by trained observers) by bias.

Problem 4. Calibrating a yardstick.
The lab presents you with 3 results: 35.96 inches, 36.01 inches, and 36.03 inches.
These indicate the yardstick deviates from 36 inches by -0.04, 0.01 and 0.03. Note that the deviations sum to
zero, so the best guess so far is that the yardstick is pretty accurate, and the chance error is in the lab's
measurement process.
A fourth measurement would most likely have a similar error -- around 0.03 inches.
So, presented with a list offering 0.01, 0.03, and 0.06 inches as our choices, we would not
expect .01 to be "most likely" (only 1 of the three measurements has a deviation that small),
nor would we expect .06 (none of the three measurements so far has a deviation that large).
The SD of the measurements is 0.0294, which also answers our question.

Problem 3. Personality test and standardized scores.

<table>
<thead>
<tr>
<th>Original score:</th>
<th>79</th>
<th>64</th>
<th>52</th>
<th>72</th>
<th>(c) ______</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized</td>
<td>1.8</td>
<td>0.8</td>
<td>(a)</td>
<td>(b)</td>
<td>- 1.4</td>
</tr>
</tbody>
</table>

First, find the mean and standard deviation. From the first two values, and the formula for Z-scores, we know:

(i) \( \frac{79 - \text{mean}}{\text{SD}} = 1.8 \), so \( 1.8 \text{ SD} = 79 - \text{mean} \) OR \( \text{mean} = 79 - 1.8 \text{ SD} \)

(ii) \( \frac{64 - \text{mean}}{\text{SD}} = 0.8 \) so \( 0.8 \text{ SD} = 64 - \text{mean} \) OR \( \text{mean} = 64 - 0.8 \text{ SD} \)

Hence \( 79 - 1.8 \text{ SD} = 64 - 0.8 \text{ SD} \) or \( 79 - 64 = 1.8 \text{ SD} - 0.8 \text{ SD} \), so that \( \text{SD} = 15 \)
and substituting back into equation (ii): \( 64 - 0.8 (15) = \text{mean} \) or \( \text{MEAN} = 52 \)

(a) \( Z\)-score = \( \frac{52 - 52}{15} = 0 \)

(b) \( Z\)-score = \( \frac{72 - 52}{15} = 20 / 15 = 1.3333 \)

(c) \( Z\)-score = \( -1.4 = \frac{X - 52}{15} \), so \( -21 = X - 52 \) or \( X = 31 \)
**Problem 4. Verbal SAT scores**

Have mean of 500 and SD of 100; assumed to follow the normal curve.

(a) Percentage of students with scores between 350 and 650?

\[
Z\text{-score of 650} = \frac{(650 - 500)}{100} = 1.5 \\
Z\text{-score of 350} = \frac{(350 - 500)}{100} = -1.5
\]

Table gives directly as 86.64 percent.

(b) Of about 1000 students with scores in the range 400-600, how many had scores from 450-550?

Answer: MORE THAN HALF, since the 450-550 range has the highest density of students.

So of the choices given, (440, 500, 560) you should choose 560.

Note that with a mean of 500 and a SD of 100, the 400-600 range is the mean +/- 1 SD.

Hence the area should be about 68 percent (68.27 percent from table A-1)

The 450-550 range is the mean +/- one-half a SD; Table A-1 reports this is 38.29 percent.

We would expect \( \frac{38.29}{68.27} = 0.56 \) percent of the students in the range 400-600 to be in the 450-550 range. The text choice of 560 was not made by accident.

**Computer notes:**

Simulation: (bind x (rnorm 5000 500 100)); (hist x); (shadebins 400 600 yellow); (shadebins 450 550)

Calculation: (anorm 1); (anorm 0.5) gives the areas under the normal curve.

**Problem 6 and 7. Men, women and math.**

At College X, math SAT scores averaged 650 for men and 600 for women.

(6) if there are 500 men and 500 women, the average math SAT for the student body is

\[
0.5 \times 650 + 0.5 \times 600 = 625
\]

The SD for the entire student body will be higher than for either group individually:

if you are at the male mean, you are 650 - 625 = 25 points away from the overall mean,

so your score will contribute to the SD of the group, and similarly for women.

(7) if there are 600 men and 400 women, the average SAT is

\[
0.6 \times 650 + 0.4 \times 600 = 630
\]

The average score moves closer to the male score.

The SD of the overall group will be greater -- will it be greater than is the case in (6)?

Modification: this will be clearer if we have a larger college with a greater difference of scores;

a smaller SD will help as well. We assume in College Y with 5000 men, 5000 women, male average SAT = 400; female average SAT = 700; SD = 100 for both.

Simulation: (bind sat0 (rnorm 5000 400 100)), (bind sat1 (rnorm 5000 700 100), (bind satall (combine sat0 sat1)).

(density-plot sat0 sat1) and (hist satall) will quite clearly show the bimodality.

My simulation gave (sd sat0) = 100.4788; (sd sat1) = 99.2223 and (sd satall) = 180.8835.

**Problem 9. Quiz scores.** A quiz with 10 questions has an average of 6.4 answers correct and a SD of 2.0.

It will therefore have an average of \( 10 - 6.4 = 3.6 \) questions incorrect; we are changing the mean by

(a) subtracting 10 from 6.4; this changes the mean but not the SD.

(b) multiplying 6.4 - 10 = -3.6 by -1; this will not change the SD either. The SD remains at 2.0.