## Chapter 1: Looking at Data – Distributions

### Introduction

*Statistics is the science of learning from data.* The first step in dealing with data is to organize your thinking about the data.

An *exploratory data analysis* is the process of using statistical tools and ideas to examine data in order to describe their main features.

### Section 1.1: Data

#### CASES, LABELS, VARIABLES, AND VALUES

Cases are the objects described by a set of data. Cases may be customers, companies, subjects in a study, or other objects.

A label is a special variable used in some data sets to distinguish the different cases.

A variable is a characteristic of a case. Different cases can have different values for the variables.

#### CATEGORICAL AND QUANTITATIVE VARIABLES

A categorical variable places a case into one of several groups or categories.

A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

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

Caution! ---- Beware of coded variable

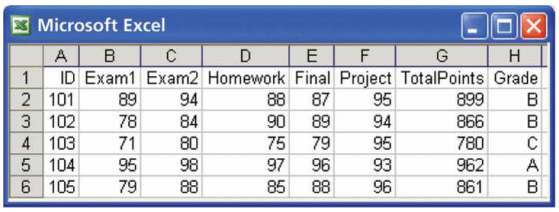
In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

Who? What cases do the data describe? How many cases appear in the data?

What? How many variables do the data contain? What are the exact definitions of these variables? In what unit of measurement is each variable recorded?

Why? What purpose do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about cases other than the ones we actually have data for? Are the variables that are recorded suitable for the intended purpose?

1.1: The following figure shows part of a data set for students enrolled in an introductory statistics class. Each row gives the data on one student. The values for the different variables are in the columns. This data set has eight variables. ID is an identifier for each student. Exam1, Exam2, Homework, Final, and Project give the points earned, out of a total of 100 possible, for each of these course requirements. Final grades are based on a possible 200 points for each exam and the final, 300 points for Homework, and 100 points for Project. TotalPoints is the variable that gives the composite score. It is computed by adding 2 times Exam1, Exam2, and Final, 3 times Homework plus 1 times Project. Grade is the grade earned in the course. This instructor used cut-offs of 900, 800, 700, etc. for the letter grades.




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | ID | Exam1 | Exam2 | Homework | Final | Project | TotalPoints | Grade |
| 2 | 101 | 89 | 94 | 88 | 87 | 95 | 899 | B |
| 3 | 102 | 78 | 84 | 90 | 89 | 94 | 866 | B |
| 4 | 103 | 71 | 80 | 75 | 79 | 95 | 780 | C |
| 5 | 104 | 95 | 98 | 97 | 96 | 93 | 962 | A |
| 6 | 105 | 79 | 88 | 85 | 88 | 96 | 861 | B |

a) Who, what, and why for the statistics class data. Answer the who, what, and why questions for the statistics class data set.
b) Give values of the variables Exam 1, Exam 2, and Final for the student with ID equal to 104.
c) A student whose data do not appear on the spreadsheet scored 83 on Exam 1, 82 on Exam 2, 77 for Homework, 90 on the final, and 80 on the project. Find TotalPoints for this student and give the grade earned.

1.2: Employee application data. The personnel department keeps records on all employees in a company. Here is the information that they keep in one of their data files: employee identification number, last name, first name, middle initial, department, number of years with the company, salary, education (coded as high school, some college, or college degree), and age.
(a) What are the cases for this data set?
(b) Describe each type of information as a label, a quantitative variable, or a categorical variable.

Homework/Page 8:1.11

1.2: Displaying distributions with graphs

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called exploratory data analysis.

Categorical variables: Pie charts, Bar graphs

The distribution of a categorical variable lists the categories and gives either the count or the percent of cases who fall in each category.

A bar graph is a picture consisting of horizontal and vertical axes with rectangle (or rectangular objects) that represent the amount (or percent) of the categories of a variable. The categories of the variables are listed along one axis and the frequencies along the other.

Pie chart

A circle or pie is divided into pieces corresponding to the categories of the variable so that the size of the slice is proportional to the percent of the cases who fall in each category.
Pie charts require that you include all the categories that make up a whole. Use them only when you want to emphasize each category's relation to the whole.

1.3: Mobile browsing and iPhones. Users of iPhones were asked to respond to the statement, "I do a lot more browsing on the iPhone than I did on my previous mobile phone" and responded as follows.

| Response | Percent (%) |
|---|---|
| Strongly agree | 54 |
| Mildly agree | 22 |
| Mildly disagree | 16 |
| Strongly disagree | 8 |

(a) Make a bar graph to display the distribution of the responses.
(b) Display the distribution with a pie chart.

Quantitative variables: stemplots

A stemplot (Stem-and-Leaf Plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

To make a stemplot

1. Separate each observation into a *stem* consisting of all but the final (right most) digit and a *leaf*, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line to the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

1.4: Make a stemplot. Here are the scores on the first exam in an introductory statistics course for 30 students in one section of the course:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 73 | 92 | 85 | 75 | 98 | 93 | 55 | 80 | 90 | 92 | 80 | 87 | 90 | 72 |
| 65 | 70 | 85 | 83 | 60 | 70 | 90 | 75 | 75 | 58 | 68 | 85 | 78 | 80 | 93 |

Use these data to make a stemplot.

Stemplots do not work well for large data sets, where each stem must hold a large number of leaves. Fortunately, there are two modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations.

You can double the number of stems in a plot by splitting each stem into two: One with leaves 0 to 4 and the other with leaves 5 through 9.

When the observed values have many digits, it is often best to trim the numbers by removing the last digit or digits before making a stemplot.

1.5: Diabetes and glucose. People with diabetes must monitor and control their blood glucose level. The goal is to maintain "fasting plasma glucose" between about 90 and 130 milligrams per deciliter (mg/dl). Here are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 141 | 158 | 112 | 153 | 134 | 95 | 96 | 78 | 148 |
| 172 | 200 | 271 | 103 | 172 | 359 | 145 | 147 | 255 |

Make a stemplot of these data (You will want to trim and also split stems.) Are there outliers?

You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution.
If a stemplot has fewer than about five stems, you should usually split the stems unless there are few observations.
If there are many stems with no leaves or only one leaf, trimming will reduce the number of stems.
When you wish to compare two related distributions, a back – to – back stemplot with common stems is useful

1.6: Baby Ruth's home run counts from 1920 to 1934: 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22
Mark McGwire's home run counts from 1987 to 1998: 49 32 33 39 22 42 9 9 39 52 58 70
Make a back-to-back stemplot. Write the stems as usual, but with a vertical line both to their left and to their right. On the right, put leaves for Ruth. On the left, put leaves for McGwire. Arrange the leaves on each stem in increasing order out from the common stem.

1.7: Compare glucose of instruction and control groups**.** The study described in the previous exercise also measured the fasting plasma glucose of 16 diabetics who were given individual instruction on diabetes control.
Here are the data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 128 | 195 | 188 | 158 | 227 | 198 | 163 | 164 |
| 159 | 128 | 283 | 226 | 223 | 221 | 220 | 160 |
| 141 | 158 | 112 | 153 | 134 | 95 | 96 | 78 | 148 |
| 172 | 200 | 271 | 103 | 172 | 359 | 145 | 147 | 255 |

Make a back-to-back stemplot to compare the class and individual instruction groups.

Histograms
A histogram breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes of equal width.

Histograms
For quantitative variables that take many values
1. Divide the range of the data into *classes* of equal width.
2. Count the number of individuals in each class. These counts are called frequencies, and a table of frequencies for all classes is a frequency table.
3. Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero.

1.8: Acidity of rainwater. Changing the choice of classes can change the appearance of a histogram. Here is an example in which a small shift in the classes, with no change in the number of classes, has an important effect on the histogram. The data are the acidity levels (measured by pH) in 105 samples of rainwater. Distilled water has pH 7.00. As the water becomes more acidic, the pH goes down. The pH of rainwater is important to environmentalists because of the problem of acid rain.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4.33 | 4.38 | 4.48 | 4.48 | 4.50 | 4.55 | 4.59 | 4.59 | 4.61 | 4.61 |
| 4.75 | 4.76 | 4.78 | 4.82 | 4.82 | 4.83 | 4.86 | 4.93 | 4.94 | 4.94 |
| 4.94 | 4.96 | 4.97 | 5.00 | 5.01 | 5.02 | 5.05 | 5.06 | 5.08 | 5.09 |
| 5.10 | 5.12 | 5.13 | 5.15 | 5.15 | 5.15 | 5.16 | 5.16 | 5.16 | 5.18 |
| 5.19 | 5.23 | 5.24 | 5.29 | 5.32 | 5.33 | 5.35 | 5.37 | 5.37 | 5.39 |
| 5.41 | 5.43 | 5.44 | 5.46 | 5.46 | 5.47 | 5.50 | 5.51 | 5.53 | 5.55 |
| 5.55 | 5.56 | 5.61 | 5.62 | 5.64 | 5.65 | 5.65 | 5.66 | 5.67 | 5.67 |
| 5.68 | 5.69 | 5.70 | 5.75 | 5.75 | 5.75 | 5.76 | 5.76 | 5.79 | 5.80 |
| 5.81 | 5.81 | 5.81 | 5.81 | 5.85 | 5.85 | 5.90 | 5.90 | 6.00 | 6.03 |
| 6.03 | 6.04 | 6.04 | 6.05 | 6.06 | 6.07 | 6.09 | 6.13 | 6.21 | 6.34 |
| 6.43 | 6.61 | 6.62 | 6.65 | 6.81 | | | | | |

(a) Make a histogram of pH with 14 classes, using class boundaries 4.2, 4.4,…, 7.0. How many modes does your histogram show? More than one mode suggests that the data contain groups that have different distributions.
(b) Make a second histogram, also with 14 classes, using class boundaries 4.14, 4.34,…, 6.94. The classes are those from (a) moved 0.06 to the left. How many modes does the new histogram show?
(c) Use your software's histogram function to make a histogram without specifying the number of classes or their boundaries. How does the software's default histogram compare with those in (a) and (b)?

How many intervals? ---- One rule is to calculate the square root of the sample size, and round up.
Size of intervals? ---- Divide the range of the data (max.–min.) by number of intervals desired, and round to convenient number
Pick intervals so each observation can only fall in exactly one interval (no overlap)

Examining Distributions
In any graph of data, look for the overall pattern and for striking deviations from the pattern.
You can describe the overall pattern of a distribution by its shape, center, and spread.
An important kind of deviation is an outlier, an individual value that falls outside the overall pattern.

Tails: The extreme values of a distribution are in the tails of the distribution. The high values are in the upper, or right, tail and the low values are in the lower, or left, tail.

Shape of the Data
Does the distribution have one or several major peaks, called modes?
A distribution with one major peak is called unimodal.
Symmetric or skewed
Is it approximately symmetric or is it skewed in one direction?

A distribution is symmetric if values smaller and larger than its midpoint are mirror images of each other. It is skewed to the right if the right tail (larger values) is much longer than the left tail (smaller values).

1.9: Describe the IQ scores. Make a graph of the distribution of IQ scores for the seventh-grade students. Describe the shape, center, and spread of the distribution, as well as any outliers. IQ scores are usually said to be centered at 100. Is the midpoint for these students close to 100, clearly above, or clearly below?

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 111 | 107 | 100 | 107 | 114 | 115 | 111 | 97 | 100 | 112 | 104 | 89 | 104 |
| 102 | 91 | 114 | 114 | 103 | 105 | 106 | 113 | 109 | 108 | 113 | 130 | 128 |
| 128 | 118 | 113 | 120 | 132 | 111 | 124 | 127 | 128 | 136 | 106 | 118 | 119 |
| 123 | 124 | 126 | 116 | 127 | 119 | 97 | 86 | 102 | 110 | 120 | 103 | 115 |
| 93 | 72 | 111 | 103 | 123 | 79 | 119 | 110 | 110 | 107 | 74 | 105 | 112 |
| 105 | 110 | 107 | 103 | 77 | 98 | 90 | 96 | 112 | 112 | 114 | 93 | 106 |

<u>Time Plots</u>

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable being measured is on the vertical scale.

Look for an overall pattern (*trend*) and deviations from this trend. Connecting the data points by lines may emphasize this trend.

Look for patterns that repeat at known regular intervals (seasonal variations).

1.10: The Boston Marathon. Women were allowed to enter the Boston Marathon in 1972. Here are the times (in minutes, rounded to the nearest minute) for the winning women from 1972 to 2012:

| Year | Time | Year | Time | Year | Time | Year | Time |
|---|---|---|---|---|---|---|---|
| 1972 | 190 | 1983 | 143 | 1994 | 142 | 2005 | 145 |
| 1973 | 186 | 1984 | 149 | 1995 | 145 | 2006 | 143 |
| 1974 | 167 | 1985 | 154 | 1996 | 147 | 2007 | 149 |
| 1975 | 162 | 1986 | 145 | 1997 | 146 | 2008 | 145 |
| 1976 | 167 | 1987 | 146 | 1998 | 143 | 2009 | 152 |
| 1977 | 168 | 1988 | 145 | 1999 | 143 | 2010 | 146 |
| 1978 | 165 | 1989 | 144 | 2000 | 146 | 2011 | 142 |
| 1979 | 155 | 1990 | 145 | 2001 | 144 | 2012 | 151 |
| 1980 | 154 | 1991 | 144 | 2002 | 141 | | |
| 1981 | 147 | 1992 | 144 | 2003 | 145 | | |
| 1982 | 150 | 1993 | 145 | 2004 | 144 | | |

Make a graph that shows change over time. What overall pattern do you see? Have times stopped improving in recent years? If so, when did improvement end?

Homework:
Pages 25 – 29: 1.25, 1.27, 1.29, 1.31, 1.33, 1.35, 1.37, 1.39, 1.43, 1.45.

## 1.3: Describing Distributions with Numbers

A brief description of a distribution should include its shape and numbers describing its center and spread. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now we will learn specific ways to use numbers to measure the center and spread of a distribution.

### Measuring Center

The two common measures of center are the mean and the median.
The mean is the "average" and the median is the "middle value."

### The Mean $\bar{x}$

To find the mean $\bar{x}$ of a set of observations, add their values and divide by the number of observations.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean is sensitive to the influence of a few extreme observations. These may be outliers. But a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a resistant measure of center.

### The Median *M*

The median M is the midpoint of a distribution. Half the observations are smaller than the median and the other half are larger than the median.

Arrange all observations in order of size, from smallest to largest.
If the number of observations *n* is odd, the median M is the center observation in ordered list. Find the location of the median by counting *(n+1)/2* observations up from the bottom of the list.
If the number of observations *n* is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again *(n+1)/2* from the bottom of the list.
Note that *(n+1)/2* does not give the median, just the location of the median in the ordered list.

Median is more resistant than mean.

1.11: Calls to a customer service center. The service times for 80 calls to a customer service center are given in Table. Use these data to compute the median service time.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 77 | 289 | 128 | 59 | 19 | 148 | 157 | 203 |
| 126 | 118 | 104 | 141 | 290 | 48 | 3 | 2 |
| 372 | 140 | 438 | 56 | 44 | 274 | 479 | 211 |
| 179 | 1 | 68 | 386 | 2631 | 90 | 30 | 57 |
| 89 | 116 | 225 | 700 | 40 | 73 | 75 | 51 |
| 148 | 9 | 115 | 19 | 76 | 138 | 178 | 76 |
| 67 | 102 | 35 | 80 | 143 | 951 | 106 | 55 |
| 4 | 54 | 137 | 367 | 277 | 201 | 52 | 9 |
| 700 | 182 | 73 | 199 | 325 | 75 | 103 | 64 |
| 121 | 11 | 9 | 88 | 1148 | 2 | 465 | 25 |

Mean versus Median

The mean and median are the most common measures of the center of a distribution.
The mean and median of data from a symmetric distribution are close together.
If the distribution is exactly symmetric, the mean and median are exactly the same.
In a skewed distribution, the mean is farther out in the long tail than is the median [the mean is 'pulled' in the direction of the possible outlier(s)].

Measuring spread: the quartiles

The simplest useful numerical description of a distribution consist of both a measure of center and a measure of spread.

The Quartiles $Q_1$ and $Q_3$

To calculate the quartiles:
Arrange the observations in increasing order and locate the median M in the ordered list of observations.
The first quartile $Q_1$ is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
The third quartile $Q_3$ is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The Five-Number Summary

The five-number summary of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is
Minimum $Q_1$ M $Q_3$ Maximum

Boxplot

A boxplot is a graph of the five-number summary.
A central box spans $Q_1$ and $Q_3$.
A line in the box marks the median *M*.
Lines extend from the box out to the smallest and largest observations.

1.12: Find the five-number summary. Here are the scores on the first exam in an introductory statistics course for 10 students:
80 73 92 85 75 98 93 55 80 90

Find the five-number summary for these first-exam scores.

The Interquartile Range IQR

The Interquartile Range IQR is the distance between the first and third quartiles,
$IQR = Q_3 - Q_1$

The quartiles and the *IQR* are not affected by changes in either tail of the distribution. They are resistant.

The $1.5 \times IQR$ Rule For Outliers:

Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third or below the first quartile.

1.13: Find the IQR**.** Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the interquartile range and use the 1.5 × *IQR* rule to check for outliers. How low would the lowest score need to be for it to be an outlier according to this rule?

Measuring spread: the standard deviation

The variance $s^2$ of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols the variance of $n$ observations $x_1, x_2, ..., x_n$ is

$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$ or, more compactly $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$

The standard deviation $s$ is the square root of the variance $s^2$.

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

1.14: Find the variance and the standard deviation**.** Here are the scores on the first exam in an introductory statistics course for 10 students:

73 92 85 75 98 93 55 80 90

Find the variance and the standard deviation for these first-exam scores.

Properties of the Standard Deviation

*s* measures spread about the mean and should be used only when the mean is chosen as the measure of center.

*s = 0* only when there is no spread. This happens only when all observations have the same value. Otherwise s > 0. As the observations become more spread out about their mean, *s* gets larger.

*s,* like the mean, is not resistant. A few outliers can make *s* very large.

Choosing a Summary

Outliers affect the values of the mean and standard deviation.
The five-number summary is usually better than the mean and standard deviation for describing a skewed distributions or a distribution with strong outliers.
Use $\bar{x}$ and *s* only for reasonably symmetric distributions that are free of outliers

Linear Transformations

A linear transformation changes the original variable *x* into the new variable $x_{new}$ given by an equation of the form $x_{new} = a + bx$

Adding the constant *a* shifts all the values of x upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by a positive number *b* changes the size of the unit of measurement.

Linear transformations do not change the shape of a distribution.

Effect of a linear transformation

To see the effect of a linear transformation on measures of center and spread, apply these rules:
Multiplying each observation by a positive number *b* multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by *b*.
Adding the same number *a* (either positive or negative) to each observation adds *a* to measures of center and to quartiles and other percentiles but not change measures of spread.

1.15: A different type of mean**.** The trimmed mean is a measure of center that is more resistant than the mean but uses more of the available information than the median. To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Trimming eliminates the effect of a small number of outliers. Compute the 10% trimmed mean of the service time data in the following table. Then compute the 20% trimmed mean. Compare the values of these measures with the median and the ordinary untrimmed mean

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 77 | 289 | 128 | 59 | 19 | 148 | 157 | 203 |
| 126 | 118 | 104 | 141 | 290 | 48 | 3 | 2 |
| 372 | 140 | 438 | 56 | 44 | 274 | 479 | 211 |
| 179 | 1 | 68 | 386 | 2631 | 90 | 30 | 57 |
| 89 | 116 | 225 | 700 | 40 | 73 | 75 | 51 |
| 148 | 9 | 115 | 19 | 76 | 138 | 178 | 76 |
| 67 | 102 | 35 | 80 | 143 | 951 | 106 | 55 |
| 4 | 54 | 137 | 367 | 277 | 201 | 52 | 9 |
| 700 | 182 | 73 | 199 | 325 | 75 | 103 | 64 |
| 121 | 11 | 9 | 88 | 1148 | 2 | 465 | 25 |

Homework

Pages 48 – 53: 1.61, 1.63, 1.65, 1.71, 1.73, 1.88, 1.89

1.4: Density curves and Normal Distributions

We have a clear strategy for exploring data on a single quantitative variable:
Always plot your data: make a graph, usually a stemplot or histogram.
Look for the overall pattern and for striking deviations such as outliers.
Calculate an appropriate numerical summary to briefly describe center and spread.

Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

The smooth curve drawn over the histogram is called a density curve.

Density Curves

A density curve is a curve that is always on or above the horizontal axis and has area exactly 1 underneath it.
A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range

1.16: A uniform distribution**.** If you ask a computer to generate "random numbers" between 0 and 1, you will get observations from a **uniform distribution**. The figure below graphs the density curve for a uniform distribution. Use areas under this density curve to answer the following questions.
(a) Why is the total area under this curve equal to 1?
(b) What proportion of the observations lie below 0.35?
(c) What proportion of the observations lie between 0.35 and 0.65?

Median and Mean of a Density Curve

The *median* of a density curve is the equal-areas point, the point that divides the area under the curve in half
The *mean* of a density curve is the balance point, at which the curve would balance if made of solid material
The median and mean are the same for a symmetric density curve. They both lie at the center of the curve.
The mean of a skewed curve is pulled away from median in the direction of the long tail.
The mean and standard deviation computed from actual observations (data) are denoted by $\bar{x}$ and *s*, respectively.
The mean and standard deviation of the actual distribution represented by the density curve are denoted by $\mu$ ("mu") and $\sigma$("sigma"), respectively.

Normal distributions

The density curves that are symmetric, unimodal, and bell shaped are called normal curves, and they describe normal distribution.
All normal distributions have the same overall shape.
The exact density curve for a particular normal distribution is described by giving its mean and its standard deviation.

The mean is located at the center of the symmetric curve and is the same as the median.
Changing $\mu$ without changing $\sigma$ moves the normal curve along the horizontal axis without changing its spread.
The standard deviation $\sigma$ controls the spread of a normal curve.
The curve with larger standard deviation is more spread out.

Locating std dev $\sigma$ : As we move out in either direction from the center $\mu$, the curve changes from falling ever more steeply to falling ever less steeply. The points at which this change of curvature takes place are located at distance $\sigma$ on either side of the mean $\mu$.
We abbreviate the Normal distribution with mean $\mu$ and standard deviation $\sigma$ as N($\mu$,$\sigma$).

<u>The 68-95-99.7 Rule</u>
In the Normal distribution with mean $\mu$ and standard deviation $\sigma$ :
Approximately 68% of the observations fall within $\sigma$ of the mean $\mu$.
Approximately 95% of the observations fall within $2\sigma$ of $\mu$.
Approximately 99.7% of the observations fall within $3\sigma$ of $\mu$.

As the 68-95-99.7 rule suggests, all normal distributions share many common properties. In fact, all normal distributions are the same if we measure in units of size $\sigma$ about the mean $\mu$ as center. Changing to these units called standardizing

1.17: Test scores. Many states have programs for assessing the skills of students in various grades. The Indiana Statewide Testing for Educational Progress (ISTEP) is one such program. In a recent year 76,531 tenth-grade Indiana students took the English/language arts exam. The mean score was 572 and the standard deviation was 51. Assuming that these scores are approximately Normally distributed, *N*(572,51), use the 68–95–99.7 rule to give a range of scores that includes 95% of these students.

1.18: Refer to the previous exercise. Use the 68–95–99.7 rule to give a range of scores that includes 99.7% of these students.

<u>Standardizing and Z-score</u>
If *x* is an observation from a distribution that has mean $\mu$ and standard deviation $\sigma$, the standardized value of *x* is $z = \frac{x - \mu}{\sigma}$. A standardized value *x* is often called a *z-score*.

1.19: Find the *z*-score**.** Consider the ISTEP scores, which we can assume are approximately Normal, *N*(572, 51). Give the *z*-score for a student who received a score of 620.

1.20: Find the *z*-score**.** Consider the ISTEP scores which we can assume are approximately Normal, *N*(572, 51). Give the *z*-score for a student who received a score of 510. Explain why your answer is negative even though all of the test scores are positive.

Standard Normal Distribution

The standard Normal distribution is the Normal distribution N(0,1) with mean 0 and standard deviation 1.

If a variable X has any Normal distribution $N(\mu,\sigma)$ with mean $\mu$ and standard deviation $\sigma$, then the standardized variable $z = \frac{X-\mu}{\sigma}$ has the standard Normal distribution.

Cumulative proportion

When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value.

1.21: Find some proportions. Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

(a) $Z < 1.65$ (b) $Z > 1.65$ (c) $Z > -0.76$ (d) $-0.76 < Z < 1.65$

1.22: Find the proportion**.** Consider the ISTEP scores, which are approximately Normal, *N*(572, 51). Find the proportion of students who have scores less than 620. Find the proportion of students who have scores greater than or equal to 620. Sketch the relationship between these two calculations using pictures of Normal curves.

1.23: Find the proportion**.** Consider the ISTEP scores, which are approximately Normal, *N*(572, 51). Find the proportion of students who have scores between 620 and 660. Use pictures of Normal curves to illustrate your calculations.

Inverse normal calculations

Observed Value for a Standardized Score

Need to "unstandardize" the z-score to find the observed value (*x*) :

$$z = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad x = \mu + z\,\sigma$$

1.24: Find some values of *z*. Find the value *z* of a standard Normal variable *Z* that satisfies each of the following conditions. (If you use Table A, report the value of *z* that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of *z* marked on the axis.

(a) 22% of the observations fall below *z*.

(b) 40% of the observations fall above *z*.

1.25: SAT scores are reported on a scale from 600 to 2400. The SAT scores are approximately Normal with mean μ = 1509 and standard deviation σ = 321.

How low is the bottom 20%? What SAT scores make up the bottom 20% of all scores?

How high is the top 10%? What SAT scores make up the top 10% of all scores?

1.26: Find the score that 80% of students will exceed. Consider the ISTEP scores, which are approximately Normal, *N*(572, 51). Eighty percent of the students will score above *x* on this exam. Find *x*.

Normal Quantile plots

Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies.
Do normal distribution calculations to find the *z-score* at these percentiles.
Plot each data point *x* against the corresponding *z.*
If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data follows a normal distribution.
Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal distribution. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

Use of Normal Quantile Plots:

If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

1.27: Density of the earth. We expect repeated careful measurements of the same quantity to be approximately Normal. Make a Normal quantile plot for Cavendish's measurements given below. Are the data approximately Normal? If not, describe any clear deviations from Normality.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5.50 | 5.61 | 4.88 | 5.07 | 5.26 | 5.55 | 5.36 | 5.29 | 5.58 | 5.65 |
| 5.57 | 5.53 | 5.62 | 5.29 | 5.44 | 5.34 | 5.79 | 5.10 | 5.27 | 5.39 |
| 5.42 | 5.47 | 5.63 | 5.34 | 5.46 | 5.30 | 5.75 | 5.68 | 5.85 | |

Homework

Pages 73-- 77: 1.111, 1.113, 1,123, 1.115, 1.119, 1.121, 1.123, 1.125, 1.129, 1.131, 1.133, 1.135, 1.137, 1.139, 1.141, 1.143, 1.145.