1.12 (a) In addition to the two identification columns, ID *Country* and the abbreviated *Code*, there are 12 variables. We see that *Developed* is a categorical variable, while the other eleven variables are all quantitative.

(b) There are many possible answers, such as "What is the average life expectancy for all countries of the world?" or "What proportion of countries are developed?"

(c) There are many possible answers, such as "Do countries with a greater land area have a larger percent rural?" or "Do countries that spend a relatively large amount on the military spend a relatively small amount on health care?" or "Do developed countries have a longer life expectancy than developing countries?"

1.20 (a) There are 8 cases, corresponding to the 8 rowers. The two variables are number of days to cross the Atlantic and gender. Number of days to cross the Atlantic is quantitative and gender is categorical.

(b) We need two columns, one for each variable. The columns can be in either order. See the table.

Time	Gender
40	Male
87	Male
78	Male
106	Male
67	Male
70	Female
153	Female
81	Female

1.22 (a) The cases are the 41 participants.

(b) There are many variables in this study. The only categorical variable is whether or not the person participated in the meditation program. All other variables are quantitative variables. These variables include (at minimum):

- Brain wave activity before
- Brain wave activity after
- Brain wave activity 4 months later
- Immune response after 1 month
- Immune response after 2 months
- Negative survey before
- Negative survey after
- Positive survey before
- Positive survey after

(c) The explanatory variable is whether or not the person participated in the meditation program.

(d) The data set will have 41 rows (one for each participant) and at least 10 columns (one for each variable).

1.50 From the description, it appears that this method of data collection is not biased.

1.56 No. This is a volunteer sample, and there is reason to believe the participants are not representative of the population. For example, some may choose to participate because they LIKE alcohol and/or marijuana, and those in the sample may tend to have more experience with these substances than the overall population. In addition, the advertisements for the study were aired on rock radio stations in Sydney, so only those people who listen to rock radio stations in Sydney would hear about the option to participate.

1.60 (a) Since the NHANES sample is drawn from all people in the US, that is the population we can generalize to.

(b) Since the NHAMCS sample is drawn from patients in emergency rooms in the US, we can generalize the results to all emergency room patients in the US.

(c) i. NHANES: The question about an association between being overweight and developing diabetes applies to all people in the US, not just those who visit an emergency room.

ii. NHAMCS: This question asks specifically about the type of injury for people who go to an emergency room.

iii. NHAMCS: This question of average waiting time only applies to emergency room patients.

iv. NHANES: This question is asking about all US residents. Note that the proportion would be equal to one for the people sampled in NAMCS since they only get into the sample if they visit an emergency room!

2.18 (a) The table is given.

	HS or less	Some college	College grad	Total
Agree	363	176	196	735
Disagree	557	466	789	1812
Don't know	20	26	32	78
Total	940	668	1017	2625

(b) For the survey participants with a high school degree or less, we see that $363/940 = 0.386$ or 38.6% agree. For those with some college, the proportion is $176/668 = 0.263$, or 26.3% agree, and for those with a college degree, the proportion is $196/1017 = 0.193$, or 19.3% agree. There appears to be an association, and it seems that as education level goes up, the proportion who agree that every person has one true love goes down.

(c) We see that $1017/2625 = 0.387$, or 38.7% of the survey responders have a college degree or higher.

(d) A total of 1812 people disagreed and 557 of those have a high school degree or less, so we have $557/1812 = 0.307$, or 30.7% of the people who disagree have a high school degree or less.

2.20 (a) This is an observational study since the researchers are observing the results after the fact and are not manipulating the gene directly to force a disruption. There are two variables: whether or not the person has dyslexia and whether or not the person has the DYXC1 break.

(b) Since $109 + 195 = 304$ people participated in the study, there will be 304 rows. Since there are two variables, there will be 2 columns: one for dyslexia or not and one for gene break or not.

(c) A two-way table showing the two groups and gene status is shown.

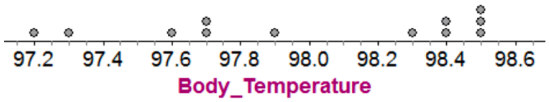
	Gene break	No break	Total
Dyslexia group	10	99	109
Control group	5	190	195
Total	15	289	304

(d) We look at each row (Dyslexia and Control) individually. For the dyslexia group, the proportion with the gene break is $10/109 = 0.092$. For the control group, the proportion with the gene break is $5/195 = 0.026$.

(e) There is a very substantial difference between the two proportions in part (d), so there appears to be an association between this particular genetic marker and dyslexia for the people in this sample. (As mentioned, we see in Chapter 4 how to determine whether we can generalize this result to the entire population.)

(f) We cannot assume a cause-and-effect relationship because this data comes from an observational study, not an experiment. There may be many confounding variables.

2.52 (a) A dotplot of the body temperatures is shown below.



(b) We compute $\overline{x} = 98.0°\text{F}$. It is the balance point in the dotplot.

(c) There are $n = 12$ data values, so the median is the average of the two middle values. We have

$$m = \frac{97.9 + 98.3}{2} = 98.1°\text{F}.$$

This is a point in the dotplot that has six dots on either side.

2.56 (a) The distribution is skewed to the left since there are many values between about 72 and 82 and then a long tail going down to the outliers on the left.

(b) Since half the values are above 72, the median is about 72. (The actual median is 71.9.)

(c) Since the data is skewed to the left, the mean will be less than the median so the mean will be less than 72. (It is actually about 68.9.)

2.60 (a) We have $\overline{x}_f = 6.40$.

(b) We have $\overline{x}_m = 6.81$.

(c) We see that $\overline{x}_m - \overline{x}_f = 6.81 - 6.40 = 0.41$. In this sample, the males, on average, spent 0.41 more hours per week exercising than the females.