1. (a) One way to summarize it is with a two-way table showing the frequency of the various pairings of values of the two variables:

	treatment	
progress	aspirin	placebo
favorable	62	45
unfavorable	17	30

Another way would be to give relative frequencies, which then presents you with a choice: "out of what whole?" One could divide all the numbers in the table above by 154, since that is the total number of participants in the study. But it may be more relevant to break apart the treatment groups, considering the 79 aspirin-takers as one "whole", and the 75 placebo-takers as another. That's how we approach things for the coming hypothesis test: summarize the data for the two groups by saying:

$62/(62+17) = 0.785 = 78.5\%$ of aspirin-takers had a favorable outcome (correspondingly, $17/(62+17) = 0.215 = 21.5\%$ had an unfavorable outcome), and
$45/(45+30) = 0.6 = 60\%$ of placebo-takers had a favorable outcome (correspondingly, $30/(45+30) = 0.4 = 40\%$ had an unfavorable outcome).

(b) One possible answer: "Does aspirin offer a higher rate (a higher proportion) of favorable outcomes to people at risk of stroke than not taking aspirin does?"

(c) Considering our populations to be

Population A: all adults at risk of stroke and taking aspirin
Population B: all adults at risk of stroke and not taking aspirin

then we will denote the proportion from population A which has a favorable outcome by p_A, and the proportion from population B which has a favorable outcome by p_B. Using these symbols for the population parameters, our null and alternative hypotheses would likely be

$$\mathbf{H}_0\colon\ p_A - p_B = 0 \qquad \text{vs.} \qquad \mathbf{H}_a\colon\ p_A - p_B > 0.$$

Note that I did a "greater-than" comparison (i.e., a one-sided alternative hypothesis) as a result of my research question, which asked if aspirin helps bring on favorable outcomes. If my question were simply whether aspirin changes the rate of favorable outcomes, then I would have written a two-sided alternative hypothesis ($p_A - p_B \neq 0$).

(d) The best estimate, from our sample, of the difference $p_A - p_B$ is the difference in sample proportions. Above, we found

$$\widehat{p}_A = 0.785 \qquad \text{and} \qquad \widehat{p}_B = 0.6,$$

so the test statistic is the difference of sample proportions

$$\widehat{p}_A - \widehat{p}_B = 0.785 - 0.6 = 0.185.$$

(e) I carry out the steps using StatKey in this video. Having obtained a P-value of 0.004, smaller than $\alpha = 0.05$, I reject the null hypothesis (that aspirin has no effect) in favor of the alternative, that aspirin does increase the rate of favorable outcomes.

(f) I gave a bit of an audio description of this process in the video. To get a **randomization sample**, one might take 154 (the total number of participants in the study) slips of paper, write the word "Aspirin" on 79 of them (79 is the sample size from Population A), and write "Placebo" on the other 75 slips. Place these slips in a bag and mix them up thoroughly. Prior to drawing slips, we write the 154 outcomes—the order in which we write them does not matter, only that there are 107 whose "progress" is labeled "favorable", and 47 labeled "unfavorable" (as in our original samples)—in a column:

treatment	progress
	favorable
	favorable
	⋮
	favorable
	unfavorable
	⋮
	unfavorable

We fill in the values in the empty "treatment" column one-by-one by drawing, without replacement, a slip from the bag, and recording the word from the slip in the next empty slot. Once the entire "treatment" column is filled in, we have something that looks very much like our original .csv file, but has been constructed so that treatment and progress have no association. This is a randomization sample.

To get a **randomization statistic**, we compute $\widehat{p}_A$ and $\widehat{p}_B$, then subtract them to get $\widehat{p}_A - \widehat{p}_B$, much as we computed them for part (d), but now using the randomization sample, not the sample data as it arose in the experiment.

To get a **randomization distribution**, we need to repeat the drawing of a randomization sample and the computing of a randomization statistic $\widehat{p}_A - \widehat{p}_B$ many, perhaps a few thousand, times. A dotplot or histogram of these randomization statistic gives us a fair idea of what values can arise, and how often they arise, when the null hypothesis of "no effect" holds.

(g) The two processes differ *only* in the generation of a sample. That is, you would generate a bootstrap sample differently than you would generate a randomization sample, but after that, the two processes are identical.

To describe the difference in the generation of a sample, first notice that there is more than one correct description for generating a randomization sample. I gave one description above, but another method might go like this:

> Take 154 slips of paper. Each slip should represent just one participant in the experiment. So, beginning at the top of the collected data, record the first participant's treatment and progress on the first slip. Record the values of treatment and progress for the 2nd participant on the 2nd slip, and proceed that way until all slips are used. Write the results on the slips in such a way that you can tear apart each slip, so that the left half contains the participant's treatment, and the right half contains the participant's progress. Tear apart all

154 slips, and put the "treatment" half-slip in one bag, the "progress" half-slip in another bag. Mix both bags thoroughly, then draw a half-slip randomly from the "treatment" bag with your left hand while drawing a half-slip from the "progress" bag with your right. Tape the two together, not putting them back in the bag. Repeat this process until the bags are empty; the 154 "mended" full slips form a randomization sample.

Now, if I modified the above description by

- writing down the result of my left- and right-hand draws, instead of taping half-slips together, and
- replacing each half-slip bag in the respective bags, and thoroughly mixing the contents of both bags before drawing again,

then after exactly 154 draws, I would have a sample I'd call a bootstrap sample instead of a randomization sample.

Note: One marker that an on-looker might use as an indication you produced a randomization sample is that there will always be exactly 107 favorable outcomes, 47 unfavorable outcomes, 79 aspirin-takers, and 75 placebo-takers. If that is true of the sample, it isn't 100% conclusive evidence that the sample is a randomization sample, as it may happen accidentally with a bootstrap sample, but it certainly is not guaranteed to happen with a bootstrap sample as it is with a randomization sample.

2. (a) The sampling distribution of $\overline{X}$ is centered on μ, the population mean height. A bootstrap distribution should be centered on the mean, also labeled $\overline{x}$, of the sample (i.e., of the actual data collected). A randomization distribution corresponding to null hypothesis $\mathbf{H}_0 : \mu = 68$ will be centered at 68.

   (b) You might take n notecards and write down the heights of the sampled females, one per card. Then shuffle the cards and draw one from somewhere in the deck, recording the height written on the card. Replace the card and shuffle again. Draw another at random, record the height written on it, put it back in the deck and reshuffle. Repeat this process exactly n times, so that the list of numbers you have recorded is the same length as the original list of sampled heights. This list you have recorded during your draws represents a bootstrap sample. From them, calculate the mean $\overline{x}$, which is the corresponding bootstrap statistic.

   (c) One way is to calculate the standard deviation of your bootstrap distribution; this serves as the approximate standard error $\text{SE}_{\overline{x}}$. Get your critical z^*-value using the theoretical normal distribution app (for an 80% confidence interval, we find $z^* = 1.282$, the 90th percentile in a standard normal distribution. Then compute the margin of error

$$(z^*)(\text{SE}_{\overline{x}}) = (1.282)(\text{SE}_{\overline{x}})$$

   and add/subtract this to your point estimate $\overline{x}$, the original sample mean, to obtain left and right boundaries for the confidence interval. This is the method I call the **centered interval** approach.

   The other way results in what is called **percentile bootstrap confidence interval**. It involves working directly with the bootstrap distribution to find its 10th and 90th

percentiles, taking the former as the lower bound and the latter as the upper bound of the corresponding 80% CI.

(d) The key is to change where the distribution is centered, from $\bar{x}$ (the sample mean) to 68 (the null value). Here are two modifications to the description of a bootstrap statistic (part (b)) that would achieve this goal. Whichever way you went, you would first need to compute the difference $d = 68 - \bar{x}$.

- You could modify the numbers on every index card, adding d to each. After doing so, each card no longer represents a sample person's height, but instead that height plus d. After that, the n draws with replacement would produce a randomization sample (no longer a bootstrap sample), and the average of those n draws would be a randomization statistic.
- Instead of modifying the numbers on the cards themselves, you could draw a bootstrap sample as described in (b), compute the mean of that bootstrap sample, and then add d to that mean. In this approach, which is probably easier than the first approach but equally effective, you never actually have a randomization sample; instead, you go from bootstrap sample to randomization statistic.

3. (a) Since the null value 0.25 lies inside the 90% CI, you can say the corresponding (2-sided) P-value is above 0.1.

   (b) $\mathbf{H}_0$: $p = 1/3$ vs. $\mathbf{H}_a$: $p \neq 1/3$? Since the null value $1/3 \doteq 0.333$ lies outside the 90% CI, you can say the corresponding (2-sided) P-value is below 0.1.

4. True or False. Place a "T" by those statements that are true without reservation. Place an "F" by those statements which are not unequivocally true.

   (a) True. This is precisely the descripor for Type I error.

   (b) False. The statement would be true if it concluded with "the chance of committing a Type I error is 5%."

   (c) False. Don't settle for any attractive-sounding but wrong temptations. A P-value represents the relatively frequency of obtaining a result at least as extreme as your sample result when the null hypothesis is true.

   (d) True.

   (e) False. We would only reject at this significance level when $P < 0.05$, which isn't true here.

   (f) False, in the sense that there are counterexamples. When $P = 0.04$, that is significant at the 5% level but not at the 1% level.

5. A confidence interval gives a range of plausable values for a population parameter. Ideally, it would be a small range, far smaller than the pre-supposed full list of possibilities. A 100% confidence intervals would always be the full list of possibilities, with nothing ruled out, which means you might as well not have bothered to do the work of data collection and analysis at all.