Problem 24-1

(a) The SE for the average is the SD of the (hypothetical) box model, which we estimate by the sample SD (here 30 inches) divided by the square root of the sample size: $30/\sqrt{25} = 6$. This the answer to (a) is ... elevation ... estimated as 81,411 inches ... off by 6 inches or so.

(b) True. 12 is 2 SE which is the right “plus or minus” for the confidence interval, and the parameter is described correctly.

(c) False. The parameter is not described correctly. The average of the 25 readings is exactly 81,411 inches. We know what that is, what we don’t know is the true elevation of the mountain peak. The latter is the parameter.

(d) False. That’s not what a confidence interval is either. The parameter is not “the next reading”. In fact the next reading (or any single reading) is likely to miss the true value by 30 inches or so not 12.

(e) False. No that’s not true either. The confidence interval is a statement about the location of the true population average (parameter). Not about how many readings were in an interval. In fact 95% of the readings are in the range parameter ± 2 SD of box, which is roughly 18,411 ± 60 but not quite because this interval is centered in the wrong place (at the observed value 18,411 rather than the true value, which we don’t know). Any interval something ± 12 is too narrow to contain 95% of the readings.

(f) False. This is close to correct, but still false. The width is about right. The interval true elevation ± 12 will contain the average of another 25 readings with 95% probability. But the centering is wrong. 18,411 is not the true elevation. Most likely it’s off by 6 inches or so.

Problem 24-3

The SE for the average is (approximately) $14/\sqrt{2,300} = 0.28$ km / sec. The confidence interval is $299,774 \pm 0.56$ km / sec.

Problem 24-4

We need to know the SD for the distance measurements.
Problem 24-5

If the distance is measured incorrectly, that introduces a bias. The speed is calculated using the formula: speed = distance/time. If the distance measurement is too high, the calculated speeds will also be too high (and similarly for too low).

Problem 24-6

The Gauss model assumes measurements are independent. If they aren’t, and Figure 1 makes it clear they aren’t in this case, if that wasn’t already clear from general knowledge, then the confidence interval based on the model is wrong. Usually too narrow. This is like Problem 21.6.

Problem 24-7

The question is when we use the “bootstrap” principle to estimate the SD of the hypothetical box model do we use the sample SD 20 or the historical SD 18? If we are sure nothing is changed in the measurement process, the estimate 18 is better since it is based on much more data. This gives and SE of 18/√37 = 2.55 and the confidence interval is 78.1 ± 5.1.

Problem 26-1

(a) True. “observed significance level” is another term for “P-value.”

(b) False. No, it’s the null that says the difference is due to chance.

Problem 26-2

(a) The null hypothesis is that the wheel works correctly, each pocket has a probability of 1/38, hence the probability of red is 18/38. The box model for the null hypothesis has 18’s and 20’s. The box model for the alternative hypothesis is any box with a greater proportion of 1’s (more reds).

(b)

Average of the box. 18/38 = 0.4737.

SD of the box. 0.3421 × 0.6379 = 0.4993.

Expected value of the sum. 3,800 × 0.4737 = 1,800.

SE of the sum. 3,800 × 0.4993 = 1,870.
**Test statistic (z).**

\[
z = \frac{\text{observed} - \text{expected}}{SE} = \frac{1.890 - 1.800}{0.78} = 2.92
\]

**P-value (P).** The normal curve tail area table says the area past 2.92 is 0.17% (between 0.19 and 0.16).

(c) Since the *P*-value is less than 1%, this is what the book calls “highly significant.” It does indeed appear that there are too many ones, hence something wrong with the wheel.

**Problem 26-4**

In order to “explain briefly” you have to actually do the calculations for a test.

**Null hypothesis.** Test scores on the final are considered to be like draws from a box with average 63 and SD 20. The average score for the TA’s students is like the average of 30 draws from this box.

**Expected value of the average.** 63, same as the average of the box.

**SE of the average.**

\[
SE \text{ of average} = \frac{\text{SD of box}}{\sqrt{\text{number of draws}}} = \frac{20}{\sqrt{30}} = 3.65
\]

**Test statistic (z).**

\[
z = \frac{\text{observed} - \text{expected}}{SE} = \frac{55 - 63}{3.65} = -2.19
\]

**P-value (P).** The normal curve tail area table says the area past 2.19 is 1.4%.

**Interpretation.** Since the *P*-value is a little over 1%, nearly into what the book calls the “highly significant” range, the yes-or-no answer to the question is “no.” The TA’s defense is no good. It does not seem to be just “chance variation.”

This doesn’t mean that the TA could not have a good *nonstatistical* defense. If the assignment of students to sections is not random, the TA could have had worse students in his or her section. The test of significance only shows that the “chance variation” theory is wrong. It doesn’t say anything about other explanations.
Problem 26-7

This question is tricky. Students generally assume questions like this have an easy quick answer which is in a box somewhere in the textbook. Not so here.

Two very famous statisticians, R. A. Fisher, who invented much of modern statistics, and W. S. Gosset, the inventor of the \( t \)-test (Section 6 of this chapter, which we skipped) got into a long argument about just this question, whether experimental randomization is really necessary (Fisher’s view) or whether systematic assignment could sometimes do a better job (Gosset’s view).

It is not clear that the systematic assignment used in the clinical trial must bias the results. Why would patients that came in on odd days be sicker or less helped by anticoagulant therapy? On the other hand, there is no way to be absolutely sure that there is no bias. There could be some subtle nonobvious way the patients that came in on odd days are different. Fisher argued that the only way to be sure of no bias is to use randomized treatment assignment. If you don’t, no one can ever be sure that failure to randomize didn’t bias the results. Gosset argued that randomization can also bias the results. The probability is low, but it can happen. Systematic assignment with no apparent bias could actually be better than randomization.

The general consensus these days is that Fisher was right, randomization is safer and better. But there are still holdouts for Gosset’s view.

The authors of the textbook say (in the Instructor’s Manual) that there was a bias. The doctors knew that patients admitted on odd days would be assigned to treatment and could delay admitting certain patients for one day to be sure they were assigned to the treatment group (how we were supposed to figure this out from the problem statement is unclear).

The authors then do a little calculation. Assuming a month with 30 days and assuming that patients arriving for treatment are equally likely to arrive on an even or odd day (Why? How about weekends and holidays?), then there should have been the same number (about 511) in the treatment and control groups instead of 580 in the treatment and 442 in the control.

If the assignment were really “like tossing a coin” there would have been 1022 coin tosses, the expected value for the number of heads would have been 511 and the standard error would have been \( \sqrt{1022 \times 0.5} = 16 \). So 580 in the treatment group is more than 4 SE above the EV and cannot be due to chance alone (assuming that patient’s choices of days to arrive are like coin flips, which they aren’t).

This doesn’t prove there was bias, but it does increase suspicion.

Problem 26-9

Again, in order to “explain briefly” you have to actually do the calculations for a test.

Also this question is tricky because you have to follow what is going on in this two-stage procedure. The computer is doing a loop within a loop. The
inner loop simulates 100 draws from the box $[0, 0, 0, 0, 1]$ and takes their sum. The outer loop repeats this procedure 144 times.

**Analysis of the inner loop.**

**Average of the box.** $1/5 = 0.20$.

**SD of the box.** $\sqrt{0.20 \times 0.80} = 0.4$.

**Expected value of the sum.** $100 \times 0.20 = 20$.

**SE of the sum.** $\sqrt{100} \times 0.4 = 4$.

**Analysis of the outer loop.**

We now have to model the outer loop as the sum or average of draws from a box. Each execution of the inner loop must be like one draw from the box for the outer loop. The result of the inner loop has expected value 20 and SE 4, so it is like a draw from a box with average 20 and SD 4. Seeing this is the crucial step. 144 repetitions of the outer loop is like 144 draws from a box with average 20 and SD 4. The observed number in question (21.13) is like the average of 144 draws from this outer loop box.

**Expected value of the average of draws of the outer loop box.** 20, same as the average of the box.

**SE of the average of draws of the outer loop box.** $4/\sqrt{144} = 1/3 = 0.3333$.

**Test statistic ($z$).**

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{21.13 - 20}{0.3333} = 3.39$$

**P-value ($P$).** The normal curve tail area table says the area past 3.39 is 0.035% (between 0.0404 and 0.0337, closer to the latter).

**Interpretation.** Since the $P$-value is less than 1%, this is what the book calls "highly significant." Since it is much less, maybe we should add another intensifier "very highly significant" or something like that. Anyway, it looks like the computer code is buggy.