Chapter 18 – Inference in Practice

18.1 (c) The customers who provide ratings can't be considered a random sample from the population of all customers who purchase a particular product. This is a voluntary response survey consisting only of those customers who choose to respond to the email. This is not an SRS. Anything we learn from this sample will not extend to the larger population. The other two reasons are valid, but less important, issues. Reason (a)—the size of the sample and the large margin of error—would make the interval less informative, even if the sample were representative of the population. Reason (b)—nonresponse—is a potential problem with every survey, but there is no particular reason to believe it is more likely in this situation.

18.2 (a) The 95% confidence interval is $\bar{x} \pm z^* \frac{s}{\sqrt{n}} = 1.92 \pm 1.96 \frac{1.83}{\sqrt{880}} = 1.92 \pm 0.1209 =$ 1.799 to 2.041 motorists. **(b)** The large sample size means that, because of the central limit theorem, the sampling distribution of $\bar{x}$ is roughly Normal, even if the distribution of responses is not. **(c)** Only people with listed telephone numbers were represented in the sample, and the low response rate (10.9% = 5,029/45,956) means that even that group may not be well represented by this sample.

18.3 Responses will vary. Some examples: The sample isn't random. Also, students on campus the day after final exams may not be a good representation of the entire student body.

18.4 (a) The 95% confidence interval for the mean weight of adult women will be $155 \pm 1.96 \frac{35}{\sqrt{398}} = 151.56$ to 158.44 pounds. **(b)** There is probably little reason to trust this interval; it is possible that many of the women either wouldn't know their current weight or would lie about it.

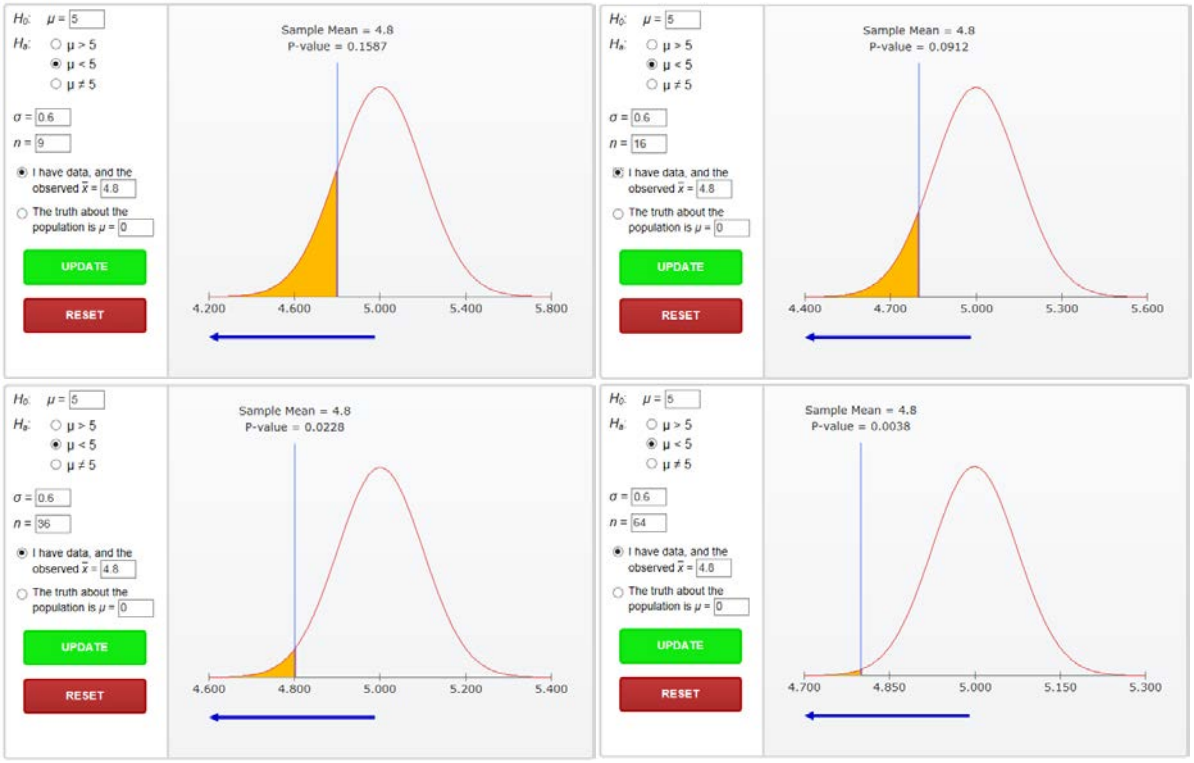
18.5 You cannot conclude this. The restaurant you work at is most likely different in many ways from the one where the experiment took place. We cannot talk about 95% of individual days from this confidence interval; the confidence interval is for the average tip, not how many days one might get that average tip.

18.6 (a) For 100 women, margin of error = $1.96 \frac{7.5}{\sqrt{100}} = 1.47$. **(b)** For 400 women, margin of error = $1.96 \frac{7.5}{\sqrt{400}} = 0.735$, and for 1600 women, margin of error = $1.96 \frac{7.5}{\sqrt{1600}} = 0.3675$. **(c)** Each time the sample size quadrupled, the margin of error was halved. This makes sense because the sample size is under a square root, and $\sqrt{4} = 2$.

18.7 (c) There is chance variation in the random selection of telephone numbers. The only source of error included in the margin of error is that due to random sampling variability. Errors due to undercoverage (such as sampling only from landline phones) and nonresponse are not included.

18.8 (a) $z = \frac{538 - 511}{120/\sqrt{50}} = 1.59$. $P(Z > 1.59) = 0.0559$; this is not quite significant at the 0.05 (5%) level. **(b)** $z = \frac{539 - 511}{120/\sqrt{50}} = 1.65$. $P(Z > 1.65) = 0.0495$; this is significant at the 5% level.

18.9 (a) and **(b)** The *P*–values and the Normal curves are shown below. We see that, as the sample size increases, the same difference between μ_0 and $\bar{x}$ goes from being not at all significant to highly significant.



18.10 The 95% confidence intervals are provided. Notice that, as the sample size increases, the margin of error becomes smaller. Also note that we would reject $H_0: \mu = 5$ with a sample size of 36 (or larger) and a two–tailed alternate hypothesis.

n	C.I. computation	Result
9	$4.8 \pm 1.96(0.6/\sqrt{9})$	4.408 to 5.192
16	$4.8 \pm 1.96(0.6/\sqrt{16})$	4.506 to 5.094
36	$4.8 \pm 1.96(0.6/\sqrt{36})$	4.604 to 4.996
64	$4.8 \pm 1.96(0.6/\sqrt{64})$	4.653 to 4.947

18.11 (a) Each test (subject) has a 1% chance of being deemed "significant" at the 1% level when the null hypothesis (no ESP) is true. With 1000 tests, we'd expect 10 such occurrences. **(b)** Retest the nine promising subjects with a different version of the test.

18.12 For a margin of error ±1, we need at least $n = (\frac{1.96 \times 7.5}{1})^2 = 216.09$, so a sample of size 217 will be needed.

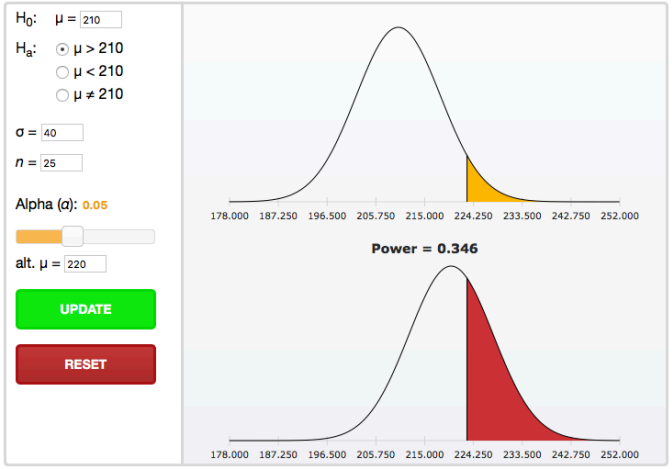
18.13 For a margin of error ±10, we need at least $n = (\frac{1.645 \times 110}{10})^2 = 327.43$, so a sample of size 328 will be needed.

18.14 (a) "Power = 0.346" means that if, in reality (unknown to the researcher), μ = 220, we will correctly reject H_0 34.6% of the time if we repeatedly sample n = 25 fourth–graders, each time conducting the significance test described. **(b)** If μ = 220, we will not reject H_0 65.4% of the time (100% – 34.6% = 65.4%) under repeated sampling, even though we should.

18.15 (a) Increase power by taking more measurements. **(b)** If you increase α, you make it easier to reject H_0, thus increasing power. **(c)** A value of μ = 225 is even further from the stated value of μ = 210 under H_0, so power increases.

18.16 The powers (obtained using the applet) are summarized in the table. The output from the applet for the first calculation from part (a) is shown.

(a)		(b)		(c)	
n	Power	μ	Power	α	Power
25	0.346	220	0.346	0.05	0.346
50	0.549	225	0.591	0.10	0.487
100	0.804	230	0.804	0.20	0.659



(a) As sample size increases (keeping everything else constant), power increases. **(b)** Keeping everything else constant, power is greater when the alternative considered is further away from 210. **(c)** Power increases when α increases, keeping everything else constant.

18.17 The table summarizes power as σ changes. As σ decreases, power increases. More precise measurements increase the researcher's ability to recognize a false null hypothesis.

σ	40	30	20
Power	0.346	0.509	0.804

18.18 (a) H_0: The patient is healthy (or "the patient should not see a doctor") and H_a: The patient is ill (or "the patient should see a doctor"). A Type I error is a false–positive, thus sending a healthy patient to the doctor. A Type II error means a false–negative—clearing a patient who should be referred to a doctor. **(b)** Answers will vary. One might wish to lower the probability of a false–negative so that most ill patients are treated, especially for serious diseases that require fast treatment. On the other hand, if resources (such as money or medical personnel) are limited, or for less serious health problems, lowering the probability of false–positives might be desirable.

18.19 (a) the data can be thought of as a random sample from the population of interest. All statistical methods are based on probability samples. We must have a random sample in order to apply them.

18.20 (c) the members of the hockey team can't be considered a random sample of all students. Especially with respect to heart rates, male athletes can't be considered representative of the population of all male students.

18.21 (b) inference from a voluntary response sample can't be trusted. To this end, inference from a voluntary response sample is never reasonable. Online Web surveys are voluntary response surveys.

18.22 (c) are in addition to the random variation accounted for by the announced margin of error. Well–designed surveys incur error due to random chance; this random variation is the only source of error accounted for in the margin of error. All forms of bias are not accounted for and are errors in addition to those due to chance.

18.23 (a) there is no control group, so the improvement might be due to the placebo effect or to the fact that many medical conditions improve over time. Because there is no control group, the researcher cannot determine if an observed improvement is due to the treatment or due to another cause.

18.24 (a) it is based on a very large random sample. The power of the test increases with sample size. That is, with a larger sample size, the small increase in life expectancy due to mild activity is more likely to be recognized.

18.25 (a) the probability that the test rejects H_0 when $\mu = 0$ is true. By definition, the significance level (α) is the probability of rejecting H_0 when H_0 is true.

18.26 (b) the probability that the test rejects H_0 when $\mu = 1$ is true. The power of the test is the probability of rejecting H_0 when H_0 is false. In this case, if $\mu = 1$, then H_0 is false, and power is the probability of rejecting H_0.

18.27 (c) describes how well the test performs when the null hypothesis is actually not true. The power of a test describes the test's ability to reject a false H_0.

18.28 We need to know that the 148 respondents were chosen at random from all general managers of three−star and four−star hotels. We also consider the possible presence of bias due to a low response rate, so we should consider the response rate.

18.29 We need to know that the samples taken from both populations (classes with attractive instructors and classes with unattractive instructors) are random. Are the samples large? Recall that if the samples are very large, then even a small, practically insignificant difference in proportion of students claiming to be highly attentive between the two samples will be deemed statistically significant.

18.30 (a) The sample described is a random sample, but women who shop at large, upscale department stores don't represent the population of all women. **(b)** Because the sample is random, the sample is likely to represent the population of all women who shop at large, upscale department stores.

18.31 Many students might be reluctant to confess that they had participated in this risky behavior. Thus, this response is likely to be biased low. The margin of error only covers random sampling errors and does not allow for this response bias.

18.32 Because we have the percents for all 13 Canadian provinces and territories, we know the exact value of μ. This assumes that the percents listed at the Web site are not estimates (though they probably are).

18.33 The effect is greater if the sample is small. With a larger sample, the impact of any one value is small.

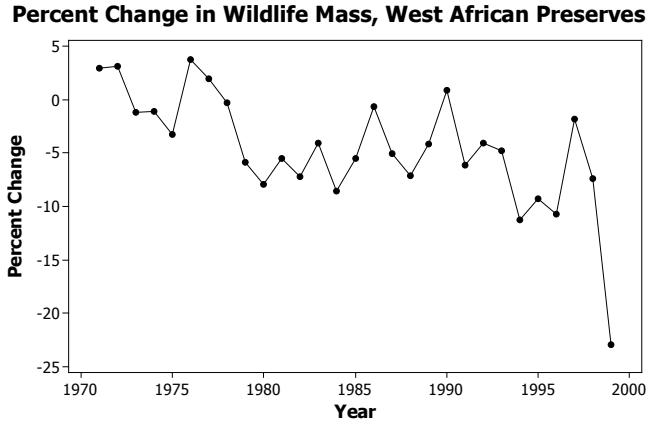
18.34 (a) A stemplot is shown. The distribution has a low outlier, which makes confidence interval methods unreliable (n = 29 observations is not a large enough sample to appeal to the central limit theorem).

Stem and Leaf

Stem	Leaf	Count
0	12334	5
-0	444321110	9
-0	998777665555	12
-1	11	2
-1		
-2	3	1
-2		

-2|3 represents -23

(b) The provided time plot shows a decreasing trend over time, so we should not treat these 29 observations as a sample coming from a single population.



18.35 Opinion—even expert opinion—unsupported by data is the weakest type of evidence, so the third description is level C. The second description refers to experiments (clinical trials) and large samples, which are the strongest evidence (level A). The first description is level B: stronger than opinion, but not as strong as experiments with large numbers of subjects.

18.36 A significance test answers only question (b). The *P*–value states how likely the observed effect (or a stronger one) is if chance alone is operating. The observed effect may be significant (it is very unlikely to be due to chance) and yet not be of practical importance. And the calculation leading to significance assumes a properly designed study.

18.37 (a) The *P*–value decreases (the evidence against H_0 becomes stronger). **(b)** The power increases (the test becomes better at distinguishing between the null and the alternative hypotheses).

18.38 It would not be reasonable to use this variable as a predictor for mortality rate. As the problem explains, the researcher tested for significance among "dozens" of candidate predictor variables. By chance alone, some of them will test as significant. The fact that mortality rates seem lower in cities with major league ballparks may just be a Type I error.

18.39 (a) The sample mean is $\bar{x} = 11.562$, so the test statistic is $z = \frac{11.562 - 10}{2.5/\sqrt{5}} = 1.4$ and the *P*–value is $P = 2P(Z \geq 1.4) = 0.1615$ (using software). This is not significant at the 5% level of significance. We would not reject 10 as a plausible value of μ, even though (unknown to the researcher) $\mu = 12$. **(b)** The small sample size makes it difficult to detect a difference that is really there.
Note: *This is an example of a test with low power; power is an optional topic.*

18.40 (a) "A significant difference ($P < 0.01$)" means that if the "affirmation training" had no effect, the chance of observing a difference in the sample performances between the two groups of women (those with and those without affirmation training) as great as that

observed would occur by chance alone less than 1 in 100 times. In other words, random chance does not really explain the observed difference in the two groups of women. **(b)** If we repeated this study, constructing a 95% confidence interval for the average difference in scores between women with training and women without training each time, then in the long run 95% of these intervals would capture the real, unknown average difference. **(c)** No, this study is not good evidence. The estimated average "improvement" is 13 points (with margin of error 8 points). We have no sense for whether 13 is a meaningful or practically important improvement. What if the gender gap is 1000 points? What if it is only 5 points (at the low end of the confidence interval)? In the former case, an improvement of 13 points would be meaningless, while in the latter case, it would be profound.

18.41 (a) "Statistically insignificant" means that the differences observed were no more than might have been expected to occur by chance, even if SES had no effect on LSAT results. **(b)** If the results are based on a small sample, then even if the null hypothesis were not true, the test might not be sensitive enough to detect the effect. Knowing the effects were small tells us that the test was not insignificant merely because of a small sample size.

18.42 $n = (\frac{2.576 \times 7}{0.1})^2 = 32{,}515.3$; take n = 32,516. This would be an unreasonable sample size, of course, and this suggests that the sample of size n = 10 used in previous exercises would be far from adequate to estimate a mean DMS threshold to within 0.1.

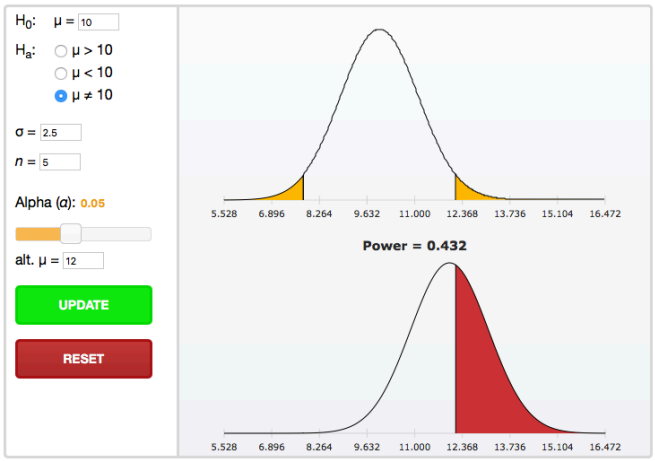
18.43 $n = (\frac{1.96 \times 3000}{600})^2 = 96.04$; take n = 97.

18.44 A low–power test may do a good job of not incorrectly rejecting the null hypothesis (that is, avoiding a Type I error), but it will often fail to reject H_0, even when it is false, simply because distinguishing between H_0 and "nearby" alternatives is difficult.

18.45 (a) This test has a 20% chance of rejecting H_0 when the alternative is true. **(b)** If the test has 20% power, then when the alternative is true, it will fail to reject H_0 80% of the time. **(c)** The sample sizes are very small, which typically leads to low–power tests.

18.46 (a) The researchers conducted a two–sided test of hypotheses at the α = 0.05 level of significance. **(b)** If there is, in fact, a clinically significant difference (a difference of at least 17 percentage points) in the presence of certain lesions 12 months after surgery, then this test would detect that difference (reject the hypothesis of no difference) 90% of the time, if the experiment was conducted repeatedly.

18.47 From the applet (screenshot provided), against the alternative μ = 12, power = 0.432.



18.48 (a) Because the alternative is $\mu > 0$, we reject H_0 at the 5% level when $z \geq 1.645$. **(b)** We reject H_0 when $3.162\bar{x} \geq 1.645$, or $\bar{x} \geq 0.5202$. **(c)** When $\mu = 0.8$, the power is $P(\bar{x} \geq 0.5202) = P(Z \geq \frac{0.5202 - 0.8}{1/\sqrt{10}}) = P(Z \geq -0.88) = 0.8106$.

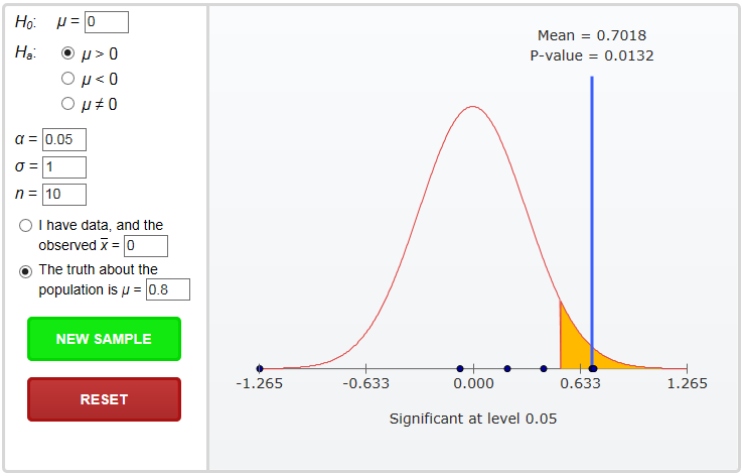
18.49 (a) The z test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 10}{2.5/\sqrt{5}} = 0.894\bar{x} - 8.944$. Because the alternative is $\mu \neq 10$, we reject H_0 at the 5% level when $z \geq 1.96$ or $z \leq -1.96$. **(b)** We reject H_0 when $z = 0.894\bar{x} - 8.944 \geq 1.96$ (that is, $\bar{x} \geq 12.197$) or $z = 0.894\bar{x} - 8.944 \leq -1.96$ (that is, $\bar{x} \leq 7.812$). **(c)** When $\mu = 12$, the power is $P(\bar{x} \geq 12.197) + P(\bar{x} \leq 7.812) = P(Z \geq \frac{12.197 - 12}{2.5/\sqrt{5}}) + P(Z \leq \frac{7.812 - 12}{2.5/\sqrt{5}}) = P(Z \geq 0.18) + P(Z \leq -3.75) \approx 0.4286 + 0 = 0.4286$.

18.50 The probability of committing a Type I error is $\alpha = 0.01$. The probability of a Type II error is 1 – power = 1 – 0.90 = 0.10.

18.51 Power = 1 – P(Type II error) = 1 – 0.49 = 0.51.

18.52 (a) P(Type I error) = P(reject H_0 when H_0 is true) = P ($\bar{x} > 0$ given that $\mu = 0$) = 0.5, because $\bar{x}$ has a Normal distribution with mean 0. **(b)** If $\mu = 0.5$, then $\bar{x}$ has a Normal distribution, with mean 0.5 and standard deviation $\sigma/\sqrt{n} = 2.5/\sqrt{25} = 0.5$. Thus P(Type II error) = P(fail to reject H_0 when specific H_a is true) = P ($\bar{x} \leq 0$ given that $\mu = 0.5$) = $P(Z \leq \frac{0 - 0.5}{0.5}) = P(Z \leq -1) = 0.1587$. **(c)** If $\mu = 1$, then $\bar{x}$ has a Normal distribution with mean 1 and standard deviation $\sigma/\sqrt{n} = 2.5/\sqrt{25} = 0.5$. Thus P(Type II error) = P ($\bar{x} \leq 0$ given that $\mu = 1$) = $P(Z \leq \frac{0 - 1}{0.5}) = P(Z \leq -2) = 0.0228$.

18.53 (a) In the long run, this probability should be 0.05. Out of 100 simulated tests, the number of false rejections will have a binomial distribution with $n = 100$ and $P = 0.05$. Most students will see between 0 and 10 rejections. **(b)** If the power is 0.812, the probability of a Type II error is 0.188. Out of 100 simulated tests, the number of false non–rejections will have a binomial distribution with $n = 100$ and $P = 0.188$. Most students will see between 10 and 29 non–rejections. One rejection result is shown below.



18.54 to **18.57** are Web–based exercises.