Final Review

(If you catch any errors in the solutions, let me know!)

1. A recent housing survey was conducted to determine the price of a typical home in Glendale, CA. Glendale is mostly middle-class, with one very expensive suburb. The mean price of a house was roughly $650,000. Which of the following statements is most likely to be true?

   (a) Most houses in Glendale cost more than $650,000.

   (b) Most houses in Glendale cost less than $650,000.

   (c) There are about as many houses in Glendale that cost more than $650,000 than less than this amount.

   (d) We need to know the standard deviation to answer this question

2. The table below shows some summary statistics of the distributions of resident tuition at public and private medical schools.

	Min	Q1	Median	Q3	Max
Private	$6,550	$30,729	$33,850	$36,685	$41,360
Public	$0	$10,219	$16,168	$18,800	$27,886

   Determine which of the following statements is true about the spread for medical school resident tuition.

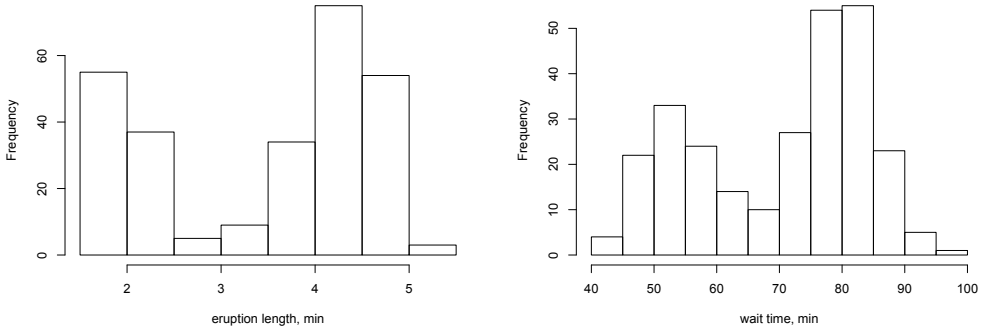
   (a) The ranges of the two distributions are roughly equal indicating that the variability is the same for the two distributions.

   (b) There is more variation in tuitions for residents at public medical schools than at private medical schools since the interquartile range is higher for public schools.

   (c) There is more variation in tuitions for residents at private medical schools than at public medical schools since there are outliers for private schools.

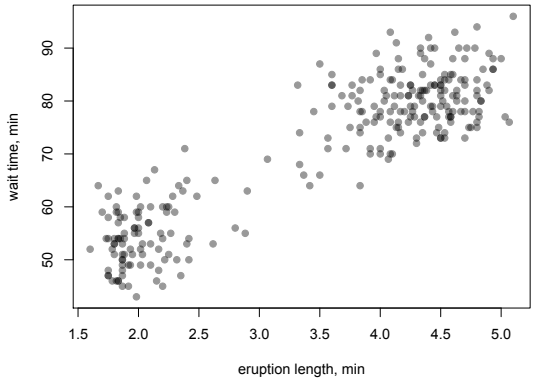
   (d) With these data, we cannot compare the variations of tuitions for residents at private and public medical schools.

3. The histograms below show the distributions of the duration of eruptions and waiting time between eruptions at the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.



(a) Comment on the modality of the two variables.

(b) Below is a scatterplot of the two variables. Describe the relationship and comment on how this scatterplot relates to the histograms.



4. It is known that 80% of people like peanut butter, 89% like jelly, and 78% like both.

(a) If we pick one person at random, what is the chance s/he likes peanut butter or jelly?

(b) How many people like either peanut butter or jelly, but not both?

(c) Suppose you pick out 8 people at random, what is the chance that exactly 1 of the 8 likes peanut butter but not jelly?

(d) Are "liking peanut butter" and "liking jelly" disjoint outcomes?

(e) Are "liking peanut butter" and "liking jelly" independent outcomes?

5. The cholesterol levels for women aged 20 to 34 follow an approximately Normal distribution with mean 185 milligrams per deciliter (mg/dl) and standard deviation 39 mg/dl. About 18.5% of women are known to have high cholesterol. What is the cutoff cholesterol level for for being considered as having high cholesterol?

6. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

7. About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical? [1]

8. Twenty-year-old men's weights have a population mean of 155 pounds and a population standard deviation of 22 pounds (www.kidsgrowth.com). The distribution is right-skewed. Suppose we take a random sample of 100 twenty-year-old men's weights.[2]

   (a) Explain why the Central Limit Theorem is applicable.

   (b) Sketch the sampling distribution of means. Label the mean, the mean $\pm$ 1, 2, and 3 standard errors, and indicate what percent of the distribution falls in each region.

   (c) What percent of random samples of 100 twenty-year-old men have means between 150.6 and 161.6 pounds?

9. It is believed that nearsightedness affects about 8% of all children. Would it be considered unusual if 21 our of 194 randomly sampled children are nearsighted? Explain your reasoning.

10. A clinical trial with 400 subjects was conducted to test whether the average weight loss using a new diet pill is significant. You know that the standard deviation of the test subject's weight loss was 5lbs. Give an example of sample mean weight loss which would result in rejecting the null hypothesis at the 5% level but not rejecting it at the 1% level. Hint: We only care if the subjects lost weight and not gained it, think about what tails are appropriate and start by writing the appropriate hypotheses.

11. In June 2002, the *Journal of Applied Psychology* reported on a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same nine commercials. One of the programs had a violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products what were advertised. Results are summarized in the table below.[3]

[1] Adapted from *Statistics: A Bayesian Perspective* by Berry.
[2] Adapted from *Introductory Statistics* by Gould and Ryan.
[3] Adapted from *Intro Stats*, by De Veaux, Velleman, and Bock.

	Program Type		
	Violent	Sexual	Neutral
No. of subjects	101	106	103
Brands recalled			
Mean	3.02	2.72	4.65
SD	1.61	1.85	1.62

For the following questions, you may assume that all assumptions and conditions necessary for inference are satisfied.

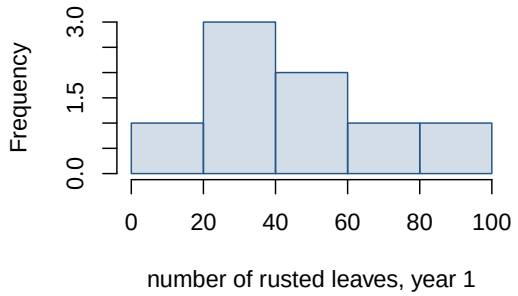
(a) Is there a significant difference in viewers' abilities to remember brands advertised in shows with **violent** vs. **neutral** content? Use a hypothesis test to answer this question and make sure to interpret your conclusion in context.

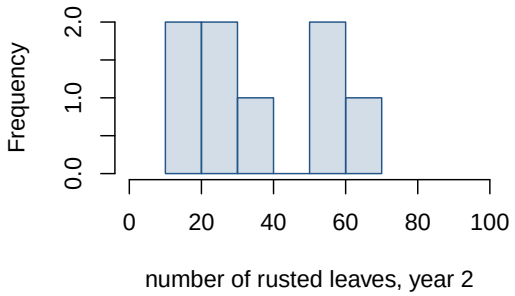
(b) Construct a 95% confidence interval for the difference in number of brand names recalled between the groups watching shows with **sexual** content and those watching **neutral** shows. Interpret your interval in context.

(c) What type of test would we use if we wanted to compare all three means simultaneously.

12. Cedar-apple rust is a (non-fatal) disease that affects apple trees. Its most obvious symptom is rust-colored spots on apple leaves. Red cedar trees are the immediate source of the fungus that infects the apple trees. If you could remove all red cedar trees within a few miles of the orchard, you should eliminate the problem. In the first year of an experiment the number of affected leaves on 8 randomly sampled trees were counted; the following winter all red cedar trees within 100 yards of the orchard were removed and the following year the same trees were examined for affected leaves. The results are recorded below:[4]

tree	number of rusted leaves: year 1	number of rusted leaves: year 2	difference year 1 - year 2
1	38	32	6
2	10	16	-6
3	84	57	27
4	36	28	8
5	50	55	-5
6	35	12	23
7	73	61	12
8	48	29	19
average	46.8	36.2	10.5
standard dev	23	19	12





[4]Adapted from http://www.physics.csbsju.edu/stats/t-test.html

(a) Write the hypotheses for testing for a difference between the average number of rusted leaves between years 1 and 2.

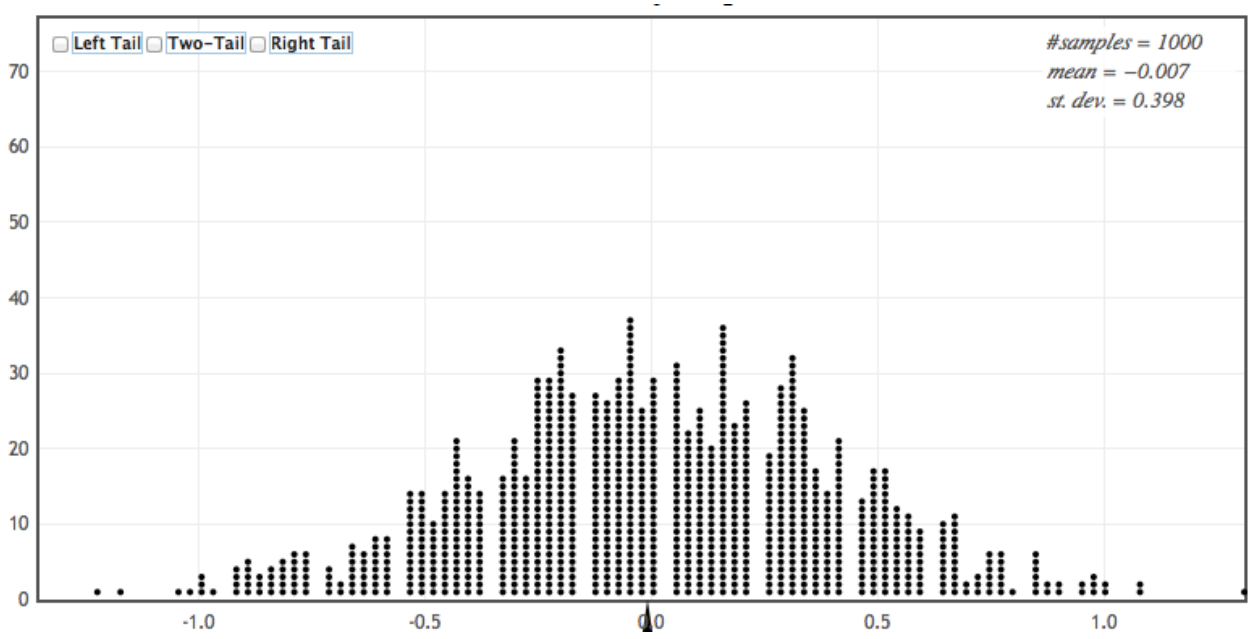
(b) Check the assumptions and conditions necessary for inference and determine the most appropriate statistical method for evaluating these hypotheses.

(c) Calculate the test statistic and the p-value.

(d) What is the conclusion of the hypothesis test?

(e) What is the p-value for testing if the number of rusted leaves has <u>decreased</u> from year 1 to 2. Give an interpretation of this new p-value.

13. Can a simple smile have an effect on punishment assigned following an infraction? "Why smiles generate leniency", LaFrance & Hecht (1995), examines the effect of a smile on the leniency of disciplinary action for wrongdoers. Participants in the experiment took on the role of members of a college disciplinary panel judging students accused of cheating. For each suspect, along with a description of the offense, a picture of one of 34 students was provided. Each student had a picture where they smiled and one where they had a neutral facial expression. A leniency score was calculated based on the disciplinary decisions made by the participants. Suppose the experimenters have prior knowledge that smiling has a positive influence on people, and they are testing to see if the average lenience score is higher for smiling students than it is for students with a neutral facial expression (or, in other words, that smiling students are given milder punishments.) In the original sample the average leniency score for smiling students was 4.912 and for students with a neutral expression was 4.118. The figure below shows the distribution of the difference between the sample means from 1,000 randomization samples.[5]



(a) Are the two groups dependent or independent?

(b) Write the appropriate hypotheses.

(c) What is the conclusion of the hypothesis test?

[5]Adapted from *Unlocking the Power of Data* by Lock, Lock, Lock, Lock, and Lock.

14. A USA Today/Gallup poll conducted between Dec. 21, 2010 - Jan. 9, 2011 asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 1,145 unemployed and 675 underemployed people were randomly sampled and surveyed. 27% of the unemployed and 25% of the underemployed people said they had major problems in relationships as a result of their employment status.

    (a) Are conditions for inference satisfied?

    (b) Construct a 95% confidence interval to estimate the difference between the true population proportions of unemployed and underemployed people who had such problems.

    (c) Conduct a hypothesis test to evaluate if these data provide convincing evidence to suggest a difference between the true population proportions of unemployed and underemployed people who had such problems. Use a 5% significance level.

15. As recently as 2008, 70% of users of social networking sites such as Facebook were 35 years old or younger. Now the age distribution is much more spread out. The table below shows the age distribution of 975 users of social networking sires from a Pew Research survey from June 2011.[6]

Age	18 - 22	23 - 35	36 - 49	50 - 65	65 +
Frequency	156	312	253	195	59

    (a) Test an assumption that users are equally likely to be in each of the five age groups listed. Show all details of the test.

    (b) Which age group contributes the largest amount to the test statistic? For the age group, is the observed count smaller or larger than the expected count?

16. A random sample of 3,052 cell phone users were asked "Do you send or receive text messages on your cell phone?" Of the 800 teens 696 said they did, and of the 2252 adults 1621 said they did. The histogram below shows the bootstrap distribution for the difference between the proportions of teens and adults who send or receive text messages on their cell phones. The plot below shows the distribution for the difference between the proportions of teens and adults ($\hat{p}_{teen} - \hat{p}_{adult}$) in 10,000 **bootstrap** samples.[7]

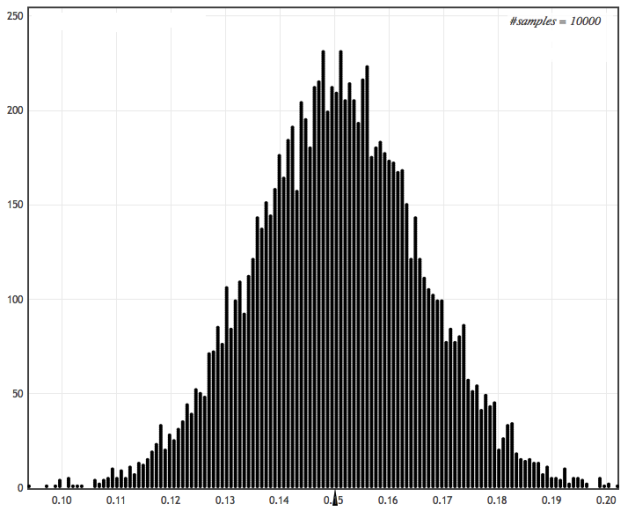
[6] From *Unlocking the Power of Data* by Lock, Lock, Lock, Lock, and Lock.
[7] Adapted from *Unlocking the Power of Data* by Lock, Lock, Lock, Lock, and Lock.

(a) Estimate the mean of the bootstrap distribution.

(b) Which of the below is the most reasonable estimate of the standard error of the difference in proportions? Explain your reasoning for your choice.

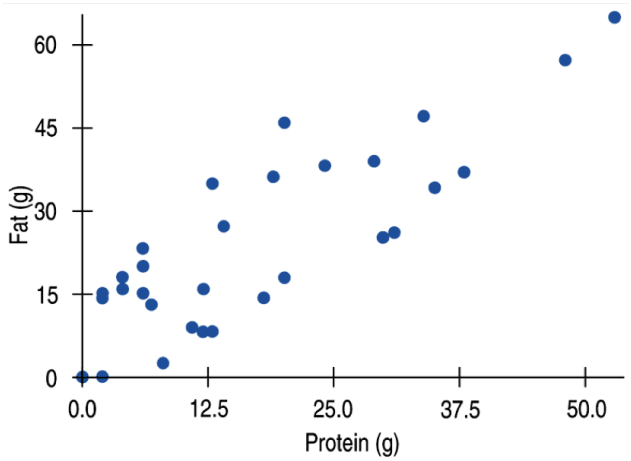
(a) 0.015 (c) 0.05
(b) 0.03 (d) 0.15

(c) Using your choice for SE in part (b), find an interval estimate for the difference in proportions of teen and adult cell phone users who send/receive text messages.



17. In a state where the political race is tight you have the job of conducting a survey on a simple random sample of voters from the state to determine who has the edge. As a pollster, you know that to give your poll credit, you need to ensure the estimate is within 4 percentage points (with 99% confidence). Determine the minimum sample size that will ensure this accuracy.

18. The mean fat content of the 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83. A scatterplot of the data is shown below.



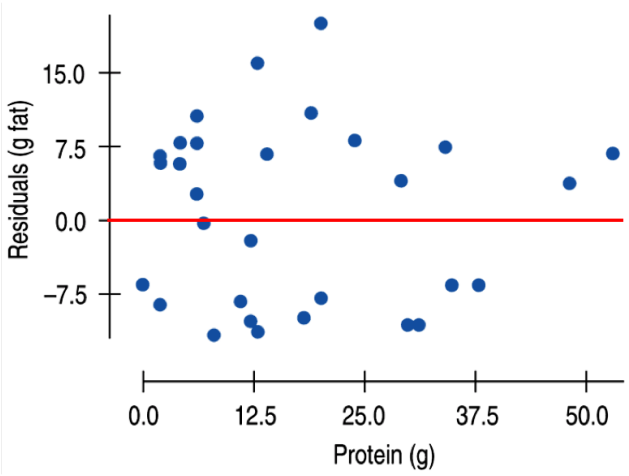
(a) Write the linear model for predicting fat content.

(b) Interpret the slope and the intercept in context.

(c) A new BK menu item has 30 grams of protein. Can we use this model to predict its fat content? If so, what is it?

(d) Another new BK menu item has 75 grams of protein. Can we use this model to predict its fat content? If so, what is it?

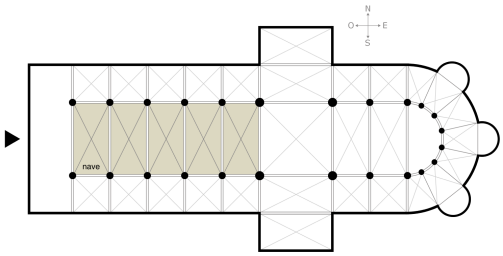
(e) Based on the residuals plot shown below, does the linear model appear to be appropriate for these data?



(f) Calculate R^2 and interpret it in context.

19. Data was collected on nave heights (ft) and total lengths (ft) of 25 English medieval cathedrals (excerpt shown below).

name	style	nave height	total length
Durham	roman	75.00	502.00
Canterbury	roman	80.00	522.00
⋮			
Old St Paul	gothic	103.00	611.00
Salisbury	gothic	84.00	473.00



(a) The regression model for predicting total length from name height and style is also given below. Which of the following is false?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.2979	81.6477	0.54	0.5929
naveheight	4.7116	1.0584	4.45	0.0002
style_roman	80.3925	32.3063	2.49	0.0209

$R^2_{adj} = 49.64\%$

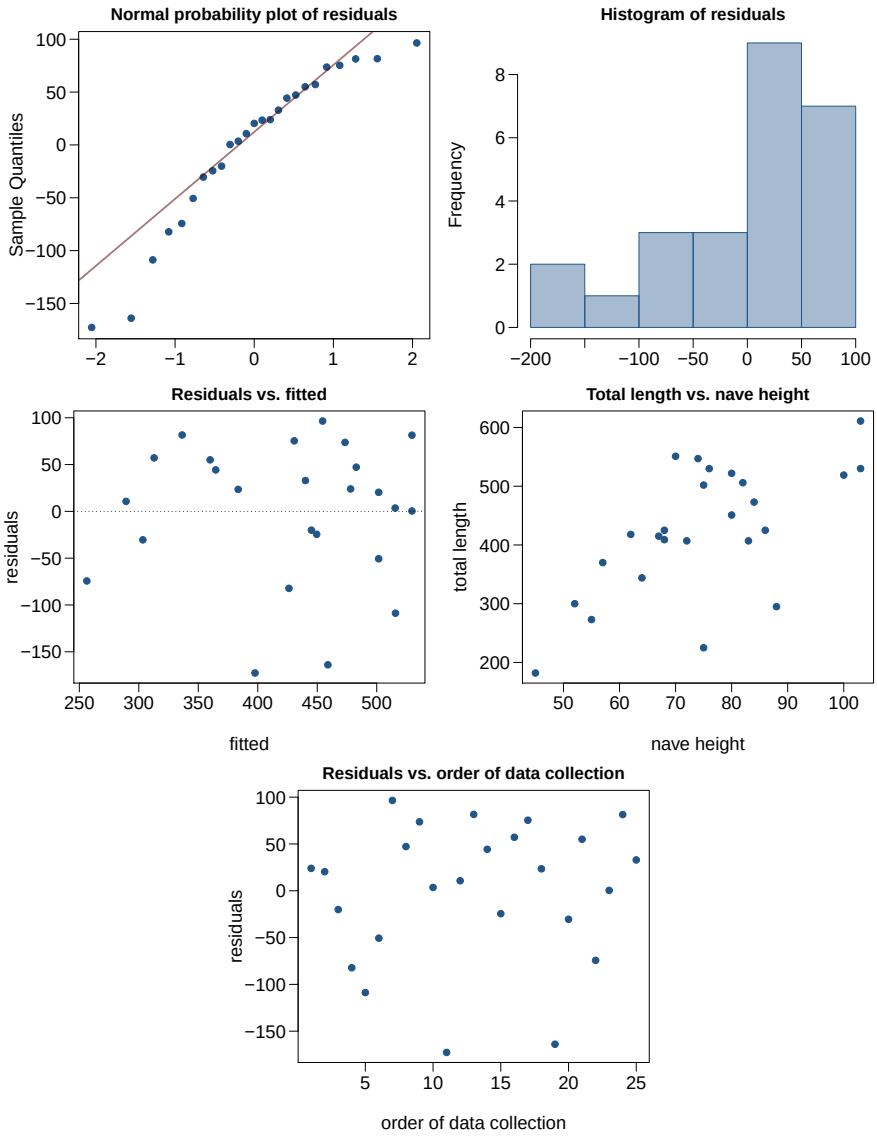
(A) Each additional foot in nave height is associated with a 4.7116 foot increase in total length.

(B) Roman cathedrals with 0 nave height are expected on average to be 44.2979 feet in total length.

(C) Roman cathedrals are expected on average to be 80.3925 feet longer than Gothic cathedrals.

(D) Both nave height and style of cathedral are significant predictors of total length.

(b) Using the plots provided below check if the assumptions and conditions for MLR is satisfied.



20. The Nielsen organization did a poll to determine whether men and women in different age groups watched different amounts of comedy television. The table below shows some summary statistics on each age/gender group.[8]

level	n	mean	sd
men 18-34	10	287	65.05
men 55+	10	171	40.81
women 18-34	10	353.70	20.78
women 55	10	356.90	39.37

(a) What method can we use to evaluate if different age/gender groups watch different amounts of comedy television on average. Explain your reasoning.

(b) Do assumptions and conditions for this technique appear to be satisfied?

21. National Health and Nutrition Examination Survey (NHANES) collects data on people's cholesterol levels and marital status, among many other variables. The data come from 940 respondents and marital status has 6 levels (divorced, living with partner, married, never

[8] Adapted from *Introductory Statistics* by Gould and Ryan.

married, separated, widowed). Below is the relevant ANOVA table.[9]

(a) Write the hypotheses for evaluating if average cholesterol level varies among people with different marital statuses.

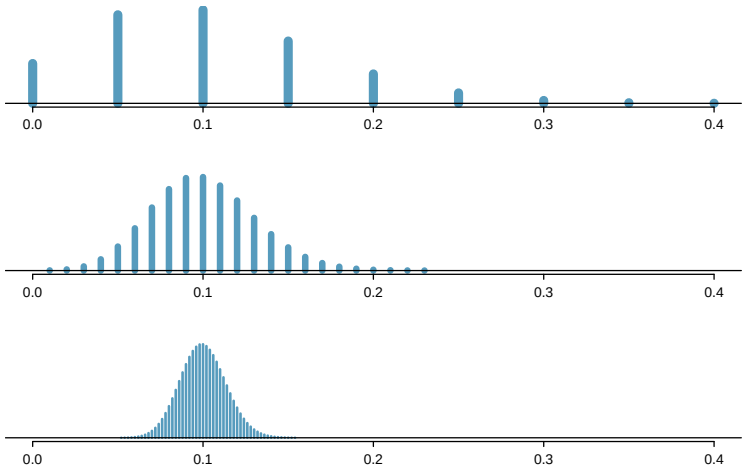
(b) Below is the relevant ANOVA table. Fill in the blanks.

	df	SS	MS	F	p-value
marital_st		89,082			<0.0001
Residuals					
Total		1,909,292			

(c) What is the conclusion of the analysis?

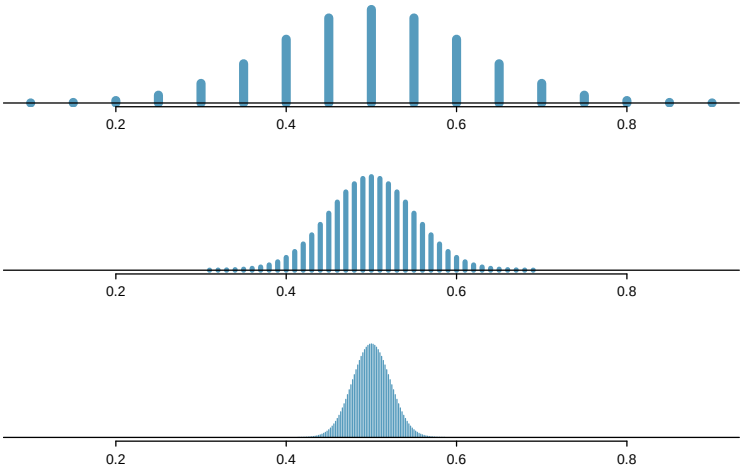
(d) Assuming you did find an association between marital status and cholesterol levels, would this association mean that marital status caused different cholesterol levels? Can you think of a confounding factor?

22. Suppose the true population proportion were $p = 0.1$. The figure below shows what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$. What does each observation in each distribution represent? Describe how the distribution of the sample proportion, $\hat{p}$, changes as n becomes larger.

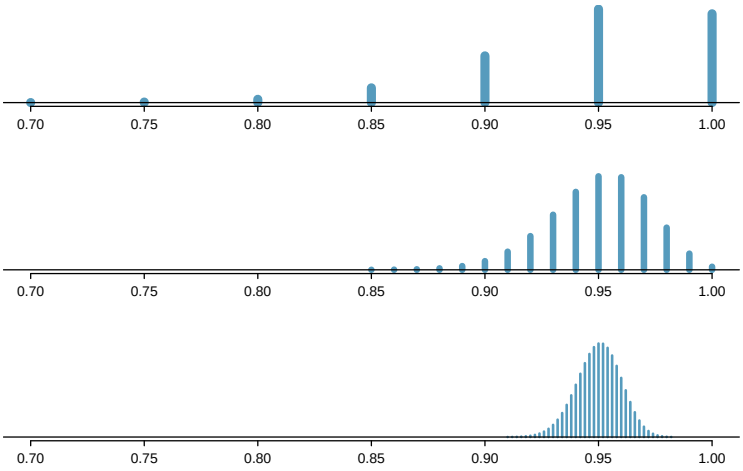


23. Suppose the true population proportion were $p = 0.5$. The figure below shows what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$. What does each observation in each distribution represent? Describe how the distribution of the sample proportion, $\hat{p}$, changes as n becomes larger.

[9] Adapted from *Introductory Statistics* by Gould and Ryan.



24. Suppose the true population proportion were $p = 0.95$. The figure below shows what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$. What does each observation in each distribution represent? Describe how the distribution of the sample proportion, $\hat{p}$, changes as n becomes larger.



25. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it isnt perfect. The study reported a standard error of about 1.2%, and a normal model may be reasonably be used in this setting.

    (a) Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

    (b) Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

        i. We can say with certainty that the confidence interval contains the true percentage of U.S. adults who suffer from a chronic illness.

        ii. If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

iii. The poll provides statistically significant evidence (at $\alpha = 0.025$) that the percentage of U.S. adults who suffer from chronic illnesses is below 50%.

iv. Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

26. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened. We are interested in evaluating whether people are equally likely to intervene when the woman is wearing a provocative or conservative outfit.

		Scenario		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.