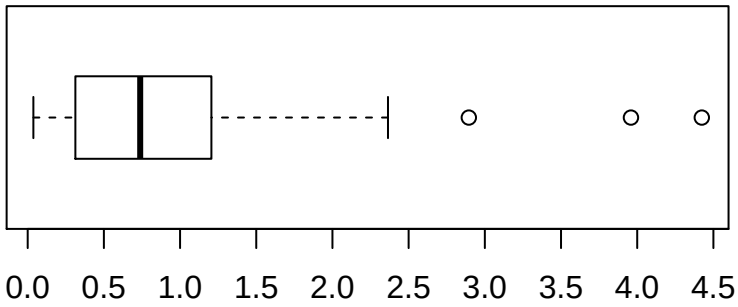
Stat 185 - Review problems for the final exam

1. Suppose that a study is conducted by asking students what score they think they are going to get on a test and then measuring their blood pressure. The goal is to see if expected score has an effect on blood pressure.
   (a) What is the predictor variable and what is the response variable? For each variable determine if it is categorical or numeric.
   (b) Is this study an observational study or a designed experiment?
2. Answer the following questions about this box and whisker plot (you only have to get reasonably close):



   (a) Estimate the mean
   (b) Estimate the IQR
   (c) Estimate the highest observation
   (d) Estimate the 4th highest observation
   (e) Estimate the lowest observation
   (f) Discuss the shape; do you think a normal distribution would be useful to model this data?
3. Draw a box and whisker plot from the following summary of a list of numeric data:

Min	1st Qu	Median	Mean	3rd Qu	Max.
-0.8356	-0.5462	0.2566	0.1322	0.5537	1.5953

4. Suppose that 20% of all people have green eyes. Use a normal distribution to estimate the probability that more than 20 of the sixty lucky students taking class with me next semester will have green eyes.

5. An organic farmer sells bags of Brussels sprouts labeled as one pound each. Let's use a t-test to investigate the farmer's claim that bags each contain one pound. I purchased one bag each of 4 weeks during the Summer farmer's market, weighed them at home and recorded the following weights:

1.05	1.15	1.2	1.1

   (a) Write down the hypothesis statement for the problem.

   (b) Write down a formula showing that the mean is $\mu = 1.125$

   (c) Write down a formula showing that the standard deviation is $\sigma = 0.06454972$.

   (d) Assuming the mean $\mu_0 = 1$ is correct, compute the t-score for the observed mean.

   (e) Compute a 95% confidence interval for the average weight of a bag of these Brussels sprouts.
   Be sure to use a t-distribution with the correct degrees of freedom.

   (f) Use a t-table to find the critical t-value for a 95% level of confidence for this problem.

6. Consider the data

x	5	8	9	5
y	6	7	9	4

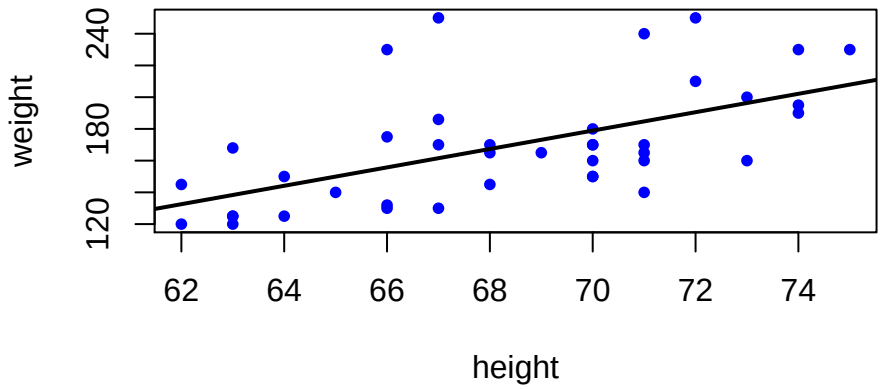
   Note that the standard deviation of x is 1.78, the standard deviation of y is 1.8, and the correlation between x and y is 0.10675.

   (a) Sketch a scatter plot of this data.

   (b) Setch the regression line for the data.

   (c) Find an equation of the regression line relating x and y.

7. The figure below shows a scatter plot of weight (in pounds) vs height (in inches) for a random sample 40 American men and the table shows linear regression table for that same data. We are curious if there is a relationship between height and weight.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-226.9576	89.4932	-2.54	0.0154
height	5.7984	1.3065	4.44	0.0001

(a) Write down the hypothesis statement that we would check with this linear model.

(b) Write out the formula for linear model.

(c) What does the model predict for the weight of a man who is 70 inches tall?

(d) State the conclusion of the hypothesis test from part (c).

8. On average, the 127 college football teams in Division I gained 5322 yards throughout the 2016 season with a standard deviation of 887 yards.

(a) Use a normal model to estimate the number of teams that rushed more than 6000 yards.

(b) A histogram for this data is shown in figure 1. Based on the histogram, do you see any issues with using a normal model for this estimate?

9. UNCA claims that 90% of it's students are from North Carolina but we suspect that it might be less than that. Suppose we draw a random sample of 50 UNCA students and find that 41 of them are in state. Let's use this to explore the claim that 90% of UNCA students are in-state vs the possibility that fewer than 90% of UNCA students are in-state.

   (a) Construct appropriate null and alternative hypotheses to explore our question.

   (b) Compute the associated p-value.

   (c) Based on a 95% level of confidence, what is the conclusion of our hypothesis test?

10. A recent study surveyed people's attitudes on the relative importance of economic issues. In particular, the study examined the question - what is more important:

    - Reducing the federal deficit or
    - Cutting taxes for corporations

    The results are summarized in table 1. Let p_D denote the proportion

	Democrats	Republicans
Deficit over taxes	97	66
Taxes over deficit	43	48

Table 1: Views on deficit reduction over tax reduction

    of democrats who think that reducing the deficit is the more important of the two issues and let p_R denote the proportion of republicans who think that reducing the deficit is the more important of the two issues. We suspect that $p_D < p_R$ so let's use a hypothesis test to examine this question.

    (a) Compute the observed proportions $\hat{p}_R$ and $\hat{p}_D$, as well as the difference
    $$\hat{p} = \hat{p}_D - \hat{p}_R.$$

    (b) Compute the standard error and the test statistic.

    (c) Use a normal table to compute the p-value.

    (d) State the conclusion of the test.

(e) Why was it OK to use a normal table?

11. A statistics professor teaching two sections of the same introductory statistics course is concerned that exam scores in one section (let's call it section A) are noticeably lower than exam scores in another section (section B). To test this hypothesis, he computed the mean and standard deviation for the combined scores on the first two exams for both classes and found the following:

    Section A: Mean=118.4, std dev = 40, number of students = 30

    Section B: Mean=133.5, std dev = 32.7, number of students = 30

    Viewing these data as samples from a random process, find

    (a) The difference between the two means.

    (b) The associated standard error and test statistic.

    (c) Compute the p-value using a normal distribution.

    (d) From the computations, can we conclude with a 95% level of confidence that section A scores lower than class B?

12. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

    (a) State the hypotheses for testing if the professor's predictions were inaccurate.

    (b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

    (c) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the t^*-value for a 95% level of confidence

    (d) Based on the t^*-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

genhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender
good	0	1	0	70	175	175	77	m
good	0	1	1	64	125	115	33	f
good	1	1	1	60	105	105	49	f

Table 2: First few lines of our CDC data

13. Table 1 shows the first few rows of our CDC data set. Match the following types of analysis with the hypothesis test that you might use to perform the analysis. In each case, write down the appropriate hypotheses.

    (a) Extract the exerany column and the smoke100 column and examine whether the proportion of exercisers exceeds the proportion of smokers.

    (b) Extract a sample of size 15 from the height column and examine whether the average height is some particular value.

    (c) Extract the weight column and examine whether the average weight is some particular value.

    (d) Extract the exerany column and the smoke100 column and examine examine the relationship between them.

    (e) Extract the gender column and examine whether the proportion of men is 1/2.

    (f) Extract the height column and the weight column and examine examine the relationship between them.

(i) One sample proportion test with the t-distribution

(ii) Linear regression

(iii) Difference of two sample proportions test with the t-distribution

(iv) χ^2 test for independence

(v) Difference of two sample means test with the normal distribution

(vi) One sample mean test with the normal distribution

(vii) χ^2 test for homogeneity

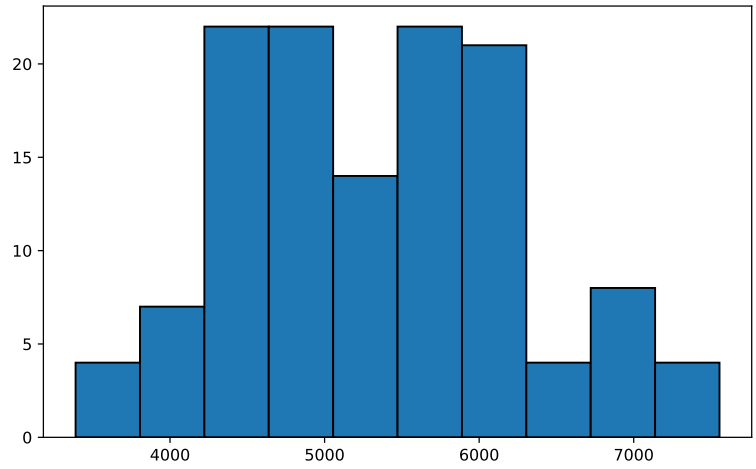


Figure 1: A histogram of total yards gained by each team for the 2016 season