(Q.1)

A researcher discovers the following information about the association between Variable X and Variable Y:

Mean of X = 50.49 Standard deviation of X = 12.83

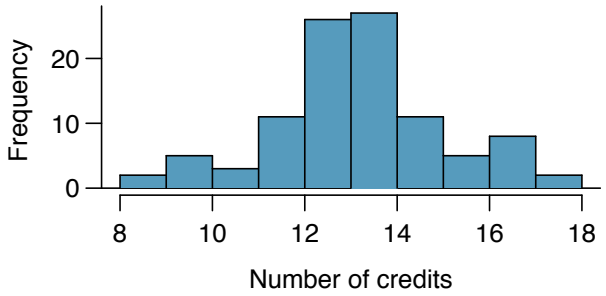
Mean of Y = 18.30 Standard deviation of Y = 4.11

r = –0.71

Calculate a and b in the regression equation ($Y' = a + bX$).

(Q.2)

7.50 College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.



Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

(e) The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original sample.

(Q.3)

Gas mileage A student runs an experiment to study the effect of three different mufflers on gas mileage. He devises a system so that his Jeep Wagoneer uses gasoline from a one-liter container. He tests each muffler 8 times, carefully recording the number of miles he can go in his Jeep Wagoneer on one liter of gas. After analyzing his data, he reports that the F-ratio is 2.35 with a P-value of 0.1199.

a) What are the null and alternative hypotheses?

b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?

c) What would you conclude?

d) What else about the data would you like to see in order to check the assumptions and conditions?

e) If your conclusion in part c is wrong, what type of error have you made?

(Q.4)

7.51 Hen eggs. The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

(Q.5)

When constructing a confidence interval for the mean with σ known, how is the standard error of the mean calculated?

When constructing a confidence interval for the mean with σ unknown, how is the standard error of the mean estimated?

(Q.6)

A sample of 25 program participants in an alcohol rehabilitation program are administered a test to measure their self-reported levels of alcohol in- take prior to entering the program. Results indicate an average (X) of 4.4 drinks per day for the sample of 25, with a sample standard deviation (s) of 1.75 drinks. Based on that information, develop a 95% confidence interval to provide an estimate of the mean intake level for the entire population of program participants (μ).

(Q.7)

Assume you've administered a worker satisfaction test to a random sample of 25 workers at your company. The test is purported to have a population standard deviation or σ of 4.50. The test results reveal a sample mean (X) of 78. Based on that information, develop an estimate of the mean score for the entire population of workers, using a 95% confidence interval.

(Q.8)

Yogurt An experiment to determine the effect of several methods of preparing cultures for use in commercial yogurt was conducted by a food science research group. Three batches of yogurt were prepared using each of three methods: traditional, ultra, and reverse os- mosis. A trained expert then tasted each of the 9 samples, presented in random order, and judged them on a scale from 1 to 10. A partially completed Analysis of Variance table of the data follows.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Ratio
Treatment	17.300			
Residual	0.460			
Total	17.769			

a) Calculate the mean square of the treatments and the mean square of the error.

b) Form the F-statistic by dividing the two mean squares.

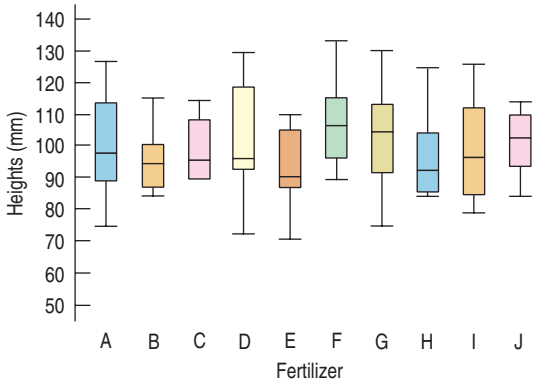
c) The P-value of this F-statistic turns out to be 0.000017. What does this say?

d) What assumptions have you made in order to answer part c?

e) What would you like to see in order to justify the conclusions of the F-test?

(Q.9)

Fertilizers A biology student is studying the effect of 10 different fertilizers on the growth of mung bean sprouts. She sprouts 12 beans in each of 10 different petri dishes, and adds the same amount of fertilizer to each dish. After one week she measures the heights of the 120 sprouts in milli- meters. Here are boxplots and an ANOVA table of the data:



Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
Fertilizer	9	2073.708	230.412	1.1882	0.3097
Error	110	21331.083	193.919		
Total	119	23404.791			

a) Check the conditions for ANOVA.
b) Write the null and alternative hypotheses.
c) What does the ANOVA table suggest about your inference? (Be sure to report this in terms of heights and fertilizers).
d) Her lab partner looks at the same data and says that he did *t*-tests of every fertilizer against every other fertilizer and finds that several of the fertilizers seem to have significantly higher mean heights. Does this match your finding in part b? Give an explanation for the difference, if any, between the two results.

(Q.10)

A data set consists of 1 million pairs of x,y observations, with the following summary statistics

x-variable: mean = 34.5, SD = 12
y-variable: mean = 480, SD = 60

A scatter plot suggests the relationship between x, y is approximately linear, with correlation r = –0.56. A data scientist, as part of an exploratory analysis, decides to standardize the variables and call them z_x, z_y.

a. Find the mean and standard deviation of z_x, z_y.
b. Find the correlation between z_x, z_y.
c. Find the line of best fit for predicting z_y from z_x.
d. If the x-value of an observation is 2 SD below the mean x-value, what is the predicted y-value?

(Q.11)
Select the right answer: A P-value indicates

A) the probability that the null hypothesis is true.
B) the probability that the alternative hypothesis is true.
C) the probability the null is true given the observed statistic.
D) the probability of the observed statistic given that the null hypothesis is true.
E) the probability of the observed statistic given that the alternative hypothesis is true.

(Q.12)
A statistics professor wants to see if more than 80% of her students enjoyed taking her class. At the end of the term, she takes a random sample of students from her large class and asks, in an anonymous survey, if the students enjoyed taking her class. Write the hypotheses she should test.

(Q.13)
Suppose that a manufacturer is testing one of its machines to make sure that the machine is producing more than 97% good parts ($H_0: p = 0.97$ and $H_A: p > 0.97$). The test results in a P-value of 0.122, and the manufacturer fails to reject the null hypothesis at a 5% significance level. Unknown to the manufacturer, the machine is actually producing 99% good parts.

a. What kind of error is this (Type I or Type II)?
b. To reduce the risk of this type of error should they increase or decrease α?
c. What is the downside of doing that?

(Q.14)
The weights of hens' eggs are normally distributed with a mean of 56 grams and a standard deviation of 4.8 grams. What is the probability that a dozen randomly selected eggs weigh over 690 grams?

(Q.15)
State the null and alternative hypotheses and explain under what situation a type I error would be committed.

1. You are testing whether the average number of cups of coffee that people drink daily has changed from 2.6 cups since 2003.
2. Architects at a firm claim their salaries are less than the State average of $48,520.

(Q.16)
What factors makes a video go viral? A researcher sets out to test the effect of the following factors on the number of views a YouTube video receives: length of the video (measured in minutes), number of likes the video received, whether the video includes kids, whether the video includes animals, what category the video falls under (the researcher examined three categories: humorous, informational, and inspirational), and, finally, the number of videos previously created by this person. Write the regression model to be estimated and clearly describe each variable in the model.

(Q.17)
An economist wishes to determine how a household's yearly spending on groceries (measured in hundreds of dollars) is influenced by yearly household income, household size, and the education level of the head of the household. Income is measured in thousands of dollars, Size is measured by number of household members, and Education is measured in years. The economist randomly selected 50 households and conducted a regression analysis on the collected data. The regression coefficients resulting from the regression analysis are shown below

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.16	0.58	-0.276	0.783814
Income	0.45	0.11	4.091	0.000160
Size	0.43	0.08	5.375	0.000002
Education	0.065	0.043	1.512	0.137052

1. Write the equation of the regression model.
2. Interpret the meaning of the coefficient of "Income" in context.
3. State the hypotheses corresponding to the P-value shown for Income.
4. Suppose we want to test whether the true slope of the Income variable is 0.2. Write the hypotheses, do the test, find the P-value, and write a conclusion.
5. Use a confidence interval to estimate the true slope of Income, and interpret its meaning in context.

(Q.18)
In class we looked at an ANOVA example about the number of job offers that a sample of students had at the time of graduation, together with their major field of study. Suppose a more complete analysis with a bigger sample gave the following summary statistics together with the ANOVA table below

major	n	mean	SD
DS	102	4.284	0.5
Biology	108	4.105	0.56
Economics	253	3.895	0.55

	Df	SS	MS	F	P-value
Factor		1.59			0.068
Residuals					
Total	462	136.66			

1. Fill in the blanks in the ANOVA table.
2. Suppose the mean value for DS was actually 4.14. How, if at all, would that change the F-ratio? Assume all other statistics remain the same.
3. Suppose we take the exact same data (same students, same job offers, etc.), but instead of grouping them by major, we group them by residency: in-state, out of state, international. We carry out a new ANOVA analysis and compare mean number of job offers by residency. Which of the following, if any, would remain the same: SS_{factor}, $SS_{residuals}$, SS_{total}? Give reasons.

(Q.19)
Suppose in the previous question we want to fit a multilinear regression model to predict job offers based on the 3 majors we considered. Write the general form of the model, and explain the meaning and possible values of the predictors.

(Q.20)
A recent medical study observed a higher frequency of heart attacks among a group of bald men than among another group of men who were not bald. Based on a P-value of 0.062 the researchers concluded there was some evidence that male baldness may be a risk factor for predicting heart attacks. Explain in this context what their P-value means.

(Q.21)
To plan the budget for next year a college needs to estimate what effect the current economic downturn might have on student requests for financial aid. Historically this college has provided aid to 35% of its students. Officials took a random sample of this year's applications to see what proportion indicate a need for financial aid. Based on these data they created a 90% confidence interval of (32%, 40%).

1. Interpret the meaning of the interval in this context. What should the college infer?
2. Suppose they decide to test the hypothesis H_0: $p=0.35$ versus H_A: $p\neq0.35$. What level of significance would match the confidence interval above and lead to the same inference?

(Q.22)
A few years back, a group of Earlham students taking a stats class collected data on smoking patterns among EC students. The table below shows frequency counts of smokers vs non-smokers in different demographic groups.

	Domestic	International
Smoker	25	8
Non-smoker	53	14

1. For each student group, find the odds ratio of smoking vs not smoking.
2. Construct a logistic regression model to predict the probability of smoking based on student group.
3. Interpret the meaning of the slope coefficient in your model in the context of odds ratios and student demographic.