Section 1

1. Use the following data from John Snow's investigation of the London Cholera Epidemic to fill in the X^2 (Chi-squared) test information in parts a through g:

Table 1: London Cholera Houses with Deaths

	Houses with death (C)	Houses without deaths (D)
Southwark & Vauxhall (A)	1262	38785
Lambeth (B)	97	26010

a. What are the row and column totals (marginal frequencies)?
A -
B -
C -
D -

b. What are the expected values?

AC -
AD -
BC -
BD -

c. What are the degrees of freedom?

d. What is the critical value for p < .05 and this degrees of freedom?
e. What is the chi-squared score?
f. Is the chi-square score greater than the critical value?
g. Do you accept/retain or reject the null hypothesis?

Section 2

2. For a paired t-test on the following data, fill in the questions:

H0: The means of the groups are equal.
Hl: The means of the groups are not equal.

Group A	Group B	Difference
47.68	48.89	
48.02	48.77	
51.11	49.76	
50.50	51.71	
48.69	49.44	
51.12	49.77	
50.24	51.45	
48.49	49.24	
48.4	47.05	

a What is the mean of the differences?

b What is the standard deviation of the differences?

c What is the sample size?

d How many degrees of freedom?

e What is the critical value for $p < .05$ and this many degrees of freedom?

f What is the t-score?

g Is the t-score greater than the critical value?

h Do you retain or reject the null hypothesis?

Section 3: Frequency Distributions and Hypothesis Testing

3 In hypothesis testing we generate a test score (t, z, F, chi-square, and others) and compare it to a critical value for the desired probability or alpha level. We reject the null hypothesis if the size of the test score is **greater/lesser** than the critical value. (Pick one)

4 What is the name of the hypothesis that the result is due to random chance or that there is no relationship between the variables?

5 What is the name for the hypothesis, also called the "research hypothesis," that matches the theory we are testing and suggests that there is a relationship between the variables?

6 Hypothesis test critical values come from probability tables based on these: _____

7 The z-score probability table is based on the _____ distribution.

8 The t-distribution is _____ in the middle and _____ in the tails than the z-distribution at small sample sizes.

9 A _____ sample t-test would be useful for comparing before and after results for the same 12 test subjects in an experiment.

10 The _____ test is used for categorical variables.

11 If the sample size is greater than 30 and the _____ standard deviation is known, we can use the z-score for hypothesis testing.

12 The _____ is used for ANOVA (Analysis of Variance) testing.

13 What is the probability level (alpha level or p-value) used in hypothesis testing in the social and life sciences?

Section 4: OLS Regression

14 The purpose of Ordinary Least Squares regression is to find a _____ , an equation that draws a _____ .

15 Ordinary Least Squares regression finds the line with the smallest possible error by minimizing the _____ of the vertical distance between fitted points and actual observations.

16 The vertical distances between fitted points on the regression line and the observed observations in the data are called _____ .

17 In the linear equation $y = \alpha + \beta X + \epsilon$, the α represents what on the line?

18 In the same equation, the ϵ, represents what?

19 What characteristic is true of the error term in OLS?

20 Which OLS assumption involves more than two X variables having linear relationships with each other and with Y?

21 The normality assumption is that Y is _____ _____ given X.

22 The homoskedasticity assumption is that _____ of the residuals is independent of the value of X.

Section 5: OLS Regression Results

You have the following results from a regression run in R for movies starring Tom Cruise in the leading role. Answer the questions:

Table 2

	Dependent variable:
	Domestic Box Office Sales ($)
Freshness Score (Rotten Tomatoes)	1,051,713.000** (458,727.900)
Constant	35,825,645.000 (31,931,528.000)
Observations	33
R^2	0.145
Adjusted R^2	0.117
Residual Std. Error	59,240,788.000 (df = 31)
F Statistic	5.256** (df = 1; 31)
Note:	*p<0.1; **p<0.05; ***p<0.01

You also have this from the regression summary:
Coefficients:
!!!!!!!!!!!!!!! Estimate !!!!! Std. Error ! t value ! Pr(>|t|)

(Intercept) 35825645 ! 31931528 ! 1.122 ! 0.2705
Freshness 1051713 ! 458728 ! 2.293 ! 0.0288 *

23 Write an equation of the form $y = \alpha + \beta X + \epsilon$ using the coefficients from the regression.

24 What is the Z-score for the variable "Freshness Score".

25 What is the t-value for the variable "Freshness Score".

26 What is the significance level for the variable "Freshness Score".

27 Do you accept or reject the null hypothesis for Freshness Score?

28 How much is one additional point of Freshness Score worth in Domestic Box Office Sales?

29 How would you describe the relationship of Freshness Score to Domestic Box Office Sale in one or two sentences?

30 Had you ever seen a Tom Cruise movie before this year?