Extra Practice Problems for the Final Exam

For Questions 1 to 2, refer to the data below. A consumer products company surveyed a large number of customers about their satisfaction with the price and quality of one of their products. Let P represent the price satisfaction; the company observed E[P]=7.6 and stdev(P)=1.8. Let Q represent the quality satisfaction; the company observed E[Q]=8.2 and stdev(Q)=1.4. It was also observed that Cor(P,Q)= −0.833.

1. Let quality satisfaction be more important so that total satisfaction is T = 2P + 3Q. What is the ~~variance~~ SD of T?

   a. 1.3
   b. 2.3
   c. 3.3
   d. 4.3

2. Which of the following is the covariance between P and Q?
   a. -2.1
   b. -4.2
   c. 2.1
   d. 4.2

   where Cov(P,Q) = SD(P)*SD(Q)*Cor(P,Q)

3. The company observed another 100 customers and found the sample mean total satisfaction of these customers. What is the distribution of the sample mean total satisfaction?

4. A sample of size n is drawn from some population, and the standard deviation (SE) of the sampling distribution of the mean is computed. A second, independent, sample of size 4n is drawn from the same population, and the new standard deviation (SE) of the sampling distribution of the mean is computed. What is the average ratio of the second SE to the first?
    a. 0.25
    b. 0.50
    c. 1.00
    d. 2.00

5. When constructing confidence intervals, as we increase the level of confidence (for example, from 90% to 95%), the width of the interval increases. Which of the following is the best explanation for this?
    a. We want to lower the proportion of intervals excluding the population parameter and hence must widen the interval.
    b. The sample size must decrease so we must widen the interval.
    c. We care less for precision and hence are willing to widen the interval.
    d. The degrees of freedom are decreasing and hence the interval gets wider.

For Questions 6 to 8, refer to the following data. A survey of 400 people living in Shenawee County was conducted and they were asked to report their annual household income. The mean was found to be $54,000.

6. A 99% CI was constructed for the population mean household income of Shenawee county based on the sample data. If the upper limit of the CI is $62,018, which of the following is nearest to the standard error of the mean?
   a. $2,673
   b. $3,114
   c. $4,091
   d. $4,874

   Assume sigma is known.

7. Assume that household incomes in Shenawee county are normally distributed. Also assume that the (sigma known) standard error of the sample mean is $1,000. What percentage of households is estimated to have income greater than $93,200?
   a. 1%
   b. 2.5%
   c. 5%
   d. 10%

   Assume now that population mean is =$54,000

8. Let μ represent the population mean household income of Shenawee county. In order to qualify for additional federal grant dollars, Shenawee county must present evidence that the mean household income is less than $60,000 dollars. Which of the following is the appropriate hypothesis test that they should do?
   a. $H_0 : \mu \leq \$60{,}000$ vs. $H_A : \mu > \$60{,}000$
   b. $H_0 : \mu = \$60{,}000$ vs. $H_A : \mu \neq \$60{,}000$
   c. $H_0 : \mu \geq \$60{,}000$ vs. $H_A : \mu < \$60{,}000$
   d. $H_0 : \mu = \$60{,}000$ vs. $H_A : \mu > \$60{,}000$

9. In a simulation experiment, the standard deviation of the profit was found to be 473.54 (million dollars), while the standard error was found to be 23.627 (also million dollars). How many trials were used?
    a. 200
    b. 400
    c. 800
    d. 1600

For Questions 10 to 11 below, refer to the following information.
Three stocks (labeled A, B, C) are being considered for a portfolio. Their pair-wise correlations are as follows: r (A,B) < 0, r(A,C) = 0 and r(B,C) > 0. The current prices of all three stocks are the same. Standard deviations of returns of stocks are the same: stdev[return$_A$]=stdev[return$_B$]=stdev[return$_C$]

10. Let a_A, a_B, and a_C be the fractions invested in each of the stocks. Which of the following portfolios has the highest variance?
    a. $a_A = 0.5$ and $a_B = 0.5$
    b. $a_A = 0.5$ and $a_C = 0.5$
    c. $a_B = 0.5$ and $a_C = 0.5$
    d. All three portfolios have the same variance

11. Suppose the expected returns of the stocks are as follows: E[B] < E[A] < E[C] Which of the following portfolios has the highest expected return?
    a. $a_A = 0.5$ and $a_B = 0.5$
    b. $a_A = 0.5$ and $a_C = 0.5$
    c. $a_B = 0.5$ and $a_C = 0.5$
    d. All three portfolios have the same expected return

12. If X and Y are independent and identically distributed, which of the following is equal to Var[2X -2Y]?
    a. 2Var[X]
    b. 4Var[X]
    c. 8Var[X]
    d. 16Var[X]

IFOCE (International Federation of Competitive Eating)
Inspired by the huge success of Nathan's annual international hot-dog eating contest, Jeff Jeffrey (a renowned food writer) was asked by the New York Times to write an article on competitive eating. He recently hired you as an intern to evaluate various beliefs and theories about the universe of competitive eating.

13. Jeffrey wants to estimate the average number of hot-dogs eaten by contestants in any hot-dog eating contest. To test this, you have randomly selected a sample of 66 contestants who participated in last year's contest. Your data yields a sample mean of 17.93 hot-dogs and a sample standard deviation of 10.59 hot-dogs. Construct a 90% confidence interval for the population mean number of hot-dogs eaten at the contest.

Jeffrey learned that the five-time Nathan's champion, Takeru Kobayashi, is touted as the Michael Jordan of competitive eating. In the last 5 years, he has won most of the hot-dog eating contests held around the world including Nathan's, and his record (55 hot-dogs in 12 minutes) is considered to be the monumental achievement in the world of competitive eating. Some people nicknamed him "125-point miracle man", and NHK (a Japanese TV station) even made a documentary on his achievements. Fascinated by Kobayashi's winning record, Jeffrey decides to find out more about him.

14. During an interview, Kobayashi claimed that he was able to eat at least 40 hot-dogs on average for the last 4 years. Jeffrey suspected that he was exaggerating his achievement as many top professional athletes do. To test this, Jeffrey selected a random sample of 40 events and looked up Kobayashi's records. The sample mean and sample standard deviation from the 40 events are 37 and 8.962, respectively. Is there strong evidence to suggest that Kobayashi exaggerated during his interview with Jeffrey? State the appropriate null and alternative hypotheses and test at $\alpha = 0.01$. Use the template to solve

The Wall Street Journal and Ross School of Business

A reporter in the Wall Street Journal, Jim Alright, is in charge of writing an article on the top 30 business schools in the US, including the Ross School of Business ranked #1 in the WSJ. The article intends to demonstrate how MBA degrees help MS (management science) professionals (in particular, Ross School BBA and MBA graduates) further their careers and increase their market value in today's business world. After interviewing 78 MS professionals with 10 or fewer years of experience after receiving an MBA degree, Jim decided to build a model explaining salary Y (in thousands of dollars) with two variables: years of experience (YR) and gender (I_F = 1 for female, 0 for male). What follows is the partially complete Excel output for the model estimated by Jim.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.75390
R Square	
Adjusted R Square	
Standard Error	20.96900
Observations	78

ANOVA

	df	*SS*	*MS*	*F*	p-value
Regression	2	43422.19	21711.09		0.00000
Residual	75	32977.58	439.70		
Total	77	76399.77			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	101.06900	5.69400	17.75009	0.00000
Yr	7.63300	0.84200		
I_F	-16.80000	4.75700	-3.53164	0.00070

15 . Write the estimated regression line and interpret the regression slopes.

16 . What proportion of variation in Y is explained by the model?

We did not cover how to calculate R-sq in Fall 2022. Ignore this question.

17. Is the model statistically significant? State the null and alternative hypotheses determine the value of the appropriate test statistic, and test using the 1% significance level.

We did not cover how to calculate F-stat in Fall 2022. Ignore this question.

18. Is there evidence suggesting that the average increase in salary per year of experience is greater than $6,500? State the null and alternative hypotheses, determine the value of the appropriate test statistic, and test using the 5% significance level. You can use the template to answer this.

Module 6 Practice

1. Following is based on American National Election Survey (ANES) data on whether survey takers voted in the 2000 presidential election or not. Many factors could influence the turn-out. In this survey, information on *age*, high school degree (*hsdeg*), college degree (*coldeg*), whether voted in 1996 election (*vote96*), and whether married (*married*) are recorded.

   When considering all variables as explanatory variables we have:

```
Logistic regression (GLM)
Data                 : Vote
Response variable    : vote00
Level                : 1 in vote00
Explanatory variables: age, hsdeg, coldeg, vote96, married
Null hyp.: there is no effect of x on vote00
Alt. hyp.: there is an effect of x on vote00

               OR coefficient std.error z.value p.value
 (Intercept)           -1.938     0.276  -7.028  < .001 ***
 age         1.011      0.011     0.004   2.782   0.005 **
 hsdeg|1     2.066      0.726     0.196   3.694  < .001 ***
 coldeg|1    2.169      0.774     0.141   5.499  < .001 ***
 vote96|1    7.148      1.967     0.130  15.147  < .001 ***
 married|1   0.912     -0.092     0.158  -0.584   0.559

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.203
Log-likelihood: -913.561, AIC: 1839.123, BIC: 1872.039
Chi-squared: 466.394 df(5), p.value < .001
Nr obs: 1,783
```

   a. Is there any need to change the model?

```
Logistic regression (GLM)
Data                 : Vote
Response variable    : vote00
Level                : 1 in vote00
Explanatory variables: age, hsdeg, coldeg, vote96
Null hyp.: there is no effect of x on vote00
Alt. hyp.: there is an effect of x on vote00

               OR coefficient std.error z.value p.value
 (Intercept)           -1.962     0.272  -7.205  < .001 ***
 age         1.010      0.010     0.004   2.756   0.006 **
 hsdeg|1     2.054      0.720     0.196   3.673  < .001 ***
 coldeg|1    2.166      0.773     0.141   5.489  < .001 ***
 vote96|1    7.095      1.959     0.129  15.176  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.203
Log-likelihood: -913.733, AIC: 1837.465, BIC: 1864.896
Chi-squared: 466.052 df(4), p.value < .001
Nr obs: 1,783
```

b. Using the model obtained, which of the following are accurate interpretations for the **estimated coefficient** of the explanatory variable *age?* Select all that apply.

   I. For a 1 year increase in age, the likelihood of voting in the 2000 presidential election decreases by 1%

   II. For a 1 year increase in age , the likelihood of voting in the 2000 presidential election increases by 1%

   III. As age increases, the likelihood of voting in the 2000 presidential election increases

   IV. As age increases, the likelihood of voting in the 2000 presidential election decreases

c. After switching the reference category of the variable *vote96* to 1 and rerunning the logistic regression, we obtain the following output where all variables are significant at 5% level. Which of the following is the most accurate interpretation of the odds ratio for the explanatory variable *vote96|0?*

I. Not voting in the 1996 election increases the odds of voting in the 2000 election by about 14%, keeping all other variables constant.
II. Not voting in 1996 election decreases the odds of voting in the 2000 election by about 14%, keeping all other variables constant.
III. Not voting in 1996 election decreases the odds of voting in the 2000 election by about 86%, keeping all other variables constant.
IV. All we can say is that not voting in the 1996 election decreases the likelihood of voting in the 2000 election but we cannot interpret the magnitude of the effect based on the corresponding odds ratio value.

```
Logistic regression (GLM)
Data                 : Vote
Response variable    : vote00
Level                : 1 in vote00
Explanatory variables: age, hsdeg, coldeg, vote96
Null hyp.: there is no effect of x on vote00
Alt. hyp.: there is an effect of x on vote00

               OR coefficient std.error z.value p.value
 (Intercept)           -0.003     0.304  -0.009   0.993
 age         1.010      0.010     0.004   2.756   0.006 **
 hsdeg|1     2.054      0.720     0.196   3.673  < .001 ***
 coldeg|1    2.166      0.773     0.141   5.489  < .001 ***
 vote96|0    0.141     -1.959     0.129 -15.176  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.203
Log-likelihood: -913.733, AIC: 1837.465, BIC: 1864.896
Chi-squared: 466.052 df(4), p.value < .001
Nr obs: 1,783
```