Instructions

Completion of this test only requires a pencil, and a calculator. Put away all other materials. Answer the questions in the space provided in the test. You will not receive credit for any work done in the scratch area. Budget your time appropriately. Remember to write down YOUR NAME and ID number. Good luck!

Multiple Choice Questions

Please write the answers in the template provide here:

<table>
<thead>
<tr>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
<th>9.</th>
<th>10.</th>
</tr>
</thead>
</table>

1) Consistency for the sample average $\bar{Y}$ can be defined as follows, with the exception of

a. $\bar{Y}$ is consistent for $\mu_Y$.

b. $\bar{Y}$ has the smallest variance of all estimators.

c. $\bar{Y} \xrightarrow{p} \mu_Y$.

d. the probability of $\bar{Y}$ being in the range $\mu_Y \pm c$ becomes arbitrarily close to one as $n$ increases for any constant $c > 0$.

Answer: b

2) The central limit theorem

a. states conditions under which a variable involving the sum of $Y_1, \ldots, Y_n$ i.i.d. variables becomes the standard normal distribution.

b. postulates that the sample mean $\bar{Y}$ is a consistent estimator of the population mean $\mu_Y$.

c. only holds in the presence of the law of large numbers.

d. states conditions under which a variable involving the sum of $Y_1, \ldots, Y_n$ i.i.d. variables becomes the Student $t$ distribution.

Answer: a

3) If variables with a multivariate normal distribution have covariances that equal zero, then
a. the correlation will most often be zero, but does not have to be.
b. the variables are independent.
c. you should use the $\chi^2$ distribution to calculate probabilities.
d. the marginal distribution of each of the variables is no longer normal.

Answer: b

4) An estimator $\hat{\mu}_Y$ of the population value $\mu_Y$ is unbiased if

a. $\hat{\mu}_Y = \mu_Y$.
b. $\bar{Y}$ has the smallest variance of all estimators.
c. $E(\bar{Y}) \rightarrow \mu_Y$.
d. $E(\hat{\mu}_Y) = \mu_Y$.

Answer: d

5) With i.i.d. sampling each of the following is true except

a. $E(\bar{Y}) = \mu_Y$.
b. $\text{var}(\bar{Y}) = \sigma^2_Y / n$.
c. $E(\bar{Y}) < E(Y)$.
d. $\bar{Y}$ is a random variable.

Answer: c

6) The $p$-value is defined as follows:

e. $p = 0.05$.
f. $\Pr_{H_0}(|\bar{Y} - \mu_{y,0}| > |\bar{Y}^{act} - \mu_{y,0}|)$.
g. $\Pr(z > 1.96)$.
h. $\Pr_{H_0}(|\bar{Y} - \mu_{y,0}| < |\bar{Y}^{act} - \mu_{y,0}|)$.

Answer: b

7) The critical value of a two-sided $t$-test computed from a large sample

e. is 1.64 if the significance level of the test is 5%.
f. cannot be calculated unless you know the degrees of freedom.
g. is 1.96 if the significance level of the test is 5%.
h. is the same as the $p$-value.
Answer: c

8) In the simple linear regression model, the regression slope

a. indicates by how many percent \( Y \) increases, given a one percent increase in \( X \).
b. when multiplied with the explanatory variable will give you the predicted \( Y \).
c. indicates by how many units \( Y \) increases, given a one unit increase in \( X \).
d. represents the elasticity of \( Y \) on \( X \).

Answer: c

9) The variance of \( Y_i \) is given by

a. \( \beta_0^2 + \beta_1^2 \text{var}(X_i) + \text{var}(u_i) \).
b. the variance of \( u_i \).
c. \( \beta_1^2 \text{var}(X_i) + \text{var}(u_i) \).
d. the variance of the residuals.

Answer: c

10) The OLS residuals

a. can be calculated using the errors from the regression function.
b. can be calculated by subtracting the fitted values from the actual values.
c. are unknown since we do not know the population regression function.
d. should not be used in practice since they indicate that your regression does not run through all your observations.

Answer: b

Analytical Questions: [40 points]

1. The following problem is frequently encountered in the case of a rare disease, say AIDS, when determining the probability of actually having the disease after testing positively for HIV. (This is often known as the accuracy of the test given that you have the disease.) Let us set up the problem as follows: \( Y = 0 \) if you tested negative using the ELISA test for HIV, \( Y = 1 \) if you tested positive; \( X = 1 \) if you have HIV, \( X = 0 \) if you do not have HIV. Assume that 0.1 percent of the population have HIV and that the accuracy of the test is 0.95 in both cases of (i) testing positive when you have HIV, and (ii) testing negative when you do not have HIV. (The actual ELISA test is actually 99.7 percent accurate when you have HIV, and 98.5 percent accurate when you do not have HIV.)
(a) Assuming arbitrarily a population of 10,000,000 people, use the accompanying table to first enter the column totals.

<table>
<thead>
<tr>
<th></th>
<th>Test Positive ((Y=1))</th>
<th>Test Negative ((Y=0))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV ((X=1))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No HIV ((X=0))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>10,000,000</strong></td>
</tr>
</tbody>
</table>

Answer:

<table>
<thead>
<tr>
<th></th>
<th>Test Positive ((Y=1))</th>
<th>Test Negative ((Y=0))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV ((X=1))</td>
<td></td>
<td></td>
<td><strong>10,000</strong></td>
</tr>
<tr>
<td>No HIV ((X=0))</td>
<td></td>
<td></td>
<td><strong>9,990,000</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>10,000,000</strong></td>
</tr>
</tbody>
</table>

(b) Use the conditional probabilities to fill in the joint absolute frequencies.

Answer:

<table>
<thead>
<tr>
<th></th>
<th>Test Positive ((Y=1))</th>
<th>Test Negative ((Y=0))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV ((X=1))</td>
<td>9,500</td>
<td>500</td>
<td><strong>10,000</strong></td>
</tr>
<tr>
<td>No HIV ((X=0))</td>
<td>499,500</td>
<td>9,490,500</td>
<td><strong>9,990,000</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>10,000,000</strong></td>
</tr>
</tbody>
</table>
(c) Fill in the marginal absolute frequencies for testing positive and negative. Determine the conditional probability of having HIV when you have tested positive. Explain this surprising result.

Answer:

<table>
<thead>
<tr>
<th>HIV (X=1)</th>
<th>Test Positive (Y=1)</th>
<th>Test Negative (Y=0)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9,500</td>
<td>500</td>
<td>10,000</td>
</tr>
<tr>
<td>No HIV (X=0)</td>
<td>499,500</td>
<td>9,490,500</td>
<td>9,990,000</td>
</tr>
<tr>
<td>Total</td>
<td>509,000</td>
<td>9,491,000</td>
<td>10,000,000</td>
</tr>
</tbody>
</table>

\[ \Pr(X=1|Y=1) = 0.0187 \] Although the test is quite accurate, there are very few people who have HIV (10,000), and many who do not have HIV (9,999,000). A small percentage of that large number (499,500/9,990,000) is large when compared to the higher percentage of the smaller number (9,500/10,000).

(d) The previous problem is an application of Bayes’ theorem, which converts \[ \Pr(Y = y | X = x) \] into \[ \Pr(X = x | Y = y) \]. Can you think of other examples where \[ \Pr(Y = y | X = x) \neq \Pr(X = x | Y = y) \]?

Answer: Answers will vary by student. Perhaps a nice illustration is the probability to be a male given that you play on the college/university men’s varsity team, versus the probability to play on the college/university men’s varsity team given that you are a male student.

2. You have obtained a sub-sample of 1744 individuals from the Current Population Survey (CPS) and are interested in the relationship between weekly earnings and age. The regression, using heteroskedasticity-robust standard errors, yielded the following result:

\[ \hat{Earn} = 239.16 + 5.20 \times Age, \quad R^2 = 0.05, \quad SER = 287.21, \]

\[ (20.24) \quad (0.57) \]

where \( Earn \) and \( Age \) are measured in dollars and years respectively.

(a) Interpret the results.

Answer: A person who is one year older increases her weekly earnings by $5.20. There is no meaning attached to the intercept. The regression explains 5 percent of the variation in earnings.

(b) Is the relationship between \( Age \) and \( Earn \) statistically significant? Is the effect of
age on earnings large?

Answer: The t-statistic on the slope is 9.12, which is above the critical value from the standard normal distribution for any reasonable level of significance. Assuming that people worked 52 weeks a year, the effect of being one year older translates into an additional $270.40 a year. This does not seem particularly large in 2002 dollars, but may have been earlier.

(c) Why should age matter in the determination of earnings? Do the results suggest that there is a guarantee for earnings to rise for everyone as they become older? Do you think that the relationship between age and earnings is linear?

Answer: In general, age-earnings profiles take on an inverted U-shape. Hence it is not linear and the linear approximation may not be good at all. Age may be a proxy for “experience,” which in itself can approximate “on the job training.” Hence the positive effect between age and earnings. The results do not suggest that there is a guarantee for earnings to rise for everyone as they become older since the regression $R^2$ does not equal 1. Instead the result holds “on average.”

(d) The average age in this sample is 37.5 years. What is annual income in the sample?

Answer: Since $\bar{Y} - \bar{X} = \bar{Y} = \bar{\beta}_0 + \bar{\beta}_1 \bar{X}$. Substituting the estimates for the slope and the intercept then results in average weekly earnings of $\text{S}434.16 or annual average earnings of $22,576.32.

**Empirical Question: [40 points]**

1. The questions for this problem are based on the following EViews output:

<table>
<thead>
<tr>
<th>Correlations:</th>
<th>SALARY</th>
<th>EDUCATION</th>
<th>EXPERIENCE</th>
<th>SENIORITY</th>
<th>GENDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALARY</td>
<td>1.000000</td>
<td>0.411985</td>
<td>0.166748</td>
<td>0.285844</td>
<td>0.474565</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>0.411985</td>
<td>1.000000</td>
<td>-0.101161</td>
<td>-0.059844</td>
<td>0.296863</td>
</tr>
<tr>
<td>EXPERIENCE</td>
<td>0.166748</td>
<td>-0.101161</td>
<td>1.000000</td>
<td>0.074666</td>
<td>0.154714</td>
</tr>
<tr>
<td>SENIORITY</td>
<td>0.285844</td>
<td>-0.059844</td>
<td>0.074666</td>
<td>1.000000</td>
<td>0.088645</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.474565</td>
<td>0.296863</td>
<td>0.154714</td>
<td>0.088645</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Statistics:</th>
<th>SALARY</th>
<th>EDUCATION</th>
<th>EXPERIENCE</th>
<th>SENIORITY</th>
<th>GENDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5420.323</td>
<td>12.50538</td>
<td>100.9269</td>
<td>16.72043</td>
<td>0.634409</td>
</tr>
</tbody>
</table>
Median  5400.000  12.00000  70.00000  15.00000  1.000000  
Maximum  8100.000  16.00000  381.0000  34.00000  1.000000  
Minimum  3900.000  8.000000  0.000000  1.000000  0.000000  
Std. Dev.  709.5872  2.282369  90.94755  10.25476  0.484206  
Skewness  0.589159 -0.486804  1.146060  0.191436 -0.558180  
Kurtosis  4.172759  2.684228  3.628779  1.881037  1.311565  
Jarque-Bera  10.70971  4.059541  21.89057  5.419842  15.87616  
Probability  0.004725  0.131366  0.000018  0.066542  0.000357  
Sum  504090.0  1163.000  9386.200  1555.000  59.00000  
Sum Sq. Dev.  46323290  479.2473  760974.0  9674.731  21.56989  
Observations  93  93  93  93  93  

**Hypothesis test:**

Hypothesis Testing for SALARY
Sample(adjusted): 1 93
Included observations: 93 after adjusting endpoints
Test of Hypothesis: Mean = 5275.000

<table>
<thead>
<tr>
<th>Method</th>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>1.975008</td>
<td>0.0513</td>
</tr>
</tbody>
</table>

(a) If the variable GENDER takes the value of 1 for males and 0 for females, what proportion of this sample is made of male workers? 0.634409

(b) If the variable EDUCATION were normally distributed, give a 95% confidence interval given the sample statistics reported

12.51 ± 1.96 × 2.28

(c) Given the data in the hypothesis test, give a 95% confidence interval for the mean SALARY level.

\[
SE = \frac{709.5872}{\sqrt{93}} = 73.58
\]

5420.323 ± 1.96 × 73.58

(d) Given the data in the hypothesis test, would you fail to reject/reject the null hypothesis? Explain.

The p-value is greater than 0.05, hence we fail to reject the null hypothesis.
(e) Would you reject the null hypothesis at a 90% confidence level? Simple yes or no: ___YES___.

(f) Given the correlations between the following pairs of variables only (that is, ignoring the remaining correlations reported): SALARY and GENDER; EXPERIENCE and EDUCATION; and EDUCATION and GENDER; discuss whether the data suggest or not that female workers are discriminated.

The correlation between SALARY and GENDER is positive, the correlation between EXPERIENCE and EDUCATION is negative, and the correlation between EDUCATION and GENDER is positive. Hence, the data does not suggest that female workers are discriminated. It is reasonable to assume that male workers are paid higher wages because they have higher education level.