Your Name: ___________________________________________

Your Section:

<table>
<thead>
<tr>
<th>Section</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhaonan (7:30)</td>
<td>Zhaonan (8:30)</td>
</tr>
<tr>
<td>Mike (11:30)</td>
<td>Chris (12:30)</td>
</tr>
<tr>
<td>Lin-Yang (2:30)</td>
<td>Yong (3:30)</td>
</tr>
<tr>
<td>Mike (11:30)</td>
<td>Jeff (9:30)</td>
</tr>
<tr>
<td>Yong (3:30)</td>
<td>Yen-Ning (1:30)</td>
</tr>
<tr>
<td>Yen-Ning (2:30)</td>
<td>Jeremy (3:30)</td>
</tr>
<tr>
<td>Lin-Yang (2:30)</td>
<td>Jeremy (4:30)</td>
</tr>
</tbody>
</table>

Note:
- Show your work on all questions. Unsupported work will not receive full credit.
- All answers should be in decimal form and should be exact, or to at least taken out to two decimal places.
- You are responsible for upholding the Honor Code of Purdue University. This includes protecting your work from other students.
- You are allowed the following aids: a one-page 8 ½” x 11” handwritten (in your handwriting) cheat sheet, a scientific calculator, and pencils.
- Instructors will not interpret questions for you. If you do have questions, wait until you have looked over the whole exam so that you can ask all of your questions at one time.
- You must show your student ID (upon request) and turn in your cheat sheet when you turn in your exam to your instructor.
- Turn off your cell phone before the exam begins!

<table>
<thead>
<tr>
<th>Question</th>
<th>Points Possible</th>
<th>Points Received</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Cheat Sheet</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
1. Choose the phrase from the following list that most accurately completes the sentences below. Each phrase should be used no more than once. (2 points each)

(A) Experiment  (G) Time Series  (M) K-th Random Sample
(B) Observational Study  (H) Cross-sectional  (N) Clustered Sample
(C) With Replacement  (I) Stem and Leaf Plot  (O) Stratified Sample
(D) Without Replacement  (J) Box plot  (P) Outlier(s)
(E) Intersection  (K) Scatterplot  (Q) Variance
(F) Union  (L) Crosstabulation  (R) Covariance

a. Aaron is trying to create a graphical representation of his data on people’s height. He wants to be able to see the shape of a distribution while still keeping all the values of the data, his friend from STAT225 suggests to him that a(n) ___I___ should do the job.

b. Becky wants to monitor her own weight for the next two months. She chooses to measure her weight once a day and the resulting record is a set of ____G_______ data.

c. Among all the different ways to measure the variability of a single variable Charley has learned in STAT 225, ONLY ____Q_______ does NOT use the same unit as the variable itself.

d. Dena is working on a data set on people’s health. She suspects that there is a linear relationship between people’s age and their chance of getting cardiovascular disease. After consulting with her STAT225 instructor, she makes a(n) ___K_____ to check her guess.

e. The staff members at School of Management want to know STAT225 student’s expected grade. They divide all STAT225 students into two groups, U.S. students and international students, and then randomly ask 100 students from each group for their expected grade. Their sampling scheme is ___O______.

f. Gerald wants to buy a new car. He went to a car dealership and test-drove 6 different cars on the dealer’s parking lot. He is sampling ____D______.

g. A pharmaceutical company wanted to evaluate the effectiveness of a new drug to lower people’s blood pressure. They gave their drug to 120 patients and recorded their temperature, pulse and blood pressure during the next 3 months. The company did a(n) _____A________ to collect data.

h. Heather wants to know the plans for spring break among her stat225 classmates. The three options are: A. go home; B. travel to some other places or C. stay on campus. From what she has learned, 300 of her classmates will either go home or travel, 200 of them will stay on campus. The number 300 represents a(n)_______F_______ of A and B.
2. A sample of 353 equities traded on the New York Stock Exchange was sorted according to which Industry the company corresponds to and the interest rates (Yield) they currently paid on their corporate debt bonds. A summary is given below in the cross-tabulation table: (3 points each)

<table>
<thead>
<tr>
<th>Yield \ Industry</th>
<th>High Tech</th>
<th>Energy</th>
<th>Manufacturing</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>65</td>
<td>45</td>
<td>25</td>
<td>135</td>
</tr>
<tr>
<td>Mid</td>
<td>47</td>
<td>A</td>
<td>34</td>
<td>B</td>
</tr>
<tr>
<td>Low</td>
<td>25</td>
<td>50</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Column Total</td>
<td>137</td>
<td>C</td>
<td>102</td>
<td>353</td>
</tr>
</tbody>
</table>

Answer the following questions making sure to show all of your work:

a. Find A, B, and C.

C=353-137-102=114
A=114-45-50=19
B=47+19+34=100

b. Compute the probability that a randomly chosen company is from energy industry and currently pays high yield on their bonds. Is it joint, marginal or conditional probability?

45/353=0.1275, Joint

c. Compute the probability that a randomly chosen company is from manufacturing industry. Is it joint, marginal or conditional probability?

102/353=0.2890, marginal

d. If we know that a company is from high tech industry, what is the probability that this company pays high yield on their bonds? Is it joint, marginal or conditional probability?

65/137=0.4745, conditional
3. A manager at a UPS branch collected information from 50 drivers on the time it takes them to make the first delivery (in minutes). The data are sorted in descending order as follows:

<table>
<thead>
<tr>
<th>102</th>
<th>72</th>
<th>60</th>
<th>56</th>
<th>50</th>
<th>41</th>
<th>37</th>
<th>35</th>
<th>34</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>70</td>
<td>58</td>
<td>56</td>
<td>44</td>
<td>39</td>
<td>36</td>
<td>35</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>86</td>
<td>67</td>
<td>58</td>
<td>55</td>
<td>43</td>
<td>39</td>
<td>36</td>
<td>35</td>
<td>32</td>
<td>27</td>
</tr>
<tr>
<td>74</td>
<td>65</td>
<td>58</td>
<td>51</td>
<td>41</td>
<td>38</td>
<td>36</td>
<td>35</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>73</td>
<td>61</td>
<td>57</td>
<td>50</td>
<td>41</td>
<td>37</td>
<td>36</td>
<td>35</td>
<td>29</td>
<td>27</td>
</tr>
</tbody>
</table>

a. Find the 5-number summary of the delivery time. (5 points)

Min: 27; Q1: 35 (13th); Q2: 41 (average of 25th and 26th); Q3: 58 (38th); max: 102

b. Find the range and the Inter-Quartile Range (IQR) of the delivery time. (2 points)

Range=102-27=75;
IQR: 58-35=23

c. It is the company’s policy that the manager has to investigate the reason for any delivery times that are considered outliers. There will be a cost of $300 to initiate the investigation plus additional $150 for each outlying case. How much will it cost the company to conduct the investigation at this branch? (3 points)

LL=Q1-1.5*IQR=35-1.5*23=0.5
UL=Q3+1.5*IQR=58+1.5*23=92.5

There are 2 outliers (95 and 102). So the cost is: 300+2*150=$600.
4. Use the data below to answer the following questions

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-6</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

a. Calculate sample standard deviation of Y ($s_Y$). (3 points)

Y-bar=-1.

$$S_Y^2=[(-6+1)^2+(-2+1)^2+(1+1)^2+(3+1)^2]/3=15.33$$

$$S_Y=\sqrt{15.33}=3.92$$

b. Given:

(i). the sample correlation between X and Y ($r_{XY}$) = 0.9791, and

(ii). the sample covariance between X and Y ($s_{XY}$) = 7, find the sample standard deviation of X ($s_X$). (3 points)

$$r_{XY}=S_{XY}/S_XS_Y,$$ so $S_X=S_{XY}/r_{XY}S_Y=7/(0.9791*3.92)=1.82$

c. Given:

(i) the sample standard deviation of Z ($s_Z$) = 2.1602,

(ii) the sample covariance of X and Z ($s_{XZ}$) = -1

(iii) the sample covariance of Y and Z ($s_{YZ}$) = -0.6667

Find out which set of variables (XY, XZ, or YZ) has the weakest correlation and justify your answer? (5 points)

$$r_{XY}=0.9791$$
$$r_{XZ}=S_{XZ}/S_XS_Z=-1/(1.82*2.1602)=-0.25$$
$$r_{YZ}=S_{YZ}/S_YS_Z=-0.6667/(3.92*2.1602)=-0.0787$$

So YZ has the weakest correlation.
5. There is a group of 15 employees at a new restaurant. (3 points each)

a. Suppose 10 of them are to be assigned as servers, 3 are to be assigned as hosts, and 2 are to be assigned as cashiers. In how many ways can the assignment be made?

\[ C_{15}^{10} C_5^3 C_2^2 = 30030 \]

b. How many possible ways are there to line up these 15 employees?

\[ 15! = 1307674368000 \]

c. What’s the probability that the 3 hosts will stand together in a lineup?

\[ \frac{13!3!}{15!} = 0.0286 \]

d. If we randomly select 3 employees from 15, what’s the probability that no cashiers are being selected?

\[ \frac{C_{13}^3}{C_{15}^3} = 0.6286 \]
6. The Lafayette High School 2011 Senior Class consists of 120 students. 110 of them are studying at least one of the three following foreign languages: Portuguese, Spanish, and French. In particular, the course enrollments are as following: 60 Portuguese, 55 Spanish, 65 French, 30 Spanish and French, and 35 French and Portuguese. You also know that there are 40 students who are taking Spanish but not Portuguese.

Compute the following probabilities making sure to show all of your work:

a. Fill in all the missing numbers in the Venn diagram. (7 points)

b. Find the probability that a student enrolled in all three language courses. (2 points)

\[
\frac{10}{120} = 0.0833
\]

c. Given that a student enrolled in Spanish, what is the probability that he/she also enrolled in French? (2 points)

\[
P(F|S) = \frac{P(F \text{ and } S)}{P(S)} = \frac{30}{55} = 0.5455
\]

d. The students in the Spanish class want to form a club, however they need to elect a President, Vice President, and Treasurer. If officers must be selected from among those students who ONLY enrolled in Spanish class, how many different officer triplets are obtainable? (2 points)

\[
P_{20}^3 = 6840
\]
7. Macro Economic Theory via Milton Friedman suggests MV = PQ, where M represents the Money Supply, V represents the Velocity of Money, P represents the Price Level, and Q represents the Real Output. An economist is curious about the relationship between the change of P (\(\Delta P\)) and the change of M (\(\Delta M\)), assuming V and the change of Q (\(\Delta Q\)) constant. She collected the following data in France for the years 1980-1990 and estimated the trend line as:

\[
\Delta P = 0.9712(\Delta M) - 0.8641 \text{ with a } R^2 = 0.6431.
\]

a. Interpret the slope in the context of this problem. (2 points)

For every unit increase in the change in money growth, on average, the change in price level will increase by 0.9712 unit.

b. It is known that change in money supply in France was 8% for the year 1991 and 25% for the year 1996. What would you predict to be the change in price level in these two years? Which of these two predictions is more reliable? (3 points)

1991: \(0.9712 \times 8 - 0.8641 = 6.9055\%\); 1996: \(0.9712 \times 25 - 0.8641 = 23.4159\%\)

The prediction for 1991 is more reliable since its \(\Delta M\) is within the range of change in money growth in the model while the \(\Delta M\) is out of range for 1996.

c. What is the Pearson correlation coefficient between change in money supply and change in price level? Is it a strong, moderate, or weak correlation? (3 points)

\[r = \sqrt{0.6431} = 0.8019, \text{ strong since it is greater than 0.8}\]

d. What does \(R^2\) mean in the context of this problem? (2 points)

Among all the variation observed in \(\Delta P\), about 64.31% is explained by having \(\Delta M\) in the trendline.
8. The grand prize at a certain carnival game is a large stuffed animal of some kind pulled out randomly from a bin without replacement. In the bin there are 6 giraffes, 4 elephants, 5 panda bears, 3 toucans, and 8 lions. Find the probability of the following situations. (3 points each)

a. The first two stuffed animals won are panda bears. (3 points)
\[ \frac{5}{26} \times \frac{4}{25} = 0.0308 \]

b. The second stuffed animal drawn is a giraffe knowing the first stuffed animal drawn was an elephant. (3 points)
\[ \frac{6}{25} = 0.24 \]

c. Of the first three animals won, at least one is a lion and at least one is a panda bear (4 points)
Let \( L = \{ \text{at least one lion} \} \) and \( P = \{ \text{at least one panda bear} \} \)
\[ P(L \text{ and } P) = P(L) + P(P) - P(L \text{ or } P) = 1 - P(L^C) + 1 - P(P^C) - [1 - P(L^C \text{ and } P^C)] \\
= 1 - \frac{C_{18}}{C_{26}} + 1 - \frac{C_{21}}{C_{26}} - 1 + \frac{C_{13}}{C_{26}} = 0.2846 \]

9. Answer the following questions:

a. A data analyst is searching for outliers in his data. He calculated the 5-number summary and found the upper limit (UL) and lower limit (LL). From his calculation, UL = 38 and LL = -10. What would be the inter-quartile-range (IQR) of the data? (3 points)
\[ \text{LL} = Q1 - 1.5 \times \text{IQR}; \quad \text{UL} = Q3 + 1.5 \times \text{IQR}; \quad \text{therefore, } \text{UL} - \text{LL} = Q3 - Q1 + 3 \times \text{IQR} = 4 \times \text{IQR}; \]
\[ \text{Since } \text{UL} - \text{LL} = 48, \quad \text{IQR} = 12 \]

b. The data analyst also found that the LL is much larger than a few actual data values and the UL is larger than all the values. What can he say about the shape of the distribution for his data, is it symmetric or skewed? If it is skewed, is it skewed to the left or to the right? (2 points)
\[ \text{It is skewed to the left since we have a few low outliers but no high outliers.} \]