Each Problem is worth 20 points. Answer five of the six problems. Problems should be answered as thoroughly as you are able. Partial credit on problems is only possible if I can locate errors in your calculations; neatness and organization of your answers is essential. Remember, answer only five of the six problems, if you chose to answer all, I will select the five with the lowest scores.

1. Employee morale is affected by working conditions and amenities at the work site. Although it is possible to overdo the amenities, better conditions – lunchrooms that are clean and have enough space for the lunch rush, sufficient numbers of clean bathrooms, good ventilation and temperature control – can be shown to improve employees views of work and employee performance.

Although it seems silly, bathrooms play a critical role in employee attitudes and the efficiency of a workplace. If there is too little bathroom space (aka, number of toilets), people spend considerable time waiting for a toilet. They are typically not happy about this and work time is lost while multiple employees wait. On the other hand, bathrooms are expensive both in terms of direct cost, floor space and cleaning time (the only thing worse than waiting for a toilet is waiting for a dirty toilet).

We are planning a new work site with about 100 employees. The planners for the site have asked about bathroom size and the number of toilets. Our records indicate that, on average, 3 toilets are in use at any one time (don’t ask how we know this).

Given the trade off between the number of toilets and the costs of bathroom facilities, we would like to have enough toilets so that employees will only have to wait for a stall 10% of the time.

A. What type of probability distribution should we use to calculate the probabilities for the number of toilets in use? Explain why this is the correct distribution.

Poisson distribution, which gives us the probability of a number of events occurring in a fixed time period.
B. Calculate the probabilities that 1, 2, 3, 4, 5 and 6 toilets are in use

<table>
<thead>
<tr>
<th>Number of Calls (aka x)</th>
<th>mean(x)^x</th>
<th>x!</th>
<th>P(x)</th>
<th>Cumulative Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.04979</td>
<td>0.04979</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0.14936</td>
<td>0.19915</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>2</td>
<td>0.22404</td>
<td>0.42319</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>6</td>
<td>0.22404</td>
<td>0.64723</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>24</td>
<td>0.16803</td>
<td>0.81526</td>
</tr>
<tr>
<td>5</td>
<td>243</td>
<td>120</td>
<td>0.10082</td>
<td>0.91608</td>
</tr>
<tr>
<td>6</td>
<td>729</td>
<td>720</td>
<td>0.05041</td>
<td>0.96649</td>
</tr>
</tbody>
</table>

C. How many toilets do we need to be sure that employees will have to wait for a toilet only 20\% of the time? How many do we need to be sure that employees have to wait only 10\% of the time.

We need to use the cumulative probabilities for this. The likelihood that 4 or fewer toilets will be in use at any time is 86\%, so that meets our 81.5\% goal (20\% of the time they have to wait). The likelihood that 5 or fewer stalls will be in use if 91.6\%, so with 6 toilets we should hit our 10\% goal.
2. A global organization with its headquarters in Germany absorbed an American firm several years ago. There is concern among those working for the American subsidiary that they are less likely to be promoted into management (even of the subsidiary) than those working for the parent company. You have been asked to analyze data on promotions to determine if those in the American subsidiary are, statistically speaking, disadvantaged with respect to promotion into management.

You know that 40% of employment in the global organization is in the American subsidiary. Considering the last 15 promotions, you find that 3 went to Americans and the remainder went to Germans. What is the likelihood that this outcome was the product of a process that was blind to nationality?

A. What type of distribution should you use to calculate this probability? Why?

   Binomial

B. What is the probability that we would have gotten the observed outcome (3 out of 15 Americans) if the selection was blind to nationality. What do you conclude from this test? Why?

\[
p(3 \text{ of } 15) = \frac{15!}{3!*12!} \times .4^3 \times .6^{12} = 6.34\%
\]
3. Extended overtime affects worker efficiency. Since Captain Jones research on the productivity differences on 8 and 12 hours shifts at Carnegie Steel in the 1880s, hundreds of studies have found that, absent particular circumstances, working beyond 40 per week results in lower productivity per hour or per worker.

We have data on hours of work and the value of output in the last hour of work for mechanical contractor (aka, plumbers and Pipefitters) across firms that work different schedules.

<table>
<thead>
<tr>
<th>Firm</th>
<th>Average Hours of Work</th>
<th>Value of Hourly Output per Employee for the Last Hour of Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>25</td>
</tr>
</tbody>
</table>

A. Calculate the covariance between average weekly hours of work and the value of output. Is the estimate consistent with the view that working longer hours causes marginal (last hours) output to decline? Explain.
B. Calculate the correlation between the average hours and the value of hourly output. How strong is the relationship between hours and value of output? What does correlation tell you about this relationship that you could not learn from the covariance measure?

Covariance and Correlation When You have 'raw' data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>(X - XBar)</th>
<th>(X -Xbar)^2</th>
<th>(Y -Ybar )</th>
<th>(Y - YBAR)^2</th>
<th>(Y - (xbar)*(ybar))</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>24</td>
<td>-6</td>
<td>36</td>
<td>7</td>
<td>49</td>
<td>-42</td>
</tr>
<tr>
<td>40</td>
<td>19</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>-2</td>
</tr>
<tr>
<td>43</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>-7</td>
<td>49</td>
<td>-14</td>
</tr>
<tr>
<td>47</td>
<td>7</td>
<td>6</td>
<td>36</td>
<td>-10</td>
<td>100</td>
<td>-60</td>
</tr>
<tr>
<td>40</td>
<td>25</td>
<td>-1</td>
<td>1</td>
<td>8</td>
<td>64</td>
<td>-8</td>
</tr>
</tbody>
</table>

| sum = 205 | 85 | sum = 78 | sum = 266 | sum = -126 |
| avg= 41    | 17 | var(x)= 19.5 | var(y)= 66.5 | COV(X,Y)= -31.5 |

std dev(x) 4.41588 std dev(y) 8.15475 Corr(X,Y)= -0.87475
4. Diversity is important to our firm and hiring a workforce that “looks like Michigan” (we won’t even begin to fully explore the implications of the slogan), is the byline of our recruiting effort. One dimension of diversity is a more balanced representation of the colleges in Michigan. Because the firm is located in Lansing, we have traditionally favored hiring graduates of the MSU and have tended to stay away from graduates of the UofM. No other school in Michigan is worth considering, and so we don’t. The schools each graduate similar numbers of undergraduates and graduate students, so the composition of the Michigan workforce with college degrees is evenly split between MSU and MU graduates. The mean for the percentage of college graduates who are from MSU for the Michigan workforce is .5 with standard deviation of .25.

Despite our stated goal of encouraging diversity in college background, the old school tie is strong among our current workforce and we are concerned that they may be continuing to favor MSU grads. We look at our records and find that 60 percent (.6) of our 25 college degree hires last year were MSU grads.

A. We wish to conduct a statistical test to determine whether we continue to hire excess numbers of individuals who have graduated from MSU. Set up the null and alternative hypotheses to test this.

\[ H_0: \mu \leq .5 \]
\[ H_A: \mu > .5 \]

B. Set up the hypothesis test and calculate the value of the t statistic.

\[ p(x > .5) = \frac{.6 - .5}{.25 / 5} = 2 \]

C. What are the critical values for a 10%, 5% and 1% test. Is it possible to reject the null in a 10% test, a 5% test, a 1% test.

We are working with a sample of 25, so we have 24 degrees of freedom. Further, as indicated in (A), this is a one tailed test. So our critical values from the t-table are:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>1.32</th>
<th>reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td>1.71</td>
<td>reject null</td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td>2.50</td>
<td>do not reject</td>
</tr>
</tbody>
</table>
5. The work of professionals has been changed by IT. The ready availability of IT and good software has permitted professionals to do many of their tasks more efficiently. For example, whereas professional once wrote memos out by hand, they now tend to compose memos directly. Similarly, they no longer use calculators to solve problems, rather they use spreadsheets and other computational software. This type of change should then result in professionals becoming more productive.

Other forces may well have worked against increased productivity. For example, whereas once professionals handed their hand written drafts to secretaries to type, they are now expected to do the typing themselves. Similarly, there used to be statistical clericals who collected the data and did calculations for professionals, such positions have now been, for the most part, eliminated. Even worse, ease of redrafting on word processors has encouraged professionals and their managers to iterate written work through more drafts than they did when typing and printing were more involved.

We are interested in whether the time spent writing memos has changed because of the introduction of IT. IBM collected data thirty years ago which indicated that the typical two page memo required 5 hours of professional’s time with a standard deviation of 1.25 hours. We have asked 28 of the professionals on our staff to track the time spent writing memos over a one month period. We find that they spend an average of 5.38 hours on a two page memo.

A. What is the null and alternative hypothesis for this study?

\[ H_0: \mu = 5 \]
\[ H_A: \mu \neq 5 \]

B. Test to determine if it is possible to reject the null in a 10%, 5% or 1% test?

\[ p(x = 5) = p(t(28 - 1) = \frac{5.38 - 5.0}{1.25 / \sqrt{28}}) = p(t(27) = 1.60 \]

Because this is a two tailed test, our critical points are 1.70 (10%) and 2.05 (5%) and 2.77 (1%). We cannot reject the null that IT has not altered the time taken to write a two page memo.
6. We are interested in the relationship between having children and FMLA use. To make things easier, we divide the workforce up into two groups: Child Status 1= those that have never had children, 2= those that have children. Additionally, we divide FMLA use into two groups: 1= those that have never used FMLA time in the past, 2= those that have used FMLA time in the past. Using the PeopleSoft™ program, we survey our employee database and find the following:

<table>
<thead>
<tr>
<th>Child Status</th>
<th>FMLA Use</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.3</td>
<td>.4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.1</td>
<td>.2</td>
<td></td>
</tr>
</tbody>
</table>

A. What is the probability that Child Status =1 and FMLA Use = 1, and what does it mean? For Child Status = 2 and FMLA Use =2? Confirm that the sum of all the probabilities is 1, and explain why this is important.

The probability that CS =1 and FMLA Use =1 is .3 or 30%. This indicates that 30% of our workforce has never had children and has never used FMLA time.

The probability that CMS =2 and FMLA Use = 2 is .2 or 20%. This indicates that 20% of our workforce have children, and has used FMLA time in the past.

Summing all the probabilities, we see that .3 + .4 + .1 + .2 = 1. This is important because is confirms that all of our employees are accounted for, and thus all possible outcomes are accounted for.
B. What is the difference between a marginal probability and a conditional probability?

A **marginal probability** is the probability that something will occur regardless, or independently, of other events. For example, the probability that someone has a child from the table above is .6, meaning that 60% of our employees have children.

A **conditional probability** is the probability something will occur, conditional on another event occurring. For example, a conditional probability from above would be the probability a person used FMLA time in the past, given (or on the condition) that they have children.

The difference between the two is that one is conditional (or dependent) on another event occurring, whereas marginal probability is unconditional. Conditional probabilities sample from a smaller portion of the population than marginal probabilities due, given the condition placed on them.

C. Calculate the probability that FMLA Use =1 given that Child Status =2. What does this probability mean?

$$\text{Prob}(\text{FMLA Use} = 1 \text{ given that } \text{CS} = 2) = \frac{.4}{.6} = .667 \text{ or 66.7%}.$$  

This means that of our employees who have children, 66.7% of them have never used FMLA time in the past.
D. Calculate the probability that Child Status = 2 given that FMLA Use = 1. What does this probability mean? Why is it different from the probability above?

\[ \text{Prob}(\text{CS} = 2 \text{ given that FMLA Use} = 1) = \frac{.4}{.7} = .571 \text{ or } 57.1\%. \]

This means that of the people who have never used FMLA in the past, 57.1% of them have children.

The two probabilities are different from one another because they are dependent on different conditions, and are thus drawing from different subsets of the population. In part C, we were drawing from the subset of the population who has children. In part D, we were drawing from the subset of the population who had never used FMLA time in the past.
Extra Credit (15 points possible)

1. (5 points)

We observe, in the population as a whole, two types of people: those who own expensive sports cars, and those who do not. In a simple random sample of 12 persons owning sports cars, we find a sample mean pulse rate of 75 (beats per minute) with sample standard deviation of 10 (beats per minute). In a simple random sample of 10 persons who do not own such sports cars, we find a sample mean pulse of 65 with sample standard deviation of 8. Do those owning sports cars have a higher pulse on average than those who do not?

The issue is whether the heart rates of the population of sports car owners and the population of non sports car owners differ. We have two samples, one from each group. So,

\[ H_0: \mu_{\text{sports}} \leq \mu_{\text{non-sports}} \]

\[ H_A: \mu_{\text{sports}} > \mu_{\text{non-sports}} \]

Our formula for this problem is:

\[
pr\left( \frac{\bar{x}_{\text{sports}} - \bar{x}_{\text{non-sports}}}{\sqrt{\frac{s_{\text{sports}}}{n_{\text{sports}}} + \frac{s_{\text{non-sports}}}{n_{\text{non-sports}}}}} \geq \frac{75 - 65}{\sqrt{\frac{10^2}{12 - 1} + \frac{8^2}{10 - 1}}} \right) = pr(t(20) \geq \frac{10}{4.025}) = pr(t(20) \geq 2.4843)
\]

our critical value for a t(20) is 10% (1.32), 5% (1.72) and 1% (2.53). We can reject the null of no difference in a 1% test.
2. (10 points)

Eight months ago, the Lancet, the British counterpart of the Journal of the American Medical Association (one of the three top general medical journals in the United States), printed an article about estimating civilian deaths in Iraq since the end of the invasion. The estimates were based on a random survey of the Iraqi population in which sites for the survey were chosen at random and then a protocol was followed within those sites to chose families at random (this is known as two stage sampling). Families where then asked how many members of the families had been killed by insurgents, Iraqi or coalition forces in the period from July 2003 to the date of the survey. The surveyors took considerable risks in doing the survey and although there were issues with one urban site (Fallujah during the interim between the murder of the American contractors and the Marine assault), overall it appears to have produced a reasonably representative sample. About 1,000 families were interviewed in the course of the survey.

Based on this sample, the researchers concluded that 89,000 Iraqi’s had died since the end of the invasion. However the 95% confidence interval (the interval which will contain the true population mean in 95 of 100 experiments) extended from 8,000 to 170,000. A 95% C.I. is constructed as \( \bar{x} \pm 1.96\times \text{SD}(\bar{x}) \). The survey was roundly criticized because it was so imprecise (and because the mean estimate was so high, it’s a long tradition to shoot the messenger).

One argument made against the survey’s conclusions was that the results really were not so bad because there was a 95% chance that only 8,000 Iraqi’s had been killed. Is it true that, based on the survey data, there is a 95% chance that only 8,000 Iraqi’s had been killed in the period since the end of the invasion? Explain your answer and support with appropriate calculations.

What we are being asked is, what is the likelihood of a population value of 8,000, given a sample mean of 89,000. The central limit theorem tells us that the likelihood of a particular (or particular range of) outcomes are normally distributed around the sample mean. How do we use this? First, the standard deviation of \( \bar{x} \) is 41,326 (=81,000/1.96). Now, we have a 95 percent interval that contains 8,000. Now lets set up a 90% interval. This would be \( \bar{x} \pm 1.645\times \text{SD}(\bar{x}) \) or, 21,017 to 156,982 (89,000 ± 1.645*41326). So the gap between the lower bound of a 95% percent interval and a 90% interval is from 8,000 to 21,017. The difference in probabilities is (95%-90%)/2 or 2.5%. If the likelihood that the population mean is between 8,000 and 21,017 is 2.5%, the likelihood of 8,000 cannot be 95%. Although the full confidence interval has a 95% likelihood of containing the population mean, the most likely outcomes are close to the sample mean and subsets of the C.I. are only part of the 95% probability.