Question 1 (6 points)

An experiment is performed in which a needle 1 inch long is tossed 1000 times onto a tabletop engraved with parallel lines exactly 1 inch apart. Of these 1000 tosses, 643 result in the needle crossing a line.

We wish to estimate the true proportion of tosses which result in a line crossing.

(a) Compute \( \hat{p} \) and its standard error (give numeric answers, not formulas). Tell whether each of these is an example of a statistic or a parameter. (1 point)

\[
\hat{p} = \frac{643}{1000} = 0.643
\]

\[
SE_\hat{p} = \sqrt{\frac{0.643(0.357)}{1000}} = 0.015.
\]

Each of these quantities is a statistic.

(b) Give a 95% confidence interval for the true population proportion. (1 point)

\[
\hat{p} \pm z^*SE_\hat{p} = 0.643 \pm 1.96(0.015) = 0.643 \pm 0.030
\]

(c) A friend performs a similar experiment, tossing a needle of unknown length onto the same lined tabletop 1000 times. She observes 625 line crossings. You wish to test

\[ H_0: \text{The true proportions in the two experiments are equal.} \]

Calculate the appropriate \( z \) statistic for the test (do not find a p-value). (1 point)

\[
z = \frac{0.643 - 0.625}{\sqrt{\frac{1268}{2000} \left(1 - \frac{1268}{2000}\right) \left(\frac{1}{1000} + \frac{1}{1000}\right)}} = 0.836
\]
Question 1 continued

It is possible to prove that the true proportion in the original 1000-toss experiment with the
1-inch needle is equal to $2/\pi$, which is 0.63662. Use this new information in parts (d) and
(e) only.

(d) Give the true value of the mean and standard deviation of the $\hat{p}$ statistic for repetitions
of the experiment. (1 point)

The mean and standard deviation of the sampling distribution of $\hat{p}$ are $p$ and $\sqrt{p(1-p)/n}$,
respectively. In this case, we know that $p = 0.63662$. Thus, we get

$$\mu_{\hat{p}} = 0.63662 \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{0.63662(0.36338)}{1000}} = 0.0152.$$  

(e) Name the theorem which tells us the approximate distribution of $\hat{p}$. Using this approx-
imate distribution, find the approximate probability that if you repeated the experiment,
you would observe a value of $\hat{p}$ less than 0.6215. (2 points)

The central limit theorem implies that $\hat{p}$ is approximately normally distributed with mean
$\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$ (see part (d)). Converting the number 0.6215 into a z-score,
then, gives $(0.6215 - 0.63662)/(0.0152) = -0.995$. We then use Table A to conclude that

$$P(\hat{p} < 0.6215) \approx P(Z < -0.995) = 0.16.$$
Question 2 (7 points)
Researchers studying vitamin C in a commodity called wheat soy blend (WSB) were concerned that some of the vitamin C content would be destroyed as a result of storage and shipment of the commodity to its final destination in Haiti. The researchers specially marked a collection of bags at the factory and took a sample from each of them to measure the vitamin C content. Five months later in Haiti they found each of the specially marked bags and took samples. The data consist of two vitamin C measures for each bag, one at the time of factory production and the other five months later in Haiti. The units are milligrams of vitamin C per 100 grams of WSB. There were 27 bags marked in the study.

The data are summarized below (you may not need all the information in the table):

<table>
<thead>
<tr>
<th>Factory</th>
<th>Haiti</th>
<th>Difference (Factory minus Haiti)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x} = 42.85$</td>
<td>$\bar{x} = 37.52$</td>
<td>$\bar{x} = 5.33$</td>
</tr>
<tr>
<td>$s = 4.79$</td>
<td>$s = 2.44$</td>
<td>$s = 5.59$</td>
</tr>
<tr>
<td>$n = 27$</td>
<td>$n = 27$</td>
<td>$n = 27$</td>
</tr>
</tbody>
</table>

(a) Set up hypotheses to examine the question of interest to these researchers. Should the alternative hypothesis be 1-sided or 2-sided? (1 point)

$H_0 :$ The mean difference (factory minus Haiti) equals 0 ($\mu_D = 0$).

$H_a :$ The mean difference is greater than 0 ($\mu_D > 0$).

This is a one-sided alternative, since we’re only concerned with a loss of vitamin C.

(b) Perform the appropriate significance test at the 5% level and explain your results. In your answer, clearly mark your test statistic and p-value. (3 points)

We perform a paired (or matched) t-test, which is the same as a one-sample t-test on the differences. Using the third column in the table above, the t statistic is

$$t = \frac{5.33 - 0}{5.59/\sqrt{27}} = 4.95$$

on 26 degrees of freedom. From table D, we find that the corresponding p-value is less than 0.0005. This leads us to reject $H_0$ and conclude that there is a significant decrease in vitamin C content.
Question 2 continued

(c) Find a 90% confidence interval for the mean vitamin C content at the factory. (2 points)

We need a t-based confidence interval since we don’t know the exact standard deviation of
the population of all possible factory measurements. On 26 df, the $t^*$ value for the 90% interval is 1.706. Thus, the CI is

$$
\bar{x} \pm t^* \left( \frac{s}{\sqrt{n}} \right) = 42.85 \pm 1.706 \left( \frac{4.79}{\sqrt{27}} \right) = 42.85 \pm 1.57.
$$

(d) State precisely what “90% confidence interval” means. (1 point)

If we resampled many times and constructed a new 90% CI each time, then 90% of these intervals would enclose the true factory mean $\mu_F$. 

**Question 3 (7 points)**

We wish to determine whether there are gender differences in the progress of students in doctoral programs in the U.S. We obtain data from University X, a major research university, which classifies all students entering its PhD programs in a given year by their status 6 years later. The categories used are as follows: completed the degree, dropped out, and still enrolled.

Below is a segment of Minitab output with some of the values covered up:

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>Dropped</th>
<th>Still</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>42.79</td>
<td>42.79</td>
<td>14.41</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>98</td>
<td>33</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>116.5</td>
<td>75.1</td>
<td>37.4</td>
<td>229.00</td>
</tr>
<tr>
<td>Male</td>
<td>53.21</td>
<td>29.94</td>
<td>16.86</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>423</td>
<td>238</td>
<td>134</td>
<td>795</td>
</tr>
<tr>
<td></td>
<td>404.5</td>
<td>260.9</td>
<td>129.6</td>
<td>795.00</td>
</tr>
<tr>
<td>All</td>
<td>50.88</td>
<td>32.81</td>
<td>16.31</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>521</td>
<td>336</td>
<td>167</td>
<td>1024</td>
</tr>
<tr>
<td></td>
<td>521.00</td>
<td>336.00</td>
<td>167.00</td>
<td>1024.00</td>
</tr>
</tbody>
</table>

Chi-Square = 13.398, DF = 2, P-Value = 0.0012

**Cell Contents --**

- % of Row
- Count
- Exp Freq

**(a)** The expected frequencies have been replaced in the output by asterisks. Fill in all of the expected frequencies by writing the correct values next to the asterisks. *(1 point)*

These have been filled in above (the asterisks are gone). For each of the six cells, the formula for expected value is (column total)× (row total)/(grand total).

**(b)** Identify the population of interest. Is the above sample a SRS from this population? Give at least one potential source of bias. *(2 points)*

The population of interest is all students in doctoral programs in the U.S. Our sample is not a SRS because a SRS is, by definition, a sample selected such that ALL possible samples of the same size from the population are equally likely. We would never get the 1024 students of University X if we took a true SRS of 1024 from the U.S. population of doctoral students. Our results may be biased because the sample only represents a single institution.
Question 3 continued

(c) State the appropriate null and alternative hypotheses for performing the test of interest. (1 point)

\[ H_0 : \text{There is no gender difference with regard to status after 6 years in a doctoral program.} \]
\[ H_a : \text{There is a difference between males and females with regard to 6-year status.} \]

(d) Give the correct degrees of freedom for the statistic and find a p-value. (1 point)

This has been added to the output on the preceding page. There are 2 degrees of freedom and Table F gives a p-value between 0.001 and 0.0025.

(e) The p-value you find in part (d) should be less than 5%, giving strong evidence against the null hypothesis. Using the row percentages as a guide, describe to a non-statistical reader the qualitative results of this study. (1 point)

While roughly equal proportions of men and women are still enrolled after 6 years, there is a wide gap in the “completed” and “dropped out” categories. A larger proportion of men than women had completed, while a larger proportion of women than men had dropped out.

(f) University X also records the age of each of its entering PhD students. Suppose we wish to determine whether there is a gender difference in average age of entrance into a PhD program. Explain why a chi-square analysis such as the one in this question would not be an appropriate way to answer this question and suggest a more appropriate type of analysis. (1 point)

A chi-square test compares two categorical variables, not a categorical and a quantitative variable. We should use a 2-sample t-test instead.