1 Solution to Problem 1.

In a sample of sixty four corn plants of a certain variety grown under laboratory conditions, the average height after 16 weeks was 1.84 m. and the sample standard deviation was .24 m.

(a) Calculate a 95% confidence interval for the population mean height of this variety of corn grown under laboratory conditions.

Solution. The sample size $n = 64$ is large enough (namely, $n \geq 30$) that we can justify the method in Box 4.1, p. 130. We have $\bar{y} = \bar{y} = 1.84$ (observed value) and $\sigma_\theta = s/\sqrt{n} = .24/\sqrt{64} = .24/8 = .03$. Also, $z(\alpha/2) = z(.025) = 1.96$, so the confidence interval is

$$1.84 \pm 1.96 \times .03 = 1.84 \pm .06 = (1.78, 1.90).$$

(b) Construct a 95% prediction interval for a single new plant of the same variety grown under the same conditions.

Solution. For the prediction interval, we can apply Box 4.10, p. 161, noting that since $n = 64$ is large we can effectively use the $z(.025)$ in place of $t(.025, 63)$, although it won’t make much difference either way. Note that $t(.025, 40) = 2.0211$ and $t(.025, 80) = 1.9901$, so we have that $1.9901 < t(.025, 63) < 2.0211$, and it is close to $z(.025) = 1.96$. Hence, the 95% prediction interval is

$$1.84 \pm 1.96 * \sqrt{1 + 1/64} \times .24 = 1.84 \pm .47 = (1.37, 2.31).$$

If we had used $t(.025, 63) = 1.998341$, we would have gotten $1.84 \pm .48 = (1.36, 2.32)$, which differs in the second decimal.

Note: Several students wrote something like “1.37 < $\mu$ < 2.31” but in fact the correct thing to write here is “1.37 < $Y$ < 2.31” with 95% confidence,” where $Y$ denotes the new observation (that hasn’t been observed yet).

(c) In a previous experiment with a different variety of corn, after 16 weeks the
average height of forty nine plants was 1.72 m. and the standard deviation was .18 m. Is there a statistically significant difference in the mean heights of the two varieties? State an appropriate null and alternative (research) hypothesis, compute an appropriate test statistic and $P$-value, and determine if you can reject the null hypothesis at the $\alpha = 0.05$ level of significance.

**Solution.** The parameter of interest is $\theta = \mu_1 - \mu_2$ where Population 1 is the first variety of corn and Population 2 is the second. We had no a priori reason that is given that one type of corn grows faster than the other, so our hypotheses to be tested are

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta \neq 0.$$ 

The point estimate is

$$\hat{\theta} = \bar{y}_1 - \bar{y}_2 = 1.84 - 1.72 = 0.12m.$$ 

The standard error is

$$\hat{\sigma}_\theta = \sqrt{s_1^2/n_1 + s_2^2/n_2} = \sqrt{.24^2/65 + .18^2/49} = 0.039.$$ 

As both sample sizes are large, we apply Box 4.1. The $z$ test statistic is

$$0.12/0.039 = 3.08.$$ 

From Table C.1, $Pr(Z > 3.08) = 0.0010$. Since this is a two sided test, we have to double this probability to get the $P$-value, i.e. the $P$-value is 0.002. This is statistically significant, i.e. we reject $H_0$ and conclude that the two varieties of corn do not grow (in the lab) at the same average rate.

**Notes:** Most of the students forgot that in a two sided test you must double the tail area to get the $P$-value. Also, many students treated it as a 2-sample t-test instead of a 2 sample z-test (Box 4.4 instead of Box 4.1). It is not wrong to do this - the t-test turns into the z-test as the sample sizes grow, but you waste a lot of
time doing all the calculations for the t-test and trying to figure out how to apply the tables when the degrees of freedom (110) doesn’t appear exactly in Table C.2.
2 Solution for Problem 2.

Below are samples of size 5 observations from 6 different distributions:

(1) Binomial with parameters \( n = 10 \) and \( \pi = .5 \);

(2) Binomial with parameters \( n = 10 \) and \( \pi = .2 \);

(3) Poisson with parameter \( \lambda = 10 \);

(4) Normal with parameters \( \mu = 30 \) and \( \sigma^2 = 100 \);

(5) Normal with parameters \( \mu = 30 \) and \( \sigma^2 = 1 \);

(6) \( F \) with parameters \( \nu_1 = \nu_2 = 4 \).

Match the samples to the distributions. The samples are:

(A) 42.40, 33.15, 28.12, 32.18, 32.40

(B) 9, 9, 11, 6, 10

(C) 28.54, 29.56, 31.17, 30.80, 30.31

(D) 0.19, 1.27, 0.24, 1.03, 6.48

(E) 1, 2, 1, 2, 2

(F) 4, 3, 4, 4, 4

Solution: The binomial and Poisson distributions result from counting: they apply to values only in the nonnegative integers. Thus, Samples B, E, and F only can correspond to distributions 1, 2, and 3. Binomial with \( n = 10 \) implies there are at most 10 successes, so the maximum value for such a distribution is 10. Since we see an 11 in sample B, this must be the Poisson distribution (which has no upper limit),
i.e. Sample B goes with Distribution 3. To finish off the discrete distributions, if $Y$ is binomial with parameters $n$ and $p$, then the average value is $np$. So if $n = 10$ and $\pi = .5$, we expect to see values around $10 \times .5 = 5$, and if $n = 10$ and $\pi = .2$ we expect to see values around $10 \times .2 = 2$. From this, we conclude that Sample F goes with Distribution A and Sample E goes with Distribution B.

Samples A, C, and D clearly go with the continuous distributions: 4, 5, and 6. If $Y$ is $N(\mu, \sigma^2)$, then values of $Y$ should be around $\mu$, and 95% of the values should be within $\mu \pm 2\sigma$. We see that Samples A and C have values around 30, so they must correspond to the normal distributions (distributions 4 and 5), and hence Sample D goes with Distribution 6. Prof. Cox mentioned in class that values of the F distribution tend to be around 1, and this is consistent with Sample D as well. Now sample A clearly has the wider range of values than Sample C, so it must go with the normal distribution having a larger variance, so we conclude Sample C goes with Distribution 5 and Sample A goes with Distribution 4.

The results are summarized in the table below:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>F</td>
<td>E</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Notes: A majority of students reversed the F and E.
3 Solution to Problem 3.

True or False: Answer “T” or “F” according as the statement is True of False.

T  F  If the P-value is 0.012, then we reject the null hypothesis at the $\alpha = 0.05$ level of significance.

Answer: TRUE Since 0.012 < 0.05, we reject the null hypothesis.

T  F  When using the $t$-distribution to construct a confidence interval for the population mean in a small sample, the most important assumption is that the population has a normal distribution.

Answer: FALSE The most important assumption is the independence assumption (random sample).

T  F  If $Y$ is a continuous random variable (i.e., its distribution is given by a probability density function), then $Pr[Y = 1] = 0$.

Answer: TRUE The area under the P.D.F. curve above $y = 1$ is 0.

T  F  Suppose we wish to test $H_0 : \theta \leq 10$ vs. $H_1 : \theta > 10$, where $\theta$ is some parameter of a population. Suppose we are told that a 95% upper confidence bound for $\theta$ is 12. Using this information, we can reject the null hypothesis at the 0.05 level of significance.

Answer: FALSE If a 95% upper confidence bound for $\theta$ is 12, then all believable values of $\theta$ are less than 12. Values too much smaller than 12 might not be believable either, but we are not given that information (on a lower confidence bound, which is what we need to decide if we can reject $H_0 : \theta \leq 10$). We cannot rule out $\theta \leq 10$ with this information.
T   F   When referring to variables, “nominal,” “qualitative,” and “ordinal” all mean the same thing.

Answer: FALSE  An ordinal variable has order in its values but a nominal variable doesn’t.
4 Solution to Problem 4.

In two independent random samples from two populations (labelled Population 1 and Population 2), we observe sample variances

\[ s_1^2 = 2.54, \quad s_2^2 = 3.92. \]

Construct a 95% confidence interval for the ratio of population standard deviations \( \sigma_1/\sigma_2 \). State what assumptions are needed to justify this confidence interval.

**Solution.** We apply Box 4.7, p. 152. We have

\[ s_1^2/s_2^2 = 2.54/3.92 = 0.6480. \]

So,

\[ s_1/s_2 = \sqrt{0.6480} = 0.805. \]

Since we aren’t given sample sizes, we cannot look up the appropriate critical values, so the best we can do is write them symbolically:

\[ \frac{0.805}{\sqrt{F(n_1 - 1, n_2 - 1, 0.025)}} < \frac{\sigma_1}{\sigma_2} < \frac{0.805}{\sqrt{F(n_1 - 1, n_2 - 1, 0.975)}}, \]

with 95% confidence. Of course,

\[ F(n_1 - 1, n_2 - 1, 0.975) = \frac{1}{F(n_2 - 1, n_1 - 1, 0.025)}, \]

which would be needed if we actually had to look up some numbers.

The assumptions would be that the two samples are random samples, that they are independent samples (independent of each other), and that the populations from which they come are both normal.

**Notes:** Clearly someone goofed up by not giving the sample sizes. I liked the solution of some students – they realized they didn’t have enough information, then said they would do the calculations based on some particular sample sizes.
5 Solution to Problem 5.

A study is conducted to investigate the effect of CO$_2$ on the growth rate of the bacterium Pseudomonas fragi. Ten samples of the bacteria are selected. Five of these samples are grown in a chamber with 0 atmospheres of CO$_2$ and the other five with 0.50 atmospheres of CO$_2$. The response is the percentage increase in the bacterial colony after 1 hour. The bacterial samples were randomly assigned to the two groups. The results are given in the table below:

<table>
<thead>
<tr>
<th>CO$_2$ level</th>
<th>Percent Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 atmospheres</td>
<td>4.5  23.5  9.2  14.1  22.4</td>
</tr>
<tr>
<td>0.50 atmospheres</td>
<td>46.2  6.6  48.0  68.4  46.9</td>
</tr>
</tbody>
</table>

The researcher believed before conducting this experiment that CO$_2$ promoted the growth of this bacteria (i.e., more atmospheres of CO$_2$ would mean higher levels of growth on the average). Explicitly define the parameter of interest. State null and alternative (research) hypotheses and carry out the appropriate test of significance. State what assumptions are needed to justify the test you perform.

Solution: Let $\mu_1$ be the population mean percentage increase in the bacterial colony after 1 hour under 0.00 atmospheres of CO$_2$, and let $\mu_2$ be the corresponding mean under 0.50 atmospheres. The parameter of interest is

$$\theta = \mu_1 - \mu_2.$$ 

The research hypothesis is that $\mu_2 > \mu_1$, i.e. we will test

$$H_0 : \theta \geq 0 \text{ vs. } H_1 : \theta < 0.$$ 

The two samples may be considered as independent: the original bacterial specimens were randomly assigned to the two groups. We will use the procedure in Box 4.4,
which requires the assumptions of (1) random samples; (2) independent samples; and (3) normal populations. Some tedious computing gives the following results:

\[
\begin{align*}
g_1 & = 14.74 \\
\bar{y}_2 & = 43.22 \\
\hat{\theta} & = -28.48 \\
s_1^2 & = 67.843 \\
s_2^2 & = 505.082 \\
\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} & = 114.585 \\
\hat{\sigma}_\hat{\theta} & = 10.704 \\
\frac{1}{k} & = 0.1978 \\
\nu & = 5 \\
t & = -2.661 \\
t(.05, 5) & = 2.0150 \\
t(.025, 5) & = 2.5706 \\
t(.01, 5) & = 3.3649.
\end{align*}
\]

Since the observed \( t \) satisfies \( t < -t(.05, 5) \) we conclude that we can reject the null hypothesis at the \( \alpha = 0.05 \) level of significance. In fact, our best estimate of the \( P \)-value is

\[
.025 < P-value < .01.
\]

Notes: Many students worked this problem as if it were paired data. Note that there is no relation between the data values in the same columns: we could have randomly rearranged these data in the rows and it would have been the same. Also, we could have had more samples in one data set than the other. One might think that
the presentation format was chosen so as to induce this confusion. In fact, it is not
totally wrong to treat independent samples as paired samples (assuming the sample
sizes are the same, of course) if one is testing for a difference in the means. The level
of significance (type I error probability) is OK, but there is less power (more type II
error probability) to detect deviations from the null hypothesis, which is a desirable
aspect. Anyway, if you did treat them as paired samples (as in Box 4.5), here are the
numbers:

\[
\begin{align*}
d & = -28.48 \\
s_d & = 27.49149 \\
t & = \sqrt{5} \times \frac{-28.48}{27.49149} = -2.31647 \\
\nu & = 5 - 1 = 4 \\
t(4,.05) & = 2.138 \\
t(4,.025) & = 2.7764.
\end{align*}
\]

Since \(-2.31647 < -2.138\), we can reject \(H_0\) at the \(\alpha = 0.05\) level of significance, and
in fact we see that the \(P\)-value satisfies

\[.025 < P-value < .05.\]

Our \(P\)-value is clearly larger than from the previous (better) two sample analysis.
This is indicative of what was said above: we did reject in this case, but if the data
were not so strongly against \(H_0\), there is the possibility that the paired sample analysis
would not find a significant difference (commit a type II error: fail to reject \(H_0\) when
it is false) whereas the two sample analysis would find a significant difference.
6 Summary of Scores and Grades.

Basic descriptive statistics of the scores:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>27</td>
</tr>
<tr>
<td>1st Quart.</td>
<td>79.75</td>
</tr>
<tr>
<td>Mean</td>
<td>79.95</td>
</tr>
<tr>
<td>Median</td>
<td>87.0</td>
</tr>
<tr>
<td>3rd Quart.</td>
<td>91.75</td>
</tr>
<tr>
<td>Max</td>
<td>97</td>
</tr>
<tr>
<td>Total N</td>
<td>22</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>17.93</td>
</tr>
</tbody>
</table>

Approximate grade assignments:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score Range</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>88-97</td>
<td>11</td>
</tr>
<tr>
<td>B</td>
<td>79-86</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>48-69</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>27</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Of course, it is the numerical grade that counts. The letter grades are an indication of how grades will be assigned in the end (when there is a total numerical grade).