1. 6.42. In a study of human blood types in nonhuman primates, a sample of 71 orangutans were tested and 14 were found to be blood type B.

   a. Construct a 95% confidence interval for the relative frequency of blood type B in the orangutan population.

   Correct:
   \[ \hat{p} = \frac{(14 + 1.96^2/2)}{(71 + 1.96^2)} = 0.2127; \]
   \[ SE = \sqrt{(0.2127)(0.7873)/(71 + 1.96^2)} = \sqrt{0.0022} = 0.0473. \]
   95% confidence interval
   \[ 0.2127 \pm (1.96)(0.0473) \]
   (0.120, 0.305) or 0.120 < \( p \) < 0.305.

   b. Interpret the interval you just computed in part (a).

   Correct:
   We are 95% confident that the proportion of type B blood in the orangutan population is no less than 0.120 and no more than 0.305.

2. 6.49(modified). The Luso variety of wheat is resistant to the Hessian fly. In order to understand the genetic mechanism controlling this resistance, an agronomist plans to examine the progeny of a certain cross involving Luso and a nonresistant variety. Each progeny plant will be classified as resistant or susceptible and the agronomist will estimate the proportion of progeny that are resistant. He'd like to report a 95% confidence interval for the proportion of resistant progeny.

   a. If the researcher believes the proportion of resistant progeny is 0.75, how many progeny does he need to classify in order to guarantee that the standard error of his estimate of this proportion will not exceed .05?

   Correct:
   Agresti-Coull Sample Size Determination
   It is necessary to use \( p_o = 0.75 \) because the researcher has an idea of what proportion are resistant.
   Since we want SE less than or equal to 0.05, then we want MOE less than or equal to \( e_o = (1.96)(0.05) = 0.098 \).
   Then, since \( (Z_{\alpha/2})^2 = 1.96^2 = 3.8416 \) and \( p_o = 0.75 \), the formula gives
   \[ n \geq \{ (3.8416)(0.75)(0.25)/(0.098)^2 \} - 3.8416 = 75 - 3.8416 = 71.1584 \]
   So choose \( n = 72 \).

   b. If the researcher has no idea what the proportion of resistant progeny may be, how many progeny does he need to classify in order to guarantee that the standard error of his estimate of this proportion will not exceed .05?

   Correct:
   Agresti-Coull Sample Size Determination
   It is necessary to use \( p_o = 0.5 \) because the proportion of progeny that are resistant is unknown in advance.
   Since we want SE less than or equal to 0.05, then we want MOE less than or equal to \( e_o = (1.96)(0.05) = 0.098 \).
   Then, since \( (Z_{\alpha/2})^2 = 1.96^2 = 3.8416 \) and \( p_o = 0.5 \), the formula gives
   \[ n \geq \{ (3.8416)(0.5)(0.5)/(0.098)^2 \} - 3.8416 = 100 - 3.8416 = 96.1584 \]
   So choose \( n = 97 \).
3. 10.58. Experimental studies of cancer often use strains of animals that have a naturally high incidence of tumors. In one such experiment, tumor prone mice were kept in a sterile environment with one group of mice maintained entirely germ free and the other group of mice exposed to the intestinal bacterium *Escherichia coli*. The accompanying table shows the incidence of liver tumors.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Number of Mice</th>
<th>Mice with Liver Tumors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germ Free</td>
<td>49</td>
<td>19</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

Let $p_1$ and $p_2$ represent the probabilities of liver tumors under the germ-free and the *E. coli* conditions, respectively.

a. Construct a 95% confidence interval for $(p_1 - p_2)$

**Correct:**

\[
\hat{p}_1 = 20/51 = 0.3922, \quad \hat{p}_2 = 9/15 = .6000
\]

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(0.3922)(0.6078)}{51} + \frac{(0.4)(0.6)}{15}} = 0.143784913.
\]

\[
(0.3922 - 0.6000) \pm (1.96)(0.143784913) = (-0.490, 0.074) \text{ or } -0.490 < p_1 - p_2 < 0.074.
\]

b. Interpret the confidence interval from part (a). That is, explain what the interval tells you about tumor probabilities.

**Correct:**

With 95% confidence, we are unsure under which condition the probability of liver tumor is larger. If the probability is higher under the germ free condition, it is by as much as 0.074. If it is higher under the *E. coli* condition, it is by as much as 0.490.

4. 10.57. Estrus synchronization products are used to bring cows into heat at a predictable time so that they can be reliably impregnated by artificial insemination. In a study of two estrus synchronization products, 42 mature cows (aged 4–8 years) were randomly allocated to receive either product A or product B, and then all cows were bred by artificial insemination. The table shows how many of the inseminations resulted in pregnancy.

<table>
<thead>
<tr>
<th>Product</th>
<th>Number of Cows</th>
<th>Number of Cows Pregnant</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>21</td>
<td>15</td>
</tr>
</tbody>
</table>

Let $p_1$ and $p_2$ represent the probabilities of pregnancy using products A and B, respectively.

a. Construct a 95% confidence interval for $(p_1 - p_2)$

**Correct:**

\[
\hat{p}_1 = 9/23 = 0.3913, \quad \hat{p}_2 = 16/23 = 0.6957
\]

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(0.6957)(0.3043)}{23} + \frac{(0.3913)(0.6087)}{23}} = 0.1399
\]

\[
(0.3913 - 0.6957) \pm (1.96)(0.1399) = (-0.58, -0.03) \text{ or } -0.58 < p_1 - p_2 < -0.03.
\]

b. Interpret the interval you just computed in part (a).

**Correct:**

We are 95% confident that the probability of mature cows (aged 4–8 years) which become pregnant by artificial insemination using Product B is larger than that of Product A by as little as 0.03 or as much as 0.58.

5. A population-based study, reported online in *Stroke: Journal of the American Heart Association*, related transient ischemic attack (TIA – a “mini-stroke”) and a person’s risk for a subsequent myocardial infarction. The estimated relative risk for patients whose TIA occurred before age 60 was calculated to be 15.1, with a 95% confidence interval of (4.11, 38.6). Interpret this interval.

We are 95% confident that the long run risk of myocardial infarction for persons with a previous TIA is 4.11 to 38.6 times greater than that for those who have not had a TIA.

Although the overall prevalence of bladder cancer is small (< 0.02%), it leads to death in approximately 20% of cases, so bladder cancer is an active research area. Several epidemiological studies suggested an association between the risk of bladder cancer and the exposure to trihalomethanes (THMs), the main disinfection by-products (DBPs) of chlorinated water. The authors analyzed data from pooling three European case-control studies from France, Finland, and Spain (5467 individuals: 2381 cases and 3086 controls). Individual exposure to THMs was calculated combining information on residential history, estimates of the average total THMs (TTHM) level in tap water at the successive residences and personal water consumption. An odds-ratio was calculated for men exposed to an average residential TTHM level > 50 µg/l (OR=1.47 (1.05; 2.05)) when compared to men exposed to levels ≤ 5 µg/l.

a. Interpret the confidence interval in the context of the setting.
We are 95% confident that the long run odds of bladder cancer for men exposed to an average residential TTHM level > 50 µg/l is 1.05 to 2.05 times the odds for men exposed to levels ≤ 5 µg/l.
b. Based on this confidence interval, can you determine whether the odds of bladder cancer are significantly larger at the 0.05 significance level? Why or why not?
Yes. The interval does not contain the number 1 (which would indicate the long run odds may be the same under either exposure level) so a non-directional test at the 0.05 level would reject the null hypothesis.
c. Why did the researchers report a confidence interval on the odds ratio rather than relative risk?
A case control study does not give an estimate of the prevalence of bladder cancer (since we choose the number of “cases” and “controls”), so we cannot estimate Pr{bladder cancer | exposure level}. So, we cannot compute the risk we are interested in. Because of the numeric equivalence of the disease odds ratio and the exposure odds ratio, we can compute an estimate of the odds ratio we want.
d. Can we use this sample odds ratio of 1.47 as an estimate of the relative risk? Why or why not?
Yes, we can since the prevalence of bladder cancer is very close to zero and this makes the odds ratio a good approximation to the relative risk. (But, if we calculated a confidence interval on the odds ratio, the interpretation should be about the odds ratio – not relative risk.)
7. 10.25. A sample of 276 healthy adult volunteers were asked about the variety of social networks that they were in (e.g., relationships with parents, close neighbors, workmates, etc.). They were then given nasal drops containing a rhinovirus and were quarantined for 5 days. Of the 123 subjects who were in 5 or fewer types of social relationships, 57 (46.3%) developed colds. Of 153 who were in at least 6 types of social relationships, 52 (34.0%) developed colds. Thus, the data suggest that having more types of social relationships helps one develop resistance to the common cold. Determine whether this difference is statistically significant. That is, use a chi-square test to test the null hypothesis that the probability of getting a cold does not depend on the number of social relationships a person is in. Use a non-directional alternative and let \( \alpha = 0.05 \).

<table>
<thead>
<tr>
<th></th>
<th>Five or Fewer</th>
<th>Six or More</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>No Cold</td>
<td>66</td>
<td>101</td>
</tr>
</tbody>
</table>

Correct:
Let \( p \) denote the probability of a cold, let 1 denote five or fewer types of social relationships, and let 2 denote six or more types of social relationships.

1) \( \alpha = 0.05 \)

2) \( H_0 \): probability of developing a cold does not depend on the number of types of social relationships (\( p_1 = p_2 \))

\( H_A \): probability of developing a cold depends on the number of types of social relationships (\( p_1 \neq p_2 \))

3) \( X^2_s = 4.355 \)

4) \( P = 0.037 \)

5) Since \( P < \alpha \), we reject \( H_0 \).

6) There is significant evidence to conclude that the probability of developing a cold depends on the number of types of social relationships.

8. 10.75. One explanation for the widespread incidence of the hereditary condition known as sickle-cell trait is that the trait confers some protection against malarial infection. In one investigation, 543 African children were checked for the trait and for malaria. The results are shown in the table. Do the data provide evidence (\( \alpha = 0.05 \)) that malaria and having the sickle cell trait are negatively associated?

<table>
<thead>
<tr>
<th></th>
<th>Heavily Infected</th>
<th>Light or Non Infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sickle Cell Trait</td>
<td>36</td>
<td>100</td>
</tr>
<tr>
<td>No Sickle Cell Trait</td>
<td>152</td>
<td>255</td>
</tr>
</tbody>
</table>

Note: This test is carried out in the same fashion as our usual test for association with a 2x2 contingency table, with the exception that our \( H_A \) is directional. I’ve already checked that the data deviate in the direction of \( H_A \), so you can proceed with the hypothesis test. Don’t forget to calculate the \( P \)-value considering a directional alternative.

Correct:

1) \( \alpha = 0.05 \)

2) \( H_0 \): Sickle-cell trait and malaria are not associated

\( H_A \): Sickle-cell trait and malaria are negatively associated

3) \( X^2_s = 5.327 \)

4) \( P = (1/2)(0.021) = 0.0105 \)

5) \( P < \alpha \), reject \( H_0 \)

6) There is significant evidence to conclude that sickle-cell trait and malaria are negatively associated.

[Remark: This finding suggests that sickle-cell trait confers some protection against malaria; however, it should be noted that the sample excludes children who died of malaria]