Solutions to Practice Final Exam for Statistics 371,

December 12, 2014

1. The confidence interval is given by formula (16.8) on page 408 of the Course Notes. Using the values from our data set, first we get for the radical term:

\[
\sqrt{500(100 + 50) - (100 - 50)^2} = \sqrt{145.29} = 12.05.
\]

Thus, the CI is

\[
100 - 50 \quad 1.96 \quad (12.05) = 0.100 \pm 0.047.
\]

2. The confidence interval is given by formula (16.8) on page 408 of the Course Notes. Using the values from our data set, first we get for the radical term:

\[
\sqrt{700(120 + 64) - (120 - 64)^2} = \sqrt{179.78} = 13.41.
\]

Thus, the CI is

\[
120 - 64 \quad 1.96 \quad (13.41) = 0.080 \pm 0.038.
\]

3. In the collapsed table we obtain

\[
\hat{p}_1 = 0.500 > 0.450 = \hat{p}_2.
\]

Thus, for Simpson’s Paradox to occur, we need to have

\[
\hat{p}_1 < \hat{p}_2
\]

in both component tables.

Let (the integer) \( x \) denote the Group 2 total in Subgroup A. Thus, we need:

\[
0.700 < 50/x \text{ or } x < 71.4 \text{ or } x \leq 71.
\]

Let (the integer) \( y \) denote the Group 2 total in Subgroup B. Thus, we need:

\[
0.300 < 40/y \text{ or } y < 133.3 \text{ or } y \leq 133.
\]

In addition, \( x + y = 200 \). Thus, the following pairs of \((x, y)\) will yield Simpson’s Paradox:

\[
(71, 129), (70, 130), (69, 131), (68, 132) \text{ and } (67, 133).
\]

4. In the collapsed table we obtain

\[
\hat{p}_1 = 0.500 > 0.473 = \hat{p}_2.
\]

Thus, for Simpson’s Paradox to occur, we need to have

\[
\hat{p}_1 < \hat{p}_2
\]

in both component tables.

Let (the integer) \( x \) denote the Group 2 total in Subgroup A. Thus, we need:

\[
0.660 < 80/x \text{ or } x < 121.2 \text{ or } x \leq 121.
\]
Let (the integer) \( y \) denote the Group 2 total in Subgroup B. Thus, we need:

\[ 0.340 < \frac{62}{y} \text{ or } y < 182.3 \text{ or } y \leq 182. \]

In addition, \( x + y = 300 \). Thus, the following pairs of \((x, y)\) will yield Simpson’s Paradox:

- \((121, 179), (120, 180), (119, 181)\)
- \((118, 182)\).

5. (a) \( \frac{218}{1500} = 0.145 \).
   (b) \( \frac{23 + 77}{1500} = \frac{100}{1500} = 0.067 \).
   (c) \( \frac{1182}{218 + 1182} = \frac{1182}{1400} = 0.844 \).
   (d) \( \frac{218}{295} = 0.739 \).
   (e) \( \frac{77}{1259} = 0.061 \).
   (f) \( \frac{241 + 77}{1500} = \frac{318}{1500} = 0.212 \).

6. (a) \( \frac{1212}{1600} = 0.758 \).
   (b) \( \frac{248 + 1212}{1600} = \frac{1460}{1600} = 0.912 \).
   (c) \( \frac{43}{43 + 97} = \frac{43}{140} = 0.307 \).
   (d) \( \frac{43}{1265} = 0.034 \).
   (e) \( \frac{248}{291} = 0.852 \).
   (f) \( \frac{291 + 97}{1600} = \frac{388}{1600} = 0.242 \).

7. (a) First, we calculate

\[
k' = \frac{250}{2} - \frac{1.282\sqrt{249}}{2} = 125 - 10.11 = 114.89.
\]

Thus, \( k = 114 \). The CI is

\[ [x_{(114)}: x_{(136)}]. \]

From the listing of the data, the lower bound is 96.6. By symmetry, the upper bound, \( u \), satisfies the equation

\[ u - 103.1 = 103.1 - 96.6 = 6.5. \]

Thus,

\[ u = 103.1 + 6.5 = 109.6, \]

and the CI is

\[ [96.6, 109.6]. \]

(b) Again, we use symmetry. The upper bound, \( u \), satisfies the equation

\[ u - 103.1 = 103.1 - 92.7 = 10.4. \]

Thus,

\[ u = 103.1 + 10.4 = 113.5. \]

(c) Again, we use symmetry. The half-width of Skylie’s interval is \( 15.4/2 = 7.7 \). Thus, her CI is

\[ 103.1 \pm 7.7. \]
8. (a) First, we calculate

$$k' = \frac{250}{2} - \frac{1.96 \sqrt{249}}{2} = \frac{125 - 15.46}{2} = 109.54.$$ 

Thus, \( k = 109 \). The CI is

$$[x(109), x(141)].$$

From the listing of the data, the lower bound is 99.6. By symmetry, the upper bound, \( u \), satisfies the equation

$$u - 113.1 = 113.1 - 99.6 = 13.5.$$ 

Thus,

$$u = 113.1 + 13.5 = 126.6,$$

and the CI is

$$[99.6, 126.6].$$

(b) Exact. Based only on the assumption that the population is a pdf, the probability given in Formula (18.7) is exact.

(c) Again, using Formula (18.7) we want to solve for \( k \) in the equation

$$1 - \frac{2k}{250} = 0.84.$$ 

This gives

$$0.16(250) = 2k \text{ or } k = 20.$$ 

Thus, the 84% PI is

$$[x(20), x(230)].$$

9. (a) From the sorted data we see that Franky’s lower bound, 92.7, is in position \( k = 111 \). Formula (18.7) on page 482 of the Course Notes states that the probability the PI will be correct is

$$1 - \frac{2(111)}{250} = 28/250 = 0.112,$$

or 11.2%.

(b) Exact. Based only on the assumption that the population is a pdf, the probability given in Formula (18.7) is exact.

(c) Again, using Formula (18.7) we want to solve for \( k \) in the equation

$$1 - \frac{2k}{250} = 0.92.$$ 

This gives

$$0.08(250) = 2k \text{ or } k = 10.$$ 

Thus, the 92% PI is

$$[x(10), x(240)].$$

10. (a) From the sorted data we see that Franky’s lower bound, 105.4, is in position \( k = 113 \). Formula (18.7) on page 482 of the Course Notes states that the probability the PI will be correct is

$$1 - \frac{2(113)}{250} = 24/250 = 0.096,$$

or 9.6%.

(b) Exact. Based only on the assumption that the population is a pdf, the probability given in Formula (18.7) is exact.

(c) Again, using Formula (18.7) we want to solve for \( k \) in the equation

$$1 - \frac{2k}{250} = 0.92.$$ 

This gives

$$0.08(250) = 2k \text{ or } k = 10.$$ 

Thus, the 92% PI is

$$[x(10), x(240)].$$
11. (a) First, we calculate the row proportions for the table of population counts: $p_1 = 6/100 = 0.06$; and $p_2 = 27/900 = 0.03$. Thus, the relative risk is $0.06/0.03 = 2.00$.

(b) I prefer doing arithmetic with integers; thus, I use the formula:

\[
\text{Odds ratio} = \frac{6(873)}{94(27)} = 2.06.
\]

(c) Trick question. With a case-control study, the data may not be used to estimate the relative risk. Here is an easy way to remember this: for these data, $\hat{p}_1 = 60/90 = 0.67$, which is a horrible estimate of $p_1 = 0.06$.

(d) The point estimate of the population odds ratio is:

\[
\frac{60(270)}{30(240)} = 2.25.
\]

Nature will note that this point estimate is a bit too large, but the researcher won’t know it.

12. (a) First, we calculate the row proportions for the table of population counts: $p_1 = 8/100 = 0.08$; and $p_2 = 18/900 = 0.02$. Thus, the relative risk is $0.08/0.02 = 4.00$.

(b) I prefer doing arithmetic with integers; thus, I use the formula:

\[
\text{Odds ratio} = \frac{8(882)}{92(18)} = 4.26.
\]

(c) Trick question. With a case-control study, the data may not be used to estimate the relative risk. Here is an easy way to remember this: for these data, $\hat{p}_1 = 105/140 = 0.75$, which is a horrible estimate of $p_1 = 0.08$.

(d) The point estimate of the population odds ratio is:

\[
\frac{105(315)}{35(245)} = 3.86.
\]

Nature will note that this point estimate is a bit too small, but the researcher won’t know it.

13. First, we note that the number of degrees of freedom is $(n - 2) = 15 - 2 = 13$. Thus, from page 1 of the exam, $t^* = 2.160$. The CI is

\[
4.947 \pm 2.160(1.325) = 4.947 \pm 2.862.
\]

14. Problem 14 is the same as problem 13.

15. From, the computer output, we find that the Fit equals 234.70 with a standard error of 6.94. Thus, the CI is

\[
234.70 \pm 2.160(6.94) = 234.70 \pm 14.99.
\]

16. From, the computer output, we find that the Fit equals 205.02 with a standard error of 4.33. Thus, the CI is

\[
205.02 \pm 2.160(4.33) = 205.02 \pm 9.35.
\]

17. From, the computer output, we find that the Fit equals 224.81 with a standard error of 4.93. In addition, $s = 14.61$. Thus, the variance of the prediction is

\[
(14.61)^2 + (4.93)^2 = 237.757.
\]

Thus, the PI is

\[
224.81 \pm 2.160\sqrt{237.757} = 224.81 \pm 33.31.
\]
18. From the computer output, we find that the fit equals 214.91 with a standard error of 3.81. In addition, \( s = 14.61 \). Thus, the variance of the prediction is

\[
(14.61)^2 + (3.81)^2 = 227.968.
\]

Thus, the PI is

\[
214.91 \pm 2.160 \sqrt{227.968} = 214.91 \pm 32.61.
\]

19. (a) We know that

\[
r^2 = R^2 = 0.518.
\]

Thus, \( r = \pm 0.720 \). Because the slope of the regression line is a positive number, we know that \( r = +0.720 \).

(b) The residual is \( y - \hat{y} \), which, in the language of Minitab, is \( y - \text{Fit} \):

\[
210 - 214.91 = -4.91.
\]

(c) We need to identify the cases for which \( y \) is larger than its fit. They are cases:

1, 2, 3, 9, 11, 14 and 15.

Thus, there are seven cases with \( e > 0 \).

(d) They are cases: 9, 11 and 15.

(e) They are cases: 4, 12 and 13.

20. (a) See the answer to 19(a).

(b) The residual is \( y - \hat{y} \), which, in the language of Minitab, is \( y - \text{Fit} \):

\[
\]

(c) We need to identify the cases for which \( y \) is smaller than its fit. They are cases:

4, 5, 6, 7, 8, 10, 12 and 13.

Thus, there are eight cases with \( e < 0 \).

21. (a) Below is the table of counts for the 100 movies. Let \( A \) denote the event that BW is the hero and let \( B \) denote the event that the hero dies.

<table>
<thead>
<tr>
<th>( A )</th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>15</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td></td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>87</td>
<td>100</td>
</tr>
</tbody>
</table>

If we divide each entry above by 100, we get the table of probabilities:

<table>
<thead>
<tr>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>( A^c )</td>
<td>0.03</td>
<td>0.72</td>
</tr>
<tr>
<td>Total</td>
<td>0.13</td>
<td>0.87</td>
</tr>
</tbody>
</table>

(b) This is

\[
P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.10}{0.13} = 0.769.
\]

If you prefer logic to formulas, we see from the table of counts that of the 13 movies in which the hero dies, 10 times it is BW; hence, the answer is 10/13, as above.

(c) This is

\[
P(A|B^c) = \frac{P(AB^c)}{P(B^c)} = \frac{0.15}{0.87} = 0.172.
\]

If you prefer logic to formulas, we see from the table of counts that of the 87 movies in which the hero does not die, 15 times it is BW; hence, the answer is 15/87, as above.
22. (a) Below is the table of counts for the 100 movies. Let $A$ denote the event that BW is the hero and let $B$ denote the event that the hero dies.

<table>
<thead>
<tr>
<th></th>
<th>$B$</th>
<th>$B^c$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>9</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>$A^c$</td>
<td>8</td>
<td>72</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>83</td>
<td>100</td>
</tr>
</tbody>
</table>

If we divide each entry above by 100, we get the table of probabilities:

<table>
<thead>
<tr>
<th></th>
<th>$B$</th>
<th>$B^c$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0.09</td>
<td>0.11</td>
<td>0.20</td>
</tr>
<tr>
<td>$A^c$</td>
<td>0.08</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>Total</td>
<td>0.17</td>
<td>0.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(b) This is

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.09}{0.17} = 0.529.$$ 

If you prefer logic to formulas, we see from the table of counts that of the 17 movies in which the hero dies, 9 times it is BW; hence, the answer is $9/17$, as above.

(c) This is

$$P(A|B^c) = \frac{P(AB^c)}{P(B^c)} = \frac{0.11}{0.83} = 0.133.$$ 

If you prefer logic to formulas, we see from the table of counts that of the 83 movies in which the hero does not die, 11 times it is BW; hence, the answer is $11/83$, as above.

23. (a) First, the number of degrees of freedom is

$$n_1 + n_2 - 2 = 10 + 10 - 2 = 18.$$ 

Thus, from the first page of the exam, $t^* = 2.101$. Next, the pooled variance equals

$$\frac{(5.63)^2 + (4.02)^2}{2} = 23.93.$$ 

Thus, $s_p = 4.89$. Raymond’s 95% CI is

$$(167.00 - 164.00) \pm \frac{2.101(4.89)}{\sqrt{(1/10) + (1/10)}} = 3.00 \pm 4.59.$$ 

(b) First, the number of degrees of freedom is

$$n - 1 = 10 - 1 = 9.$$ 

Thus, from the first page of the exam, $t^* = 2.262$.

The variance of the differences is

$$(5.63)^2 + (4.02)^2 - 2(0.400)(5.63)(4.02) = 29.75.$$ 

Thus, $s_d = 5.45$. Tom’s 95% CI is

$$3.00 \pm 2.262(5.45/\sqrt{10}) = 3.00 \pm 3.90.$$ 

(c) This is a trick question. If you read and remember the material in Section 20.4.2 you will know that Tom’s analysis is invalid. The error rate of his interval will be much larger than the nominal rate of 0.05.
24. (a) First, the number of degrees of freedom is
\[ n_1 + n_2 - 2 = 12 + 12 - 2 = 22. \]
Thus, from the first page of the exam, \( t^* = 2.074 \). Next, the pooled variance equals
\[ \frac{(5.13)^2 + (3.82)^2}{2} = 20.45. \]
Thus, \( s_p = 4.52 \). Raymond’s 95% CI is \((166.00 - 162.00)\pm 4.00 \pm 3.83.\)
(b) First, the number of degrees of freedom is
\[ n - 1 = 12 - 1 = 11. \]
Thus, from the first page of the exam, \( t^* = 2.201 \).
The variance of the differences is
\[ (5.13)^2 + (3.82)^2 - 2(0.450)(5.13)(3.82) = 23.26. \]
Thus, \( s_d = 4.82 \). Tom’s 95% CI is
\[ 4.00 \pm 2.201(4.82/\sqrt{12}) = 4.00 \pm 3.06. \]
(c) This is a trick question. If you read and remember the material in Section 20.4.2 you will know that Tom’s analysis is invalid. The error rate of his interval will be much larger than the nominal rate of 0.05.

25. (a) First, we compute
\[ \hat{p}_1 = 0.40 \text{ and } \hat{p}_2 = 0.20, \]
which yield
\[ \hat{q}_1 = 0.60 \text{ and } \hat{q}_2 = 0.80. \]
Thus, the 95% CI for \( p_1 - p_2 \) is
\[ 0.200 \pm 1.96 \sqrt{\frac{0.4(0.6)}{100} + \frac{0.2(0.8)}{100}} = 0.200 \pm 0.123. \]
(b) The study is cross-sectional.

26. (a) First, we compute
\[ \hat{p}_1 = 0.50 \text{ and } \hat{p}_2 = 0.34, \]
which yield
\[ \hat{q}_1 = 0.50 \text{ and } \hat{q}_2 = 0.66. \]
Thus, the 95% CI for \( p_1 - p_2 \) is
\[ 0.160 \pm \frac{1.96}{\sqrt{150}} \frac{0.5(0.5)}{150} + \frac{0.34(0.66)}{150} = 0.160 \pm 0.110. \]
(b) The study is cross-sectional.

27. We know from Chapter 17 that sometimes Gosset works well and sometimes it doesn’t. Of the given counts, two indicate that Gosset is working well—512 and 523—and the other two indicate that it is not.

When Gosset is performing poorly, then its performance will improve if \( n \) increases. It therefore follows that: the 718 errors are for \( n = 10 \); and the 607 errors are for \( n = 20 \). The last two cases are tricky. The observed counts of 512 and 523 indicate that the true probability of an incorrect
interval could be the target, 0.05; i.e., the difference between these counts can be attributed to random error. As a result, it is impossible to know which count goes with \( n = 40 \) and which goes with \( n = 80 \).

28. Refer to the previous solution. The exact same reasoning shows that the 732 errors are for \( n = 10 \); the 621 errors are for \( n = 20 \). Finally, we cannot assign the remaining counts, 509 and 521, to their sample sizes.

29. The proportion of successes (made shots) in the three years are: 0.765, 0.785 and 0.813.

I will define population 1 [2] to be the 1985–86 [1984-85] season. Why? This way, \( p_1 - p_2 \) is the amount Erving’s ability improved from the earlier year to the latter year. The 95% CI is \((0.785 - 0.765) \pm \)

\[
1.96 \sqrt{\frac{0.785(0.215)}{368} + \frac{0.765(0.235)}{442}} = 0.020 \pm 0.058.
\]

This interval shows that he might have improved by as little as \(-0.038\) or as much as \(+0.078\). Note that improving by a negative amount is more commonly referred to as whatever is the opposite of improving.

30. Refer to the previous solution.

I will define population 1 [2] to be the 1986–87 [1985-86] season. Why? This way, \( p_1 - p_2 \) is the amount Erving’s ability improved from the earlier year to the latter year. The 95% CI is \((0.813 - 0.785) \pm \)

\[
1.96 \sqrt{\frac{0.813(0.187)}{235} + \frac{0.785(0.215)}{368}} = 0.028 \pm 0.065.
\]

This interval shows that he might have improved by as little as \(-0.037\) or as much as \(+0.093\). Note that improving by a negative amount is more commonly referred to as whatever is the opposite of improving.