Biostats VHM 801 Course
Fall 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment 3

This solution discusses several confidence intervals in Question 1, where one appropriately chosen interval would suffice. Also the discussions of the interpretation of independence and dependence in Questions 2 and 3, as well as the explanation of different ways of performing the successive splitting in Question 5 are more detailed than expected for a 100% mark.

1. Sensitivity and specificity of RT-PCR test

For the purpose of laying out the notation and the statistical models, consider first the results for the un-preprocessed samples. The infection status effectively splits the 95 samples into two groups corresponding to infected and uninfected fish. We assume that the infection status yields a correct classification into infected and uninfected fish. If we denote the number of samples from infected and uninfected fish by n_1 and n_2 (where $n_1 + n_2 = 95$), and furthermore the number of samples that tested RT-PCR *positive* among the infected samples by X_1 and the number of samples that tested RT-PCR *negative* among the uninfected samples by X_2, our statistical models are:

$$X_1 \sim \text{Bin}(n_1, p_1), \quad \text{and} \quad X_2 \sim \text{Bin}(n_2, p_2),$$

where p_1 is the sensitivity (Se) and p_2 is the specificity (Sp) of the test. It seems reasonable to assume binomial settings for the two samples. We do not have any information to cast doubt on the independence and homogeneity (same probability) of the test results. The models for the preprocessed samples are entirely similar. The *marginal* tables below give the data for the two tests (aggregated across the values of the other test).

		Infection status	
un-preprocessed		0	1
test	0	43	9
result	1	4	39
	Total	47	48

		Infection status	
preprocessed		0	1
test	0	45	7
result	1	2	41
	Total	47	48

From the tables we can estimate the sensitivity and specificity of the tests and compute 95% confidence intervals (CIs) within each column. The conditions for the classical (normal approximation) CI will be nowhere met, because none of the samples split according to infection status have at least 15 positives and 15 negatives for either test. Our two options are therefore the "plus four" CI (based on the normal approximation after adding 2 failures and 2 successes; the sample sizes all easily exceed 10) and the "exact" (Clopper-Pearson) CI based on the binomial distribution and computed by software. As an example of the calculations, the "plus four" 95% CI for un-preprocessed sensitivity is computed as:

$$\text{plus four} \quad : \quad 41/52 \pm 1.96\sqrt{(41/52)(11/52)/52} = 0.788 \pm 0.111 = (0.677, 0.899).$$

Test	Parameter	Estimate	Plus four 95% CI	Exact 95% CI
un-preprocessed	specificity	43/47=0.915	(0.794, 0.971)	(0.796, 0.976)
un-preprocessed	sensitivity	39/48=0.813	(0.677, 0.899)	(0.674, 0.911)
preprocessed	specificity	45/47=0.957	(0.848, 0.995)	(0.855, 0.995)
preprocessed	sensitivity	41/48=0.854	(0.724, 0.930)	(0.722, 0.939)

The estimates of both Se and Sp are higher for the preprocessed than for the un-preprocessed samples, suggesting that the preprocessing was beneficial. The confidence intervals are fairly wide, reflecting the moderate sample size. The "exact" CIs are mostly wider than those obtained by the "plus four" method, as we would expect from their conservative coverage. In answer to the question about computing a two-sample z-test for comparing the sensitivities and specificities of the two tests: this procedure is *not* valid because the samples are obtained on the same fish and therefore matched or paired (*not independent*). The corresponding Pearson chi-square test (with $X^2 = z^2$) is invalid as well, as are direct comparisons of the confidence intervals (by the rules for "2 out of 3 situations"), for the same reason. The correct approach is McNemar's test, which as explained in Lecture 7 can be viewed as a sign test for binary data. This method is however not part of the course syllabus.

2. Independence between tests

In order to assess the (in)dependence of the results of two tests on the same samples, we crosstabulate the outcomes as shown in the table below. For this question, we aggregate numbers across the infection status. Both of the tests are *response* variables (their outcome is not known prior to the study), so the appropriate model is a multinomial distribution on the 4 cells defined by all possible outcomes by the two tests. The multinomial $n = 95$ (the total number of fish), and our interest in the hypothesis (H_0) of independence between the classification according to the RT-PCR results for un-preprocessed and preprocessed samples. The table also gives the expected values under the null hypothesis.

Count (expected value)		preprocessed result 0	preprocessed result 1	Total
un-preprocessed result	0	45 (28.5)	7 (23.5)	52
	1	7 (23.5)	36 (19.5)	43
Total		52	43	95

The Pearson chi-square test statistic is $X^2 = 46.9$ with 1 degrees of freedom, which is very significant in a $\chi^2(1)$-distribution. Thus, there is an overwhelming evidence of dependence between the test results. All expected counts are above 5 so the condition for the Pearson chi-square reference distribution of X^2 is met. By comparing the observed and expected values it is seen that the data have far too many outcomes with both tests either positive or negative, and too few outcomes where the tests show disagreement. If both tests were able to distinguish, not perfectly but at least to some extent, between clinically negative and positive fish, this is exactly what we would expect. Independence between two tests in a mixed population (containing both truly positive and negative subjects) would mean that at least one of the tests does not give any information about whether a subject is true positive or negative. In other words, that at least one of the tests does not work at all. In practice, this hypothesis is rarely of interest. As the test of independence of two tests in a mixed population therefore does not give any useful information, it is usually *not* computed or reported. Because the true status of samples is rarely known, it can nevertheless be a useful descriptive step to assess the presumed dependence between test outcomes.

3. Conditional independence between tests

For this question, we carry out the same analysis as in **2.**, only separately for the two subpopulations of fish that were deemed to have positive and negative infection status. The results are summarized in the tables (on the next page), which also give the Pearson chi-square statistic and associated P-value.

Infection status negative:

Count (expected value)		preprocessed result 0	preprocessed result 1	Total
un-preprocessed	0	41 (41.2)	2 (1.8)	43
result	1	4 (3.8)	0 (0.2)	4
Total		45	2	47

$X^2 = 0.19,\ \mathrm{df} = 1$
$P = 0.66$
(Fisher's exact $P = 1$)

Infection status positive:

Count (expected value)		preprocessed result 0	preprocessed result 1	Total
un-preprocessed	0	4 (1.3)	5 (7.7)	9
result	1	3 (5.7)	36 (33.3)	39
Total		7	41	48

$X^2 = 7.93,\ \mathrm{df} = 1$
$P = 0.005$
(Fisher's exact $P = 0.017$)

In neither of the tables, the conditions for using the Pearson chi-square test are met; therefore, the P-value for Fisher's exact test with a two-sided alternative is also included. Among the uninfected fish, there is no indication whatsoever of a dependence between the tests – the results correspond almost perfectly to what would be expected from independent tests. Because both tests show high specificity, there are only few samples with positive results, and the data are not very powerful to show any dependence. It can in fact be shown mathematically that when one of the tests is perfect, no dependence is possible. Therefore our lack of evidence against a dependence is not too surprising.

In the population of infected fish, both the Pearson chi-square test and Fisher's exact test give significance, and the latter is the more trustworthy of the two. By comparing observed and expected counts we see that the agreement between the tests is somewhat better than expected if they were independent. Our interpretation is that the two tests to some extent detect the same positive subjects. That is indeed what we would expect because it is the RT-PCR same test, just with a preprocessing step added for one version of it. From that perspective, it is perhaps more surprising that the dependence is not stronger; for example, we may think say that 8 out 48 samples showing disagreement in the results is a bit disappointing. Finally, in answer to the question posed, when two tests are strongly, positively dependent, say in the infected subpopulation, it implies that after having performed the first test the second test will not improve detection of infected subjects much, because it will test positive on more or less the same infected subjects as the first test. Although beyond the scope of the course, it can also be said that using two strongly dependent tests in combination is less effective than combining independent tests.

4. Two-way table analysis for Study B

The two-way table is a cross-classification of 1165 preschool children according to two variables: illness during the past year and nutrition status. Here it seems that both variables were responses determined from a survey of the children. There is no indication that any of the two variables could have been (partly) controlled, even though the illness refers to a past period. The distribution of the variables likely reflects the target population (preschool children living under these conditions). Therefore, the data collection leads to a multinomial distribution on the 16 cells formed by combining all levels of the two variables (type II model). This would assume a multinomial setting for the $n = 1165$ children, each with independent classifications on the 16 cells and the same probabilities.

In the analysis of two response variables it is possible and legitimate to *condition* on the value of one of them if the primary interest is on analysing one variable as a function of the other one. This is similar

to what we do in linear regression for the purpose of predicting one response variable from another response variable. Loosely stated, we regard the values of one variable (and hence its distribution) as fixed for the purpose of the analysis. For the present data, it is not clear that there is particular interest in conditioning on illness or nutrition, so it seems most natural to keep our model II.

The multinomial model parameters are the probabilities for each of the 16 table cells which are estimated by the overall (table) sample proportions: $\hat{p}_{ij} = N_{ij}/n$. It is often of greater interest to compute conditional probabilities along either rows or columns because these will more clearly show any deviations from homogeneity. The table below shows the observed counts and the within-row/column proportions, computed by dividing the counts by the row or column totals, respectively. In formulas, $\hat{p}^r_{ij} = N_{ij}/N_{i\cdot}$ and $\hat{p}^c_{ij} = N_{ij}/N_{\cdot j}$.

Count (N_{ij}) (within-row/column prop.: $\hat{p}^r_{ij}$; $\hat{p}^c_{ij}$)		Nutritional status				Total
		Normal	I	II	III and IV	
Illness:	URI	95 (.21; .33)	143 (.32; .41)	144 (.32; .39)	70 (.15; .42)	452 (1;.39)
	Diarrhea	53 (.18; .18)	94 (.32; .27)	101 (.34; .28)	48 (.16; .29)	296 (1;.25)
	URI and diarrhea	27 (.14; .09)	60 (.32; .17)	76 (.40; .21)	27 (.14; .16)	190 (1;.16)
	None	113 (.50; .39)	48 (.21; .14)	44 (.19; .12)	22 (.10; .13)	227 (1;.19)
Total		288 (.25; 1)	345 (.30; 1)	365 (.31; 1)	167 (.14; 1)	

The table shows that among the rows (illness categories) it is very much the last category (None) that stands out with a very different within-row distribution than the three other categories. Also, the within-column distribution for Normal nutritional status appears quite different from the other nutrition categories. These patterns do not not suggest that homegeneity among rows or columns (or independence in Model II) will describe the data well.

In model II, the statistical hypothesis of interest (H_0) is that of independence between illness and nutrition classifications, and the alternative (H_a) is the opposite of this, i.e. dependence. We test the hypothesis by a X^2-statistic for which the observed and expected values are given in the table below, in addition to the Pearson chi-square cell contributions ($(N_{ij}-e_{ij})^2/e_{ij}$).

Count (N_{ij}) (expected: e_{ij}; X^2-contrib.)		Nutritional status			
		Normal	I	II	III and IV
Illness:	URI	95 (111.7; 2.51)	143 (133.9; 0.62)	144 (141.6; 0.04)	70 (64.8; 0.42)
	Diarrhea	53 (73.2; 5.56)	94 (87.7; 0.46)	101 (92.7; 0.74)	48 (42.4; 0.73)
	URI and diarrhea	27 (47.0; 8.49)	60 (56.3; 0.25)	76 (59.5; 4.56)	27 (27.2; 0.00)
	None	113 (56.1; 57.7)	48 (67.2; 5.50)	44 (71.1; 10.3)	22 (32.5; 3.41)

The Pearson test statistic equals $X^2 = 101.29$, which in a χ^2-distribution with df $= (4-1)\cdot(4-1) = 9$ gives $P < 0.0005$ (Minitab). The test statistic is very strongly significant, and there is clear evidence against the null hypothesis of independence between illness and nutritional status. All expected counts (e_{ij}) are very large ($\gg 5$), so there is no concern about the validity of the χ^2-distribution and its P-value. The by far largest Pearson chi-square contribution is from the cell (None, Normal), but there are relatively large contributions also in other cells, mostly in the last row and first column. As further exploration is deferred to the next question, our conclusion here will just be the very strong evidence of a dependence between the two classifications.

5. Splitting of two-way table for Study B

With the strongly significant X^2-test from above, it is of interest to try to partition (split) the $4\times4-$table into smaller tables that allow for easier interpretation. We saw above the major inhomogeneities in the within-row and within-column distributions to involve the None illness status and the Normal nutritional status, respectively. Therefore the most obvious choices for a first split are to omit either the last row or the first column; both of these operations will be described below (although in practice, one will choose one of them and proceed from there).

Initial splitting by Illness = None:
Splitting by the last row, leaves us with two tables: a reduced $3\times4-$table and a collapsed $2\times4-$table.

Count (expected e_{ij})		Nutritional status			
		Normal	I	II	III and IV
Illness:	URI	95 (84.3)	143 (143.1)	144 (154.7)	70 (69.9)
	Diarrhea	53 (55.2)	94 (93.7)	101 (101.3)	48 (45.8)
	URI and diarrhea	27 (35.5)	60 (60.2)	76 (65.0)	27 (29.4)
		Pearson $X^2 = 6.35$, df $= 6$, $(P = 0.39)$			

Count (expected e_{ij})		Nutritional status			
		Normal	I	II	III and IV
Illness:	some	175 (231.9)	297 (277.8)	321 (293.9)	145 (134.5)
	None	113 (56.1)	48 (67.2)	44 (71.1)	22 (32.5)
		Pearson $X^2 = 95.53$, df $= 3$, $P < 0.0005$			

It is seen that the large X^2-value from the full table is captured almost entirely by the second (collapsed) table ($101.29 \approx 6.35 + 95.53$), and the first (reduced) table has only little indication of an association between the two variables (and no significance whatsoever for its X^2-test). This means that we can focus our further exploration on the collapsed table. Even without the X^2-contributions being shown, it is obvious from comparisons of the observed and expected counts that the first column has the strongest discrepancies. This agrees with our previous observation that both the last row and the first column were inhomogeneous from the rest. The natural next splitting step is therefore to separate out the first column from the table, leading to a reduced $2\times3-$table and a collapsed $2\times2-$table. As described in the assignment, the P-values obtained from χ^2-distributions with the reduced df (shown in parenthesis) are not strictly valid, but a conservative P-value assessment can be obtained from the original $\chi^2(9)$-distribution (not in parenthesis).

Count (expected e_{ij})		Nutritional status		
		I	II	III and IV
Illness:	some	297 (300.2)	321 (317.6)	145 (145.3)
	None	48 (44.8)	44 (47.4)	22 (21.7)
		Pearson $X^2 = 0.55$, df $= 2$, $(P = 0.76)$		

Count (e_{ij})	Nutritional status	
	Normal	I–IV
some	175 (231.9)	763 (706.1)
None	113 (56.1)	114 (170.9)
Pearson $X^2 = 95.13$, df $= 1$, $P < 0.0005$		

Once again, the split captured almost all the association in the collapsed table. Effectively we have in these three steps decomposed the original table with a very strong association into two tables with no noticeable association and one $2\times2-$table containing almost all the association. We can now interpret the original association in terms of the effects observed in these three tables. However, before turning to the interpretations, we will consider how the splitting procedure develops if one instead starts by splitting out a column.

Initial splitting by Nutrition = Normal:
Splitting by the first column, leaves us with two tables: a reduced $4 \times 3-$table and a collapsed $4 \times 2-$table.

Count (expected: e_{ij})		Nutritional status I	II	III and IV
Illness:	URI	143 (140.4)	144 (148.6)	70 (68.0)
	Diarrhea	94 (95.6)	101 (101.1)	48 (46.3)
	URI and diarrhea	60 (64.1)	76 (67.8)	27 (31.0)
	None	48 (44.8)	44 (47.4)	22 (21.7)
		Pearson $X^2 = 2.59$, df $= 6$, $(P = 0.86)$		

Count (e_{ij})	Nutritional status Normal	I–IV
URI	95 (111.7)	357 (340.3)
Diarrhea	53 (73.2)	243 (222.8)
URI and diarrhea	27 (47.0)	163 (143.0)
None	113 (56.1)	114 (170.9)
Pearson $X^2 = 98.59$, df $= 3$, $P < 0.0005$		

Also after this split, the reduced table has virtually no association, and the vast majority of the original X^2-statistic is carried over to the collapsed table. As before, we continue by further splitting this table, now by separating out the last row, which visibly has the largest discrepancies between observed and expected values.

Count (expected e_{ij})		Nutritional status Normal	I–IV
Illness:	URI	95 (84.3)	357 (367.7)
	Diarrhea	53 (55.2)	243 (240.8)
	URI & diar.	27 (35.4)	163 (154.6)
Pearson $X^2 = 4.25$, df $= 2$, $(P = 0.12)$			

Count (e_{ij})	Nutritional status Normal	I–IV
some	175 (231.9)	763 (706.1)
None	113 (56.1)	114 (170.9)
Pearson $X^2 = 95.13$, df $= 1$, $P < 0.0005$		

We end up with the same $2 \times 2-$table accounting for the vast majority of the initial X^2-value, as well as a reduced $3 \times 2-$table failing to show any significant association (although its P-value is not as large as in previous tables without any indication of association). Because the P-values cannot be interpreted strictly at the $\alpha = 0.05$ significance level anyway (unless viewed with the initial 9 degrees of freedom), it seems fair to interpret this reduced table as showing no noteworthy association.

We have therefore, by both splitting procedures, reduced the initial significance to a $2 \times 2-$table with a very strong association. This table expresses that children without illness are far more likely to have normal nutritional status ($\hat{p}^r_{21} = 113/(113+114) = 0.50$) than those with some illness ($\hat{p}^r_{11} = 175/(175+763) = 0.19$). Conversely we could also, by conditioning on columns, say that children with some nutritional deficiency are far more likely to have some illness ($\hat{p}^c_{12} = 763/(763+114) = 0.87$), compared to those with normal nutritional status ($\hat{p}^c_{11} = 175/(175+113) = 0.61$). The lack of significance in the other tables express (in order of appearance, *i-iv*) that, *i)* there is no clear association between type of illness and nutritional status (or presence of nutritional deficiency, *iv*) for children with an illness, and *ii)* there is no clear association between degree of nutritional deficiency and occurrence of illnesss (or type of illness, *iii*), among children with inadequate nutritional status. In other words, what matters is whether the child has an illness or not, and whether the child is nutritionally deficient or not, and those two events are strongly dependent, as described.