Biostatistics VHM 8010 Course
Fall 2024, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment III

The 10% version of the home assignment only included Questions 1−3. The solution includes several confidence intervals in Question 1 where one appropriately chosen interval would suffice. The discussion of the interpretation of independence and dependence in Questions 2 and 3 is also more detailed than expected for a 100% mark. The solution includes, for Question 5, an explanation of McNemar's test which was not required as part of the answers.

As discussed in the introduction to the assignment, we consider the fish sampled as representative of two populations: fish (cages) showing clinical signs corresponding to "true positives" (diseased), and fish (cages) showing no clinical signs corresponding to "true negatives" (non-diseased). We disregard the fact that several fish were sampled from each cage, which might lead to correlations in outcome between fish from the same cage. Therefore, our assumption is that of simple random (or i.i.d.) samples from the respective populations.

Question 1. Sensitivity and specificity

To compute the sensitivity and specificity, we analyze the two tests separately and also split the samples into the two populations of clinically positive and negative fish. The *marginal* tables below give the data for the two tests (aggregated across the values of the other test). Throughout binomial models $B(n, p)$ and binomial settings are assumed.

		Clinical status	
IFAT		0	1
test	0	458	30
result	1	24	129
Total		482	159

		Clinical status	
Histopath		0	1
test	0	350	43
result	1	132	116
Total		482	159

From the tables we can immediately estimate the sensitivity and specificity of the tests and compute 95% confidence intervals (CIs). The table below shows classical (normal approximation) CIs, "plus four" CIs (based on the normal approximation after adding 2 failures and 2 successes) and "exact" CIs based on the binomial distribution and computed by software. As an example of the calculations, the classical and "plus four" CIs for IFAT sensitivity are computed as:

$$\text{classical} \quad : \quad 129/159 \pm 1.96\sqrt{(129/159)(30/159)/159} = 0.811 \pm 0.061 = (0.751, 0.872),$$
$$\text{plus four} \quad : \quad 131/163 \pm 1.96\sqrt{(131/163)(32/163)/163} = 0.804 \pm 0.061 = (0.743, 0.865).$$

Test	Parameter	Estimate	Classical 95% CI	Plus four 95% CI	Exact 95% CI
IFAT	specificity	458/482=0.950	(0.931,0.970)	(0.926,0.967)	(0.927,0.968)
IFAT	sensitivity	129/159=0.811	(0.751,0.872)	(0.743,0.865)	(0.742,0.869)
Histopath	specificity	350/482=0.726	(0.686,0.766)	(0.685,0.764)	(0.684,0.766)
Histopath	sensitivity	116/159=0.730	(0.661,0.799)	(0.655,0.793)	(0.654,0.797)

Note that the "exact test" above uses the Clopper-Pearson method, the default in Minitab up to version 22 (and in Stata as well), rather than the alternative versions of Minitab version 22. For these quite large samples, the three sets of CIs are almost the same, and should all be considered as "valid" because the conditions for their use are all satisfied. The classical (normal approximation) CI without the plus four correction is *generally* considered inferior, and the intervals are slightly off those by the two other methods, except for the Histopath specificty (with large numbers of both positives and negatives). It is seen that the IFAT test has higher values of both sensitivity and specificity than the Histopathology test, and would therefore intuitively seem as a more accurate test.

Question 2. Independence between tests

In order to assess the (in)dependence of the results of two tests on the same samples, we cross-tabulate the outcomes as shown in the table below. For this question, we aggregate numbers across the clinical 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 $n = 641$ (the total number of fish), and our interest in the hypothesis (H_0) of independence between the classification according to the IFAT and Histopath test results. The table also gives the expected values under the null hypothesis.

Count (expected value)		Histopath result 0	Histopath result 1	Total
IFAT	0	344 (299.2)	144 (188.8)	488
result	1	49 (93.8)	104 (59.2)	153
	Total	393	248	641

The (Pearson) chi-square test statistic is $X^2 = 72.7$ with 1 degrees of freedom, which is very significant in a $\chi^2(1)$-distribution. Thus, there is an overwhelming evidence against independence (H_0) and in favour of dependence between the test results (H_a). All expected counts are way above 5 so the condition for the 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 often *not* computed or reported.

Question 3. Conditional independence between tests

For this question, we carry out the same analysis as in Question 2, only separately for the two subpopulations of clinically positive and negative fish. The results are summarized in the tables below which also give the chi-square statistics and associated P-values.

Clinical status negative:

Count (expected value)		Histopath result 0	Histopath result 1	Total
IFAT	0	332 (332.6)	126 (125.4)	458
result	1	18 (17.4)	6 (6.6)	246
	Total	350	132	482

$X^2 = 0.07$
df $= 1$
$P = 0.79$

Clinical status positive:

Count (expected value)		Histopath result 0	Histopath result 1	Total
IFAT	0	12 (8.1)	18 (21.9)	30
result	1	31 (34.9)	98 (94.1)	129
	Total	43	116	159

$X^2 = 3.15$
$\text{df} = 1$
$P = 0.076$

In both tables, the conditions for using the chi-square test are met. Among the clinically negative 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. In the population of clinically positive fish, the chi-square test is non-significant but close to significance ($P = 0.076$). It is therefore fair to say that there is some indication of a dependence between the tests. 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. Such a behaviour is often seen with tests that biologically work in similar ways, but this is not the case for IFAT and Histopathology. Another possible explanation is that the degree of "infectedness" (disease) plays a role for the tests to detect an infected subject. For the clinically negative fish, no indication of such an effect is seen; in practice, conditional dependence between tests is most commonly found in the infected subpopulation. 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.

Question 4. Kappa calculations

For this and the next question, the calculations are separate for the two subpopulations. However, the interest is in agreement, irrespective of whether the two tests "get it right" or wrong. We consider again first the clinically negative samples, in order to lay out the notation. Let X_{ij} denote the number of samples (out of $n = 482$) for which the IFAT test result is i and the Histopath result is j, where $i, j = 0, 1$ and $0 \sim$ a negative test result, and $1 \sim$ a positive test result. The implicit statistical model for the kappa (κ) calculations is a multinomial distribution for the entire 2×2 table (in our terminology, a model for examining independence, or model II). The expected agreement is calculated under the assumption of independence, and these values can be computed from the expected counts in the tables of Question 3. The calculations are shown in the table below, including also the data for the clinically positive samples (with the same notation and statistical model).

Statistic	Formula	Clinically negative samples	Clinically positive samples
number of obs.	n	482	159
pos. agreement	X_{11}/n	$6/482 = 0.400$	$98/159 = 0.616$
neg. agreement	X_{00}/n	$332/482 = 0.337$	$12/159 = 0.075$
obs. agreement	$(X_{11} + X_{00})/n$	$338/482 = 0.737$	$110/159 = 0.692$
exp. agreement pos.	$X_{\cdot 1}X_{1\cdot}/n^2$	$6.6/482 = 0.014$	$94.1/159 = 0.592$
exp. agreement neg.	$X_{\cdot 0}X_{0\cdot}/n^2$	$332.6/482 = 0.690$	$8.1/159 = 0.051$
exp. agreement	sum af above two	$339.2/482 = 0.704$	$102.2/159 = 0.643$
κ (kappa)	see formula in text	-0.008	0.137

The agreement between the two tests is poor for clinically negative samples and slight for clinically positive samples. These results match well the previous findings that the dependence between the two tests is at most weak for clinically positive samples, and that the two tests are essentially independent for the clinically negative samples. Our calculation of kappa did not include a standard error, so we have no associated statistical inference with these values, but the test for kappa being equal to zero is the Pearson X^2 test of Question 3.

Question 5. Proportions of positive results

An illustration of McNemar's test, as applied to the current situation, is included here, following the description on slide 7L−10. This yields an exact P-value, computed from a binomial distribution. Another version of McNemar's test as a chi-square test can be found in many sources, but it is simply inferior to the exact test and is therefore *not* recommended under any circumstances. The P-value provided by Minitab is for the exact test.

The analysis is based entirely on the disagreeing samples. In the notation from Question 4, these are X_{01} and X_{10}, the latter of which being the count of samples that tested positive for IFAT and negative for Histopath. Under the hypothesis that the probability of testing positive is the same by the two tests, the probability of a sample with a disagreement to be IFAT positive (and Histopath negative) is 0.5. This is because the probability of IFAT testing positive (and Histopath negative) must be *the same* as the probability of IFAT testing negative (and Histopath positive), and for samples with a disagreement these are the only two possibilities, and therefore the probabilities must equal 0.5. Under the hypothesis, it furthermore holds that $X_{10} \sim \text{Bin}(n_d, 0.5)$ where $n_d = X_{10}+X_{01}$ is the total number of samples with disagreement. The statistical analysis consists therefore in testing $\text{H}_0 : p = 0.5$ against $\text{H}_a : p \neq 0.5$ in a binomial model, by an exact test in the binomial distribution (using a z-test will correspond to the above-mentioned chi-square test version of McNemar's test, but there is (still) no reason to use an approximate test). The calculations are shown below for both the clinically negative samples and the clinically positive samples. The exact probabilities from the binomial distributions are only accessible using software (in Minitab, e.g. using the `Probability Distribution Plot` menu).

Statistic	Formula	Clinically negative samples	Clinically positive samples
number of obs.	n_d	$126 + 18 = 144$	$18 + 31 = 49$
count for IFAT pos.	X_{10}	18	31
proportion	X_{10}/n_d	$18/144 = 0.125$	$31/49 = 0.633$
tail probability		$\text{P}(X_{10} \leq 18) < 0.000001$	$\text{P}(X_{10} \geq 31) = 0.04272$
P-value	$2\times$ tail prob.	< 0.001	0.085

The calculations for McNemar's test show a strong evidence against the hypothesis of equal proportions of positives for the clinically negative samples, whereas there is not quite enough evidence for clinically positive samples. This means that the specificities of the two tests differ significantly, but we do not have enough evidence to say that the sensitivities differ as well. The results from Question 1 showed the IFAT test to have the highest values of both sensitivity of specificity, so we do have strong evidence to say that the IFAT test has better characteristics than the histopathology test. Finally, we saw in Questions 3 and 4 that the tests are only weakly related in the two subpopulations with known disease status. Because the tests therefore do *not* seem to make the same mistakes (by giving either a false positive or a false negative result), we can say that the tests seem to pick up different characteristics of the infected and non-infected samples.