Homework #3

Student name: *Alex Towell* (*atowell@siue.edu*)

Course: *STAT 579 - Discrete Multivariate Analysis* – Professor: *Dr. Andrew Neath*
Due date: *02/11/2021*

Question 1

A diagnostic test is used to detect Covid antibodies in test subjects. Consider a 2×2 table in which the row variable is the true status (row 1 = antibodies, row 2 = no antibodies) and the column variable is the test result (1 = positive, 2 = negative). Then $\pi_{1|1}$ is the sensitivity and $\pi_{2|2}$ is the specificity. Let ρ be the prevalence of the disease in the testing population.
The standard test for antibodies is estimated to have sensitivity $\pi_{1|1} = .850$ and specificity $\pi_{2|2} = .995$. Consider the test applied to a population with prevalence $\rho = .10$.

(a) Compute the joint probabilities for the 2×2 table.

(b) Compute PVP $:= P(A|+)$, the predictive value positive.

(c) Provide an interpretation of the result in (b), stated in the context of the problem.

Answer. Let X denote the explanatory variate (input) with a support $\{A, NA\}$ respectively for *antibodies* and *no antibodies*. Let Y denote the response variate with a support $\{+, -\}$ respectively for a positive and negative test result.

(a) The *original* contingency table is given in the form of conditional probabilities for positive or negative, e.g., $\pi_{1|1} := P(+|A)$. The *conditional* probability table is given by table 1.

		diagnostic	
		$+$	$-$
antigen	A	$\delta = \pi_{1\|1} = .850$	$1 - \delta = \pi_{2\|1} = .150$
	NA	$1 - \gamma = \pi_{1\|2} = .005$	$\gamma = \pi_{2\|2} = .995$

Table 1: Antigen-diagnostic conditional probability table

It is given that the prevalance $P(A) = \rho = .1$. Therefore, by the laws of probability, the joint probability $\pi_{11} = P(A, +) = P(+|A)P(A)$. (Note that when you see something like $P(A, +)$, this is shorthand for $P(X = A \cap Y = +)$.) Since $P(NA) = 1 - P(A) = 1 - \rho = .9$, the remaining cells in the joint probability contingency table can be derived in a similiar way,

$$\begin{aligned}
\pi_{11} &= \pi_{1|1}\rho = .850 \times .10 = .085\,,\\
\pi_{12} &= \pi_{2|1}\rho = (1 - .850) \times .10 = .015\,,\\
\pi_{21} &= \pi_{1|2}(1 - \rho) = (1 - .995) \times (1 - .10) = .0045\,,\\
\pi_{22} &= \pi_{2|2}(1 - \rho) = .995 \times (1 - .10) = .8955\,.
\end{aligned}$$

The joint probability table is given by table 2.

(b) We are interested in PVP $:= P(A|+)$. This is trivial to construct from table 2. By Bayes theorem, $P(A|+) = P(A, +)/P(+)$, where $P(+)$ is the marginal $\pi_{+1} = P(A, +) + P(NA, +) = \pi_{11} + \pi_{21}$. Thus,

$$\begin{aligned}
\text{PVP} &= \pi_{11}/(\pi_{11} + \pi_{21})\\
&= .085/(.085 + .0045) \approx .95\,,
\end{aligned}$$

		diagnostic	
		$+$	$-$
antigen	A	$\pi_{11} = .0850$	$\pi_{12} = .0150$
	NA	$\pi_{21} = .0045$	$\pi_{22} = .8955$

Table 2: antigen-diagnostic joint probability table

which may be rewritten in terms of sensitivity and specificity,

$$\text{PVP} = \frac{\delta\rho}{\delta\rho + (1-\gamma)(1-\rho)} = \frac{.850 \times .1}{.850 \times .1 + .005 \times .9} \approx .95\,.$$

(c) We estimate that there around a 95% chance that a person who tests positive for antibodies using this diagnostic test has the Covid antibodies.

On the assumption that if people who have antibodies have or at some point had Covid, then we may rephrase this as there is a 95% chance that a person who tests positive has or had Covid.

Question 2

Table 3 is cross sectional data based on the records of traffic accidents.

(a) Compute an estimate of the relative risk.

(b) Provide an interpretation of your result, stated in the context of the problem.

	injury	
safety equipment	**fatal**	**non-fatal**
none	1,601	162,527
seatbelt	510	412,368

Table 3: Cross-sectional data of traffic accidents

Answer.

(a) The relative risk (with respect to a fatality occuring given that a seat belt is either used or not used) is defined as

$$\text{RR} := \frac{\pi_{1|1}}{\pi_{1|2}}\,. \tag{1}$$

We are given cross-sectional count data in the above table, in which the safety equipment use X and the injury Y are simultaneously observed. We can use cross-sectional data to estimate the joint probability distribution of X and Y, $\{\pi_{ij}\}$. From the joint probability distribution, we can estimate the conditional distributions $\{\pi_{j|i}\}$. A point estimator of $\{\pi_{j|i}\}$ is given by $\{\hat{\pi}_{j|i} = n_{ij}/n_{i+}\}$, and therefore

$$\hat{\text{RR}} = \frac{\hat{\pi}_{1|1}}{\hat{\pi}_{1|2}} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{1601/(1601+162527)}{510/(510+412368)} \approx 7.9\,.$$

(b) We estimate that if a person is involved in a traffic accident, those who do not wear a seatbelt are 8 times more likely to experience a fatality than those who do wear a seat belt.

Question 3

Explain the difference between a prospective study and a retrospective study. What parameters can be estimated from a prospective study? What parameters can be estimated from a retrospective study?

Answer. In a prospective study, the input variable (sometimes denoted the explanatory variable) is fixed by the sampling design, so only the response variable is observed from experimentation. For instance, let X denote the input variable and Y denote the response variable. Then, we observe the conditional relation $Y = j$ given $X = i$, whose probability we denote by $\pi_{j|i}$. We cannot estimate other kinds of probabilities, like the joint probability or the marginal probabilities, but the conditional probability $\pi_{j|i}$ is the probabilistic relation of primary interest in such a study. Naturally, we can also estimate any characteristic that is a function of $\pi_{j|i}$, like relative risk, difference of proportions, and odds ratio.

In a retrospective study, the response variable is fixed by the sampling design, so only the input variable is observed from experimentation and thus we may estimate the conditional probabilities $\pi_{i|j}$ but not the conditional probability of primary interest, $\pi_{j|i}$. We may also estimate the odds ratio since by definition it is determined by the conditional probabilities in either direction, i.e., it is invariant to rows and columns. (Additionally, when $\pi_{1|i}$ is small, the odds ratio is also a reasonable estimator of the relative risk.)