Introduction to Biostatistics – Mat 2379B

Solutions to Assignment 1

MAT 2379, Introduction to biostatistics

Solution to Assignment 1

Due date: Friday September 19, 2014 at 3:00 p.m.

$Total = 100$ marks

(1) **[15 marks]** *Describe the outcome space for each of the following random experiments.*

(a) *A candy bar with a 20.4 gram label is selected at random from a production line and is weighed.*

(b) *A coin is tossed 10 times, and the sequence of heads (marked 1) and tails (marked 0) is observed.*

(c) *A rat is selected at random from a cage, and its sex is determined.*

(d) *The province of Ontario selects 6 digits at random for one of the lottery games.*

(e) *A student is selected at random from a biostatistics class, and the student's age in years is determined.*

Solutions and grading comments. Every item below is worth 3 marks.

(a) The weight of a candy bar, ideally, is 20.4 g, but in practice there will always be a small error. Choose the sample space so that every reasonable value of weight falls within it. It should be a concrete interval, reasonably wide, and one should remember that having a sample space that is too big is not a problem (some outcomes will happen with zero probability, and that's all). There is no unique answer, the TA has to determine in each case whether the answer is reasonable. E.g. the positive real line, $[0, +\infty)$, could be a solution, or something like $[10, 50]$ (as a bar is highly unlikely to go below 10 g and above 50 g, for example).

(b) $S + \{0, 1\}^{10}$, the space of all binary strings of length 10.

(c) $S = \{M, F\}$.

(d) Here the formulation of the problem is not very precise. Taken literally, it has to be interpreted as $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}^6$. Some students have guessed that I meant 6 out of 49, in which case the sample space could be $S = \{1, 2, \ldots, 49\}^6$, or even $\{0, 1, 2, \ldots, 100\}^6$, or even $\mathbb{N}^6$, where $\mathbb{N}$ is the set of all natural numbers. All those answers will be accepted as correct.

(e) Again, here there is no uniquely determined answer. The sample space should include a range of ages within which every imaginable university student, past and present, would fall. For instance, $\{10, 11, \ldots, 200\}$ would surely do, in my opinion, while, for instance, $\{17, 18, \ldots, 25\}$ is way too restrictive. The TA will have to judge in every case separately, and possibly give partial marks. □

(2) **[15 marks]** *An ecologist is studying the effects of predation on the behaviour of the common rabbit. During one hours of observation, she spots a golden eagle with probability* 0.2*, a common rabbit with probability* 0.5*, and both with probability* 0.15*.*

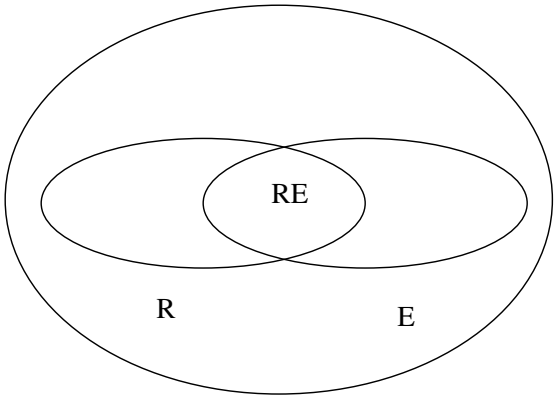
(a) *Find the probability that during one hour the ecologist sees both a golden eagle and a common rabbit.*

(b) *Find the probability that during one hour the ecologist sees neither.*

(c) *Find the conditional probability that the ecologist sees a common rabbit if she has seen an eagle. How can you interpret the result? Compare with the total probability to see a rabbit.*

Solutions and grading comments. Let us begin by mentioning that I have obviously mistyped the question (a), because, taken literally, it requires the answer 0.15. Let us therefore give to (a) just one point, and 7 points each to (b) and (c).

(b) Denote R the event that during one hour of observation the ecologist sees a rabbit, and E that she sees an eagle. We have therefore: $P(R) = 0.5$, $P(E) = 0.2$, and $P(RE) = 0.15$.



□

With the help of a Venn diagram above, and using the includion / exclusion principle we determine:

$$P(R \cup E) = P(R) + P(E) - P(AE) = 0.5 + 0.2 - 0.15 = 0.55,$$

and so

$$P((R \cup E)^c) = 1 - P(R \cup E) = 0.45.$$

The answer: 0.45.

(c) The conditional probability in question is $P(R|E)$, and it is determined from the definition:

$$P(R|E) = \frac{P(RE)}{P(E)} = \frac{0.15}{0.2} = 0.75.$$

This probability is higher than the probability just to see a rabbit which is $P(R) = 0.5$. The interpretation is quite easy: the presence of an eagle suggests that it is likely following a rabbit somewhere on the ground.

(3) **[10 marks]** *(Problem 3.2 from the recommended textbook). The official languages in Canada are English and French. Ottawa is a multicultural city whose residents have a diverse linguistic background. According to a 2006 Statistics Canada census, 59.9 % of the residents of the city of Ottawa speak only English, 1.6 % speak only French, and 1.3 % speak neither one of the official languages.*

(a) *What is the percentage of the city of Ottawa residents who speak at least one of the official languages?*

(b) *What is the probability that a randomly chosen resident of the city of Ottawa speaks both official languages?*

Solution and grading remarks. (a) The percentage is 100%-1.3%=98.7%.

(b) Let E be the event that the person speaks English and F be the event that the person speaks French. We know that $P(E \cap F') = 0.599$, $P(F \cap E') = 0.016$ and $P(E \cup F) = 0.987$. From the Venn diagram, we have:

$$P(E \cup F) = P(E \cap F') + P(F \cap E') + P(E \cap F).$$

Hence

$$\begin{aligned} P(E \cap F) &= P(E \cup F) - P(E \cap F') - P(F \cap E') \\ &= 0.987 - 0.599 - 0.016 = 0.372 \end{aligned}$$

Grading: 3 points for (a) and 7 points for (b). □

(4) **[15 marks]** *(Problem 3.5 from the recommended textbook). Consider 16 gallons of genetically modified tomatoes. Suppose that 75 % of those tomatoes have an increased resistance to pests, 50 % were engineered to have a longer shelf life, and 30 % have an increased resistance to pests and were engineered to have a longer shelf life. If one of these tomatoes is randomly selected, compute the probability that the tomato*

(a) *has an increased resistance to pests, but was not engineered to have a longer shelf life;*

(b) *has an increased resistance to pests or was engineered to have a longer shelf life;*

(c) *does not have an increased resistance to pests, but was engineered to have a longer shelf life;*

(d) *does not have an increased resistance to pests and was engineered to have a longer shelf life.*

Solution and grading remarks. When retyping the problem, I have copied (c) and pasted it into (d), but neglected to add the word "not": (d) should say "does not have an increased resistance to pests and was **not** engineered to have a longer shelf life." Some students have looked up the right formulation in the book, some did not; if taken literally, my problem should really give the same answer to (d) as to (c). This will be accepted. The right answer (below) to (d) will be accepted, too. The solution below applies to the correct formulation of (d). Suggested notes: (a), (b), (c) - 4 poionts each, (d) - 3 (no matter if the corrected form, or just repeating the answer in (c)).

Let A be the event that the tomato has an increased resistance to pests and B the event that the tomato has a longer shelf life. We know that $P(A) = 0.75$, $P(B) = 0.5$

and $P(A \cap B) = 0.3$.
(a) $P(A \cap B') = P(A) - P(A \cap B) = 0.75 - 0.3$.
(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.75 + 0.5 - 0.3 = 0.95$.
(c) $P(A' \cap B) = P(B) - P(A \cap B) = 0.5 - 0.3 = 0.2$.
(d) $P(A' \cap B') = 1 - P(A \cup B) = 1 - 0.95 = 0.05$. □

(5) **[15 marks]** *(Problem 4.2 from the recommended textbook). The nuchal translucency test is a special ultrasound scan which is widely used as a screeding test for Down's syndrome in early pregnancy. The test measures the fluid under the skin at the back of the baby's neck and can be used to determine the risk of having a baby with Down's syndrome. The following table gives the test results for a sample of 1,000 pregnant women, with the age between 35 and 40:*

	Down syndrome baby	*Normal baby*	*Total*
Test +	*3*	*50*	*53*
Test -	*2*	*945*	*947*
Total	*5*	*995*	*1,000*

Calculate
(a) *the false positive rate and false negative rate of the test,*
(b) *the sensitivity and specificity of the test,*
(c) *the positive predictive value and negative predictive value of the test.*

Solution. (a) The false positive rate is: $P(\text{test} + | \text{true}-) = 50/995 = 0.05$. The false negative rate is: $P(\text{test} - | \text{true}+) = 2/5 = 0.4$.

(b) The sensitivity is $P(\text{test} + | \text{true}+) = 3/5 = 0.6$. The specificity is $P(\text{test} - | \text{true}-) = 945/995 = 0.95$

(c) The positive predictive value is $P(\text{true} + | \text{test}+) = 3/53 = 0.06$. The positive predictive value is $P(\text{true} - | \text{test}-) = 945/947 = 0.998$.

Grading: 5 points for each one of (a),(b),(c). □

(6) **[15 marks]** *(Problem 4.3 from the recommended textbook). A screeding test which measures the level of a prostate specific antigen (PSA) is a commonly used tool for the detection of prostate cancer. Men with PSA levels greater than 10 (ng/ml) have on average a chance of 67 % of prostate cancer, whereas men whose PSA levels are between 4 and 10 have on average a 25 % chance of having prostate cancer. For those whose PSA levels are below 4 %, the average risk of developing prostate cancer is only 5 %. Suppose that 15 % of men have PSA levels greater than 10, 10 % of men have PSA levels between 4 and 10, and 75 % of men have PSA levels lower than 4.*
(a) *What is the probability that a randomly chosen man will develop prostate cancer?*
(b) *What is the probability that a randomly chosen man has a PSA level greater than 10, given that he has prostate cancer?*

Solution. (a) Let F be the event that a randomly chosen man has prostate cancer. Let A be the event that his PSA level is greater than 10, B be the event that his PSA level

is between 4 and 10, and C be the event that his PSA level is smaller than 4. We know that $P(F|A) = 0.67$, $P(F|B) = 0.25$ and $P(F|C) = 0.05$. By the total probability rule,

$$\begin{aligned} P(F) &= P(F|A)P(A) + P(F|B)P(B) + P(F|C)P(C) \\ &= (0.67)(0.15) + (0.25)(0.10) + (0.05)(0.75) \\ &= 0.1005 + 0.025 + 0.0375 = 0.163. \end{aligned}$$

(b) By the Bayes' rule,

$$P(A|F) = \frac{P(F \cap A)}{P(F)} = \frac{P(F|A)P(A)}{P(F)} = \frac{0.1005}{0.163} = 0.62.$$

Grading: 7 points for (a) and 8 for (b). □

(7) **[15 marks]** *(Problem 8.5 from the recommended textbook). According to recent estimates, only 45 % of people in Africa have access to safe drinking water, this being the major cause of many waterborne diseases. The incidence rate of waterborne diseases in communities which do not have access to safe drinking water is 88 %, whereas in communities which do have access to safe drinking water, this rate is 22 %.*

(a) *What is the incidence rate of waterborne diseases in Africa?*

(b) *A patient suffering from a waterborne disease is randomly chosen from a clinic in an African village. What is the probability that this patient did not have access to safe drinking water?*

Solution and grading remarks. Again, when retyping the problem, I have given the probability of 0.22 instead of 0.32, so the problem is *different* from that in the book. However, both answers will be accepted as correct, because some students have followed the problem as stated in the book. So, below I give *two* solutions.

Solution to the problem as stated in the book. A random person is chosen in Africa. Let A be the event that the person had access to safe drinking water, and B the event that the person suffers from waterborne disease. We know that $P(A) = 0.45$, $P(B|A) = 0.32$ and $P(B|A') = 0.88$.

(a) By the total probability rule,

$$P(B) = P(B|A)P(A) + P(B|A')P(A') = (0.32)(0.45) + (0.88)(0.55) = 0.144 + 0.484 = 0.628$$

The incidence rate of waterborne diseases in Africa is 62.8%.

(b) By the Bayes' rule,

$$P(A'|B) = \frac{P(B|A')P(A')}{P(B)} = \frac{0.484}{0.628} = 0.77$$

Solution to the problem as stated in my Assignment. A random person is chosen in Africa. Let A be the event that the person had access to safe drinking water, and B the event that the person suffers from waterborne disease. We know that $P(A) = 0.45$, $P(B|A) = 0.22$ and $P(B|A') = 0.88$.

(a′) By the total probability rule,

$$P(B) = P(B|A)P(A) + P(B|A')P(A') = (0.22)(0.45) + (0.88)(0.55) = 0.583$$

The incidence rate of waterborne diseases in Africa is 58.3%.

(b′) By the Bayes' rule,

$$P(A'|B) = \frac{P(B|A')P(A')}{P(B)} = \frac{0.484}{0.583} = 0.83$$

Grading scheme: 8 points for (a) and 7 points for (b), and again, both versions as stated above will be accepted as correct. □