

COMPLEMENTARY EXERCISES WITH PROBABILITY THEORY

EX 1

Assume a certain type of respiratory allergy usually affects 1 person out of 20, while food intolerance affect 3.5% of people. Assuming that the two events are independent, what is the probability of having both problems? and of having at least one? and of having only one?

EX 2

In a study on the Body Mass Index in a certain population it was estimated that 33% of people had normal weight, 50% were overweight and 17% were obese. In these 3 groups the probability of onset of a certain type of cardiovascular disease is respectively 1%, 3% and 6%. Assuming that the population size is 10,000 people, compute:

- a) How many obese people there were in the population;
- b) What is the probability that an individual chosen at random from the population has normal weight and will be affected by that kind of cardiovascular disease;
- c) Which is the probability that an individual chosen at random from the population will have that kind of disease and thus how many diseased people we expect to have in the population.

EX 3

In Mediterranean anaemia a parent who is a disease-free carrier (microcitemic) transmits the genetic abnormality to child with 50% probability; if the child receives the abnormality from both parents he/she will suffer from Mediterranean Anemia; if the child receives the abnormality from only one parent, he/she will be a disease-free carrier (microcitemic).

- a) What is the probability that the child of two disease-free carriers will be sick? and that he/she will be microcitemic? and that he won't inherited the abnormality?
- b) Consider now the descendents of the child of two disease-free carriers, say we call him Bob. Bob is having children with an "healthy" woman; what is the probability that Bob's child will be a disease-free carrier?

Suggestion for the approach: there are 3 possible situations, being Bob a disease-free carrier / sick / healthy, the probability of transmission depends on the situation, but all situations are possible ...

EX 4

In conception, a certain type of couple has a probability of having a child with an inborn disease equal to 15%; a diagnostic test during pregnancy is used to diagnose the event with certainty; the test is given to 8 pregnant women. What is the probability that 2 cases will be diagnosed? and the probability that no couple will be diagnosed?

EX 5

The death rate for A H1N1 influenza is 1 death out of 10,000 sick people, what is the probability that among 100 sick people there won't be any death?

EX 6

Assume weight at birth is distributed according to a Normal distribution with mean 2.8 and variance 0.16, what is the probability that a new-born child has a weight less than 1.8 kg?.

EX 7

Assuming that the prevalence of microcitemia in a region is 15%, and giving a diagnostic test for microcitemia to 100 children, if sensitivity is 80% and specificity is 98, how many positive tests should we find?

EX 8

The Skin-prick-test is often used for food allergies screening; it is a skin test that provides a result in 15 minutes, the minimum sensitivity is 90%, but specificity is low, around 60%.

What is the main problem with this test? For each sentence below, say if it is True or False.

- a) The test gives too many false negative
- b) The test gives too many false positive
- c) The test leads often to a wrong diagnosis of food allergy, while the person is not allergic
- d) It is highly probable that a person with allergy is negative

Additionally, according to some studies, food allergies are observed in 6% of children and 1.5% of adults. Say whether each of the following sentences is True or False and, if False, suggest a correction for the underlined parts.

- e) A child given the Skin-prick-test has a 6% probability of having a positive test
- f) A child given the Skin-prick-test has a 90% probability of having a positive test
- g) An allergic adult given the Skin-prick-test has positive test with probability equal to 90%.
- h) A child who gets a positive Skin-prick-test has 13% probability of being allergic
- i) The probability of being allergic when the Skin-prick-test is positive is the same in adults and children

EX 9

In a lab they have developed a new diagnostic test for celiac disease based on a blood sample. They need to estimate sensitivity and specificity by a controlled experiment. They take 20 patients with biopsy-proven celiac disease ("diseased") and 60 patients who were checked due to symptoms but were found not having the celiac disease ("not-diseased"). All 80 subjects are given the new test, results are as follows:

	Test +	Test -	Tot
Diseased	17	3	20
Not diseased	5	55	60
tot	22	58	80

Compute sensitivity and specificity and the positive predictive value, assuming that the prevalence of celiac disease is 8%. Compare the latter result with what you might obtain from the table: is there a difference, and why?

EX 10

suppose each person has a probability equal to 0.4% (4 per thousand) to develop type I diabetes during the entire life. This disease has a probable genetic component; in fact a person whose sibling has type I diabetes has 5% risk of developing type I diabetes, and this risk is 30% if the affected sibling is a syngenic twin.

- a) If John and Mary are a married couple, what is the probability that both develop type I diabetes?
- b) If John and Mary are sibling, , what is the probability that both develop type I diabetes?

SOLUTIONS

EX 1 (allergy and intolerance)

The events we are considering and the data of the exercise are:

A = Respiratory Allergy, $p(A)=1/20 =0.05$

I = Food Intolerance, $p(I)=0.035$.

A and I independent events

Let us now write the exercise questions in terms of probabilities and use the formulas to solve:

- Probability of having both: $p(A \& I)$ [intersection] = [due to independence] = $p(A) \cdot p(I)=0.05 \cdot 0.035=0.00175$ (approximately: about 2 per one thousand)
- Probability of at least one: $p(A \text{ or } I)$ [union] = $p(A)+p(I)-p(A\&I) = 0.05+0.035-0.00175=0.08325$ (about 8.3%)
- Probability of one only: we can see two approaches.

The first illustrates a **general approach, using the decomposition of one event of interest**.

In fact the event we are interested in is (A & not I) or (I & not A) – correct?

Now notice that this is the union of two incompatible events, thus its probability is:

$$p((A \& \text{not } I) + p(I \& \text{not } A) = (*)$$

The events “not I” and “not A” are respectively the complement of I and A, thus:

$$p(\text{not } I)=1-0.035=0.965 ; p(\text{not } A)=1-0.05=0.95$$

Since A and I are independent events, also their complement events are independent and e.g. A is independent of “not I”. Thus the probability of an intersection like (A & not I) is the product of the probabilities of the two events, and thus:

$$(*) = p(A) \cdot p(\text{not } I) + p(I) \cdot p(\text{not } A) = 0.05 \cdot 0.965 + 0.035 \cdot 0.95 = 0.04825 + 0.03325 = 0.0815$$

According to another and more intuitive approach: “only one” means “one or the other but not both”, and, also by using a picture, this suggests to compute the probability as:

$$pr(\text{one or the other}) - pr(\text{both}) = pr(\text{union}) - pr(\text{intersection}) = 0.08325 - 0.00175 = 0.0815$$

EX 2 (Cardiovascular disease and BMI)

The events and probabilities we know are:

N=Normal weight, $p(N)=0.33$; S=overweight, $p(S)=0.5$; O=Obese, $p(O)=0.17$. Notice that these events are not overlapping (incompatible), and all together (their union) represent the space of all events Ω ($N \cup S \cup O = \Omega$); in terms of probabilities, we have that $p(N)+p(S)+p(O)=1$.

Additionally, we consider the event M which is the onset of a specific kind of cardiovascular disease. We know from the exercise text the probability of M under different conditions – i.e. we know the following conditional probabilities:

$$p(M|N)=0.01 \quad p(M|S)=0.03 \quad p(M|O)=0.06$$

Finally, we know the population size $n=10,000$ people.

Questions:

a) number of obese people: $p(O) \cdot n = 0.17 \cdot 10,000 = 1,700$

b) $p(N \cap M) = p(N) \cdot p(M|N) = 0.33 \cdot 0.01 = 0.0033 = 3.3 \text{ ‰}$ (per one thousand)

c) $p(M)$: this is another example of use of a decomposition of an event: to approach the question we need to consider that the probability of M depends on the class of BMI (and we know such conditional probabilities) and that the composition of the population according to the class of BMI is given by the probabilities that we know too:

$$p(M) = p((M \cap N) \cup (M \cap S) \cup (M \cap O))$$

these events are incompatible (since N , S and O incompatible) $\rightarrow$

$$p(M) = p(M \cap N) + p(M \cap S) + p(M \cap O)$$

the first was computed at step b, similarly:

$$p(M) = p(M|N) \cdot p(N) + p(M|S) \cdot p(S) + p(M|O) \cdot p(O) = 0.0033 + 0.015 + 0.0102 = 0.0285 \approx 3\%$$

$$\rightarrow \text{number of people affected by disease} = p(M) \cdot n = 0.0285 \cdot 10,000 = 285$$

- Notice that we applied a sort of weighted average of the conditional probabilities of disease onset given the class of BMI, with weights given by the prevalence of the class of BMI (the sum of the weights being 1 as it should be). If we had proportionally more obese than people with normal weight, since the probability of M given O is larger than given N , overall $p(M)$ would have been larger.

- Notice that in the sum that returns $p(M)$ the larger contribution comes from the group of overweight people, despite their probability of onset of the disease is not the largest. The Risk Ratios are $RR_{OvsN}=6$, $RR_{SvsN}=3$. The mechanism is that even a low risk of disease applied to a large number of people (the overweight are the largest group) imply a large risk for the population. This kind of reasoning is key in epidemiology and public health.

EX 3 (Transmission of Mediterranean anaemia)

Let us use P for the event "disease-free carrier" (microcitemic), M for the event "affected of Mediterranean anaemia", and S for "healthy" (limited to this kind of disease). Notice that the latter is the complement to the union of the first two, and the three events are incompatible.

The other event of interest is T =transmission of the abnormality; according to the text $p(T|P)=0.50$, and, as it is intuitive, $p(T|M)=1$, $p(T|S)=0$.

a)

- child is sick if there was transmission from both parents; transmission from one parent is independent of transmission from the other parent; thus $p(\text{child is sick} | \text{both parents disease-free carriers}) = p(T|P) \cdot p(T|P) = 0.25$;

- $p(\text{child is disease-free carrier} | \text{both parents disease-free carriers}) = 2 \cdot p(T|P) \cdot (1-p(T|P)) = 0.5$. In fact we must consider the event that father transmits and mother doesn't, or the other way round - the two cases have the same probability, that's why we multiply it by 2;

- $p(\text{child is healthy} | \text{etc}) = p(\text{no transmission from father} \& \text{no transmission from mother}) = (1-p(T|P)) \cdot (1-p(T|P)) = 0.25$

b) Bob's child has a healthy mother, thus he/she will be a disease-free carrier only if he/she inherits from Bob (event: transmission); transmission can happen only if Bob is disease-free carrier (P) or sick (M). Also here we make a decomposition of the event of interest (T), and compute its probability as a sort of weighted average of the probability of transmission given P and given M :

$$p(T) = p(T|M) \cdot p(M) + p(T|P) \cdot p(P) = p(T|M) \cdot 0.25 + p(T|P) \cdot 0.50 = 1 \cdot 0.25 + 0.50 \cdot 0.50 = 0.50$$

Notice: there should be a third addendum since there is a third situation, that Bob is sane, but in that case the probability of transmission is zero; in other terms, the third addendum is equal to zero:

$$p(T|S) \cdot p(S) = 0 \cdot p(S)$$

EX 4 (Inborne disease)

This is a typical problem with a Binomial distribution: we are interested in a certain event called “success” (here, diagnosis of inborn disease) with known probability $\pi=0.15$, and we repeat the observation in $n=8$ “trials” (women, or couples). We count X =number of successes in 8 trials.

The questions ask respectively $\Pr(X=2)$ and $\Pr(X=0)$

$$p(X=2) = \binom{8}{2} 0.15^2 \cdot 0.85^6 = \frac{8 \cdot 7 \cdot 6 \cdot \dots \cdot 2 \cdot 1}{(2 \cdot 1) \cdot (6 \cdot 5 \cdot \dots \cdot 2 \cdot 1)} 0.15^2 \cdot 0.85^6 = 0.237$$

$$p(X=0) = \binom{8}{0} 0.15^0 \cdot 0.85^8 = \frac{8!}{0!(8-0)!} 1 \cdot 0.85^8 = 0.272$$

EX 5 (Death due to A H1N1 influenza)

The information we have is $p(\text{death}) = \pi = 0.0001$. We want to know the probability that X =number of deaths in $n=100$ sick people is 0. This is a typical problem with Binomial distribution, anyway the probability of success π is very small and the number of trials is very high, thus we can use the Poisson formula (the Poisson is also called the distribution “of rare events”).

The parameter for the Poisson is $\lambda = n \cdot \pi = 0.01$ (it represents the average number of deaths expected in the population with 100 individuals).

$$p(X=0) = \frac{e^{-0.01} 0.01^0}{0!} = e^{-0.01} = 0.99$$

EX 6 (Weight at birth)

This is a simple exercise of usage of the Normal curve as a theoretical tool. Parameters are: mean $\mu = 2.8$, variance $\sigma^2 = 0.16$

Thus, standard deviation is $\sigma = 0.4$.

For computing $p(X < 1.8)$ we look in the tables for a standard Normal $N(0,1)$ what is the probability that Z is < the standardized value z :

$$z = \frac{1.8 - 2.8}{0.4} = -2.5$$

Remind that the table returns $\Phi(z)$, the area under the curve from -infinite to z only for positive z ; the area we are looking for is an area in a tail and it is equal to the area from 2.5 to infinite, i.e. it is equal to $1 - \Phi(2.5)$.

On our table (*pay attention if you are using another table, not all texts report this kind of area*) crossing the row for 2.5 with the column for 0 we find $\Phi(2.5) = 0.994 \rightarrow$ the probability we want is $0.006 \approx 6\%$.

EX 7 (Diagnostic test for microcitemia)

Use P for Positive test, N for negative test, M for “disease”, which in this case is microcitemia.

$p(P | M) = \text{sensitivity} = 80\%$

$p(N | \text{not } M) = \text{specificity} = 98\%$

$p(M) = 15\%$

The number of positive tests is the sum of True Positive (Positive tests among children with microcitemia) plus False Positive (Positive tests among children without microcitemia). Children with microcitemia are expected to be 15 (15% of the sample, 15 out of 100!), the other 85 are not microcitemic. Thus:

- the expected number of True Positive is: $p(P | M) \cdot 15 = 0.8 \cdot 15 = 12$
- the expected number of False Positive is: $p(P | \text{non } M) \cdot 85 = (1-0.98) \cdot 85 = 1.7$
- in total we expect about 14 Positive tests.

EX 8 (screening for food allergies)

sensitivity = 90% = (introducing some notation) $p(TP | AL)$

specificity = 60% = $p(N | \text{not } AL)$

In the second part of the exercise the test gives the prevalence of AL which depends on age: it is 0.06 in children and 0.015 in adults

- a) The test gives too many false negative → **False**: false negative are 10% (1-sensitivity)
- b) The test gives too many false positive → **True**: false positive are 40% (1-specificity)
- c) The test leads often to a wrong diagnosis of food allergy, while the person is not allergic → **True**, it states the same as point b
- d) It is highly probable that a person with allergy is negative → **False**, it states the same as point a
- e) A child given the Skin-prick-test has a 6% probability of having a positive test → **False**. Compute this probability as:

We restrict to children now, thus $\text{pr}(AL)=0.06$.

$$\begin{aligned} \text{Pr}(TP) &= \text{Pr}(TP \& AL) \text{ or } (TP \& \text{not } AL) = \\ &= \text{Pr}(TP \& AL) + \text{Pr}(TP \& \text{not } AL) = \\ &= \text{Pr}(TP | AL) \cdot \text{Pr}(AL) + \text{Pr}(TP | \text{not } AL) \cdot \text{Pr}(\text{not } AL) = \\ &= \text{sens.} \cdot 0.06 + (1-\text{spec}) \cdot (1 - 0.06) = \\ &= 0.9 \cdot 0.06 + 0.4 \cdot 0.94 = 0.43 \end{aligned}$$

Thus the correction is 43% instead of 6%.

- f) A child given the Skin-prick-test has a 90% probability of having a positive test → **False**. The correction is "an allergic child" in place of "child" (*the following sentence suggested the correction*)
- g) An allergic adult given the Skin-prick-test has positive test with probability equal to 90%. → **True**
- h) A child who gets a positive Skin-prick-test has 13% probability of being allergic → **True**. This is the result of the calculation of the predictive value of a positive test:

We restrict to children now, thus $\text{pr}(AL)=0.06$.

$$\begin{aligned} \text{Pr}(AL | TP) &= \text{sens.} \cdot \text{prev.} / [\text{sens.} \cdot \text{prev.} + (1-\text{spec}) \cdot (1 - \text{prev})] = \\ &= 0.9 \cdot 0.06 / 0.9 \cdot 0.06 + 0.4 \cdot 0.94 = 0.126 \end{aligned}$$

- i) The probability of being allergic when the Skin-prick-test is positive is the same in adults and children → **False**, since the prevalence is different in children and adults. In adults the prevalence, i.e. the prior probability of allergy, is lower, and thus also the posterior probability will be lower. We can verify this by computing the predictive value of a positive test for adults:

This time we restrict to adults, thus $\text{pr}(AL)=0.015$.

$$\begin{aligned} \text{Pr}(AL | TP) &= \text{sens.} \cdot \text{prev.} / [\text{sens.} \cdot \text{prev.} + (1-\text{spec}) \cdot (1 - \text{prev})] = \\ &= 0.9 \cdot 0.015 / 0.9 \cdot 0.015 + 0.4 \cdot 0.985 = 0.033 \end{aligned}$$

For an adult who has a positive test the probability of being allergic is only 3.3%, compared to 12.6% in children.

EX 9 (experiment on a diagnostic test)

From the table, estimated parameters are:

$$\text{sensitivity} = 17 / 20 = 0.85$$

$$\text{specificity} = 55 / 60 = 0.92$$

For the positive predictive value of the test, using the prevalence = 0.08 and the Bayes formula, we get:

$$p(M | T+) = \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})} = \frac{0.85 \cdot 0.08}{0.85 \cdot 0.08 + (1 - 0.92) \cdot (1 - 0.08)} = 0.47$$

Thus a subject who gets a positive result with the new test has 47% probability of being celiac - which means that the biopsy will give positive result with 47% probability.

If we used the data from the table instead:

$$p(M | T+) = (\text{true positive}) / \text{positive} = \text{tp} / (\text{tp} + \text{fp}) = 17 / 22 = 0.77$$

The results are different, and the second one is **WRONG** since the second approach does not use the correct value of the prevalence, but the value obtained from the table, which is wrong since the relative numbers of patients with disease 20/80 does not represent the population, but was chosen by the investigators for their convenience (it could be by chance that true prevalence = estimated prevalence, or it could be done on purpose; generally, it is not, and often the ratio between diseased and not diseased is 1:2 or 1:3).

So with wrong prevalence higher than true (25%) we get wrong predictive value, 77%, higher than true.

To complete the illustration, **let's see an example where the estimated prevalence is correct.**

A large observational study (see courses in epidemiology) considered a population of 100 people exposed to a chemical material that could induce a certain blood cancer. All people were given a diagnostic screening test to assess whether they had absorbed large quantities of the chemical material, being thus classified into "at risk" if positive and "not at risk" if the test was negative. Among the 100 people, 22 had that type of cancer. The crosstab was:

	Test +	Test -	Tot
Diseased	13	9	22
Not diseased	25	53	78
tot	38	62	100

Estimates:

$$\text{sensitivity} = 13 / 22 = 0.59$$

$$\text{specificity} = 53 / 78 = 0.68$$

$$\text{estimated prevalence: } 22/100 = 22\%$$

This estimate correctly returns the overall prevalence in the population. Thus we can use both procedures to compute the positive predictive value of the test:

- either the Bayes formula:

$$p(M | T+) = \frac{sens \cdot prev}{sens \cdot prev + (1 - spec) \cdot (1 - prev)} = \frac{0.59 \cdot 0.22}{0.59 \cdot 0.22 + (1 - 0.68) \cdot (1 - 0.22)} = 0.34$$

- or the (faster) formula that looks the percentage in the table:

$$(\text{true positive}) / \text{positive} = 13 / 38 = 0.34$$

Let's see also the predictive value of a negative test, what is the probability that a subject with negative screening test is diseased? =

$$p(M | T-) = \frac{(1 - sens) \cdot prev}{(1 - sens) \cdot prev + spec \cdot (1 - prev)} = \frac{(1 - 0.59) \cdot 0.22}{(1 - 0.59) \cdot 0.22 + 0.68 \cdot (1 - 0.22)} = 0.14$$

EX 10 (type I diabetes)

The probability of disease p(D) is in general 0.004. It anyway depends on what we know:

$$p(D | \text{a sibling is diseased}) = 0.05$$

$$p(D | \text{a syngenic twin is diseased}) = 0.30$$

Let us use the symbols DJ and DM respectively for "John is diseased" and "Mary is diseased", we need p(DJ & DM) under two different situations.

a) married

the two events are independent, thus:

$$= p(DJ) \cdot p(DM) \text{ with } p(DJ) = p(DM) = p(D) = 0.004$$

$$\text{thus: } \Pr(DJ \& DM) = 0.004 \cdot 0.004 = 0.000016$$

b) sibling (notice: being a male and a female they cannot be syngenic twins)

this time the two events are not independent because for example $p(DJ | DM) = 0.05 \neq 0.004$

$$p(DJ \& DM) = p(DM) \cdot p(DJ | DM) = 0.004 \cdot 0.05 = 0.0002$$