Problem set 1: Probabilities and medical tests

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Useful probability distributions, mean and variance

1.1 Bernoulli distribution

Consider a random variable X that can take two values, 0 and 1. In other words, the set of values that can be taken by X is $\Omega = \{0, 1\}$. Assume that the probability $P(X = 1)$ that it takes the value 1 is equal to p. This is called a Bernoulli random variable.

a) Express $P(X = 0)$ as a function of p.

b) Express the mean $\langle x \rangle$ of X as a function of p. We recall its definition:

$$\langle x \rangle = \sum_{x \in \Omega} x P(X = x) \,. \tag{1}$$

c) Express the variance of X as a function of p. We recall the definition of the variance:

$$\mathrm{var}(X) = \left\langle (x - \langle x \rangle)^2 \right\rangle = \langle x^2 \rangle - (\langle x \rangle)^2 \,. \tag{2}$$

Show that the last two expressions are equivalent.

d) For what value of p is the variance of X maximal? What distribution does this correspond to?

e) Consider a coin, flip (toss) it: it can land on heads (0) or tails (1). If we denote by X the random variable corresponding to the result of the coin flip, exactly two possibilities exist, either $X = 0$ or $X = 1$. For a coin, what does $p = 1/2$ mean? And $p \neq 1/2$?

1.2 Binomial distribution

a) Consider a coin, modeled as a Bernoulli random variable (see previous section) with $p = 1/2$. If it is flipped twice, what is the probability to obtain "heads" twice? "Heads" once and "tails" once?

b) Now consider a coin, still modeled as a Bernoulli random variable, but without assuming that $p = 1/2$. Let us denote by M the random variable corresponding to the number of times that one obtains "tails" if the coin is flipped N times. What values can this random variable take?

c) What is the probability $P(M = m)$, noted $P(m)$ for brevity, to obtain exactly m times "tails" and $N - m$ times "heads"? You can first consider examples such as $m = 0$ and $m = 1$. Show that

$$P(m) = \frac{N \times (N-1) \times (N-2) \times \cdots \times (N-m+1)}{m!} p^m (1-p)^{N-m} = \binom{N}{m} p^m (1-p)^{N-m} \,, \tag{3}$$

where we have used the notation

$$\binom{N}{m} = \frac{N!}{m!(N-m)!} \,, \text{ with } N! = N \times (N-1) \times (N-2) \times \cdots \times 3 \times 2 \times 1 \,. \tag{4}$$

The probability distribution in Eq. 3 is called the binomial distribution.

d) Calculate the probability to obtain 6 times "tails" out of $N = 10$ flips if $p = 1/2$. Using Python, draw 10000 samples in a binomial distribution with $N = 10$ and $p = 1/2$ (this means 10000 values of m, i.e. this corresponds to doing 10000 times the experiment of flipping the coin 10 times), and plot the histogram corresponding to this data.

e) Calculate the probability to obtain 600 times "tails" out of $N = 1000$ flips if $p = 1/2$. Compare it to the result of the previous question. Using Python, draw 10000 samples in a binomial distribution with $N = 1000$ and $p = 1/2$, and plot the histogram corresponding to this data. Compare it to the histogram from the previous question. Imagine that you flip a coin 1000 times and obtain 600 times "tails": what do you think about this coin?

f) Express the mean of M as a function of p and N. This can be done using a simple reasoning or a calculation (note that the calculation is a bit formal). Test your result in Python by calculating the mean of the samples you produced in the two questions above, and of another one with $p \neq 1/2$.

1.3 Poisson distribution

In the case of rare events, specifically if $p \ll 1$ while $N \gg 1$ such that $\lambda = Np$ is finite, the binomial distribution Eq. 3 simplifies to

$$P(m) = \frac{\lambda^m e^{-\lambda}}{m!}\,. \tag{5}$$

This distribution is called the Poisson distribution. Consider a random variable M following this distribution, i.e. such that $P(M = m) = P(m)$ is given by Eq. 5. Here M can take all integer values m from 0 to ∞.

a) Check that this probability distribution is normalized.

b) Based on the previous part about the binomial distribution, what value do you expect for the value of the mean of M? Also calculate the mean of M directly, and check that it is consistent with this expectation.

c) Using Python, draw 10000 samples in a Poisson distribution with $\lambda = 5$. Compute the mean and variance of this sample, and check that your result for the mean is consistent with the one obtained above. Plot the histogram corresponding to this data.

d) Using Python, draw 10000 samples in a binomial distribution with $N = 1000$ and $p = 5/1000$. Compute the mean and variance of this sample, and plot the histogram corresponding to this data. Compare to the results from the previous question, in light of the introduction to section 1.3 above.

2 HIV evolution and treatment

Retroviruses such as HIV have RNA genomes. When infecting host cells (T cells in the case of HIV), they use enzymes (called reverse transcriptases) to reverse-transcribe their RNA genomes into DNA, which is then integrated into the host cell genome and replicated along with it. The genome of the HIV virus comprises about 10^4 base pairs. Reverse transcriptases are quite error-prone. The probability of errors in reverse transcribing the HIV genome is about one error for every 3×10^4 nucleotide. We will assume that each error replaces a DNA base (A, T, C or G) by one of the three other bases, chosen uniformly at random. Each time an HIV virus infects a T cell, the reverse transcription step creates opportunities for such errors, which are then inherited by the offspring viruses. The total population of infected T cells in an infected patient's blood is of order 10^7 cells. For simplicity, we will assume that each infected T cell is infected by a wild-type/standard HIV virus, so that mutations do not accumulate (in fact, they can accumulate, but if most mutants are deleterious then it is nevertheless likely that a new infected T cell is infected by a wild-type virus).

Consider a drug that is perfectly effective at stopping replication of wild-type/standard HIV (an ideal drug!). Assume that one particular mutation makes HIV resistant to it, so that it no longer stops replication. For instance, imagine that the standard base of DNA at site number 220 in the HIV genome is A, but if it is C, then HIV becomes resistant to the drug.

a) Consider an HIV virus that infects a T cell. What is the probability that at least one mutation occurs at the reverse transcription step?

b) Find the probability that the particular mutation that confers resistance to the drug will occur upon a given T cell infection by an HIV virus.

c) Estimate the number of T cells that possess this particular mutation in an infected patient. What should happen if the patient is treated with this drug?

d) Now imagine that there are two drugs with similar properties, but such that the particular mutation giving resistance to each of them is different. Find the probability that both particular mutations that confer resistance to the two drugs will occur upon a given T cell infection by an HIV virus.

e) Estimate the number of T cells that possess these two particular mutations in an infected patient. What should happen if the patient is treated with these two drugs simultaneously? Comment on the interest of bitherapy and tritherapy.

3 Cancer screening

The hemoccult test can be used to detect colorectal cancer. Imagine that mass screening is conducted with this test in a segment of the population where 0.3% of individuals are expected to have this disease. People who are sick have a 50% chance to test positive. Among those who do not have the disease, 3% nevertheless test positive.

a) A given patient tests positive. Calculate the probability that this person is sick. Comment on the result.

b) A given patient tests negative. Calculate the probability that this person is sick.

4 COVID-19 testing

A specific brand of rapid self-tests ("lateral flow devices") for COVID-19 is estimated to have about 78.7% sensitivity and 99.7% specificity [1]. Recall that sensitivity is the proportion of patients who test positive among sick patients, while specificity is the proportion of patients who test negative among healthy patients.

a) An early recommendation to doctors interpreting COVID-19 test results reads: "A positive (...) test for COVID-19 test has more weight than a negative test" [2]. Explain this recommendation in light of the specificity and sensitivity of the test.

b) Express the sensitivity and specificity in terms of conditional probabilities.

c) Express the probability $P(s|p)$ that a person who tests positive is really sick, as a function of sensitivity, specificity and of the probability $P(s)$ that the patient is sick *a priori* (i.e. before the result of the test is known). For instance if we don't know anything about the patient, $P(s)$ may be taken as the proportion of sick people in the population.

d) Express the probability $P(s|n)$ that a person who tests negative is in fact sick, as a function of sensitivity, specificity and of the probability $P(s)$ that the patient is sick *a priori*.

e) In Python, using the expressions obtained at the two previous questions, plot $P(s|p)$ and $P(s|n)$ as a function of the probability $P(s)$ that the patient is sick *a priori*, for $P(s)$ ranging between 0 and 1. Show them on the same plot. Comment on the curves obtained, in particular, on how much each of them deviates from the diagonal $y = x$.

f) A person with no symptoms and no known exposure risks to COVID-19 tests positive. Calculate the probability $P(s|p)$ that this person is sick, in a population where 300/100,000 individuals are expected to have the disease. Comment on the result.

g) Same question as above if the incidence rate is 4,000/100,000.

h) A medical doctor with high exposure risks to COVID-19 develops characteristic symptoms but tests negative. Calculate the probability $P(s|n)$ that this doctor has COVID-19 nevertheless, if their probability of being sick *a priori* is estimated to be 90%. Imagine that after two days the symptoms get milder, so that this doctor feels well enough to return to work, in contact with immuno-compromised patients: do you think this would be a good idea?

5 Doping

A cyclist tested positive for illegal steroid drugs in the 2006 Tour de France. The lab said that the test was highly unlikely to be positive unless the subject had taken illegal steroid drugs, and the cyclist was accused of doping. This cyclist was tested 8 times during the race, and a total of 126 tests were made on all cyclists.

a) Assume that the cyclist is innocent, but the false-positive rate of the test is 2%, meaning that $FP/(FP + TN) = 0.02$ where FP is the number of false positive people (innocent but testing positive) and TN is the number of true negative people (innocent and testing negative). What is the probability that at least one of the 8 tests is positive?

b) In the same case, what is the probability that at least 2 of the 8 tests is positive? Comment on this result.

c) For the same test, assume that all cyclists are innocent. What is the probability that at least one of them tests positive at least once? Comment on this result.

d) In fact, what we would like to know is the probability that a cyclist is guilty given that the test is positive. Express this probability as a function of the false-positive rate defined above and of the false-negative rate $FN/(FN + TP)$ where FN is the number of false negative people (guilty but testing negative) and TP is the number of true positive people (guilty and testing positive). Simplify this expression assuming that the false-negative rate is zero. What additional information do we need to compute the probability that a cyclist is guilty given that the test is positive? Why is it difficult to know it in practice?

6 Additional problem: Non-invasive prenatal testing

This problem was previously given as part of the final exam of this class.

Non-invasive prenatal testing (NIPT) aims to estimate the risk that a baby will be born with certain genetic anomalies. Our blood contains small fragments of DNA that are freely circulating, i.e. not within cells. These DNA fragments arise when cells die. During pregnancy, some of these DNA fragments come from cells from the placenta, whose DNA is usually identical to that of the embryo. Thus, analyzing DNA fragments in the blood of a pregnant patient provides an opportunity for early detection of certain genetic anomalies in the embryo. It is a non-invasive procedure, as it just requires a blood test.

NIPT is most often used to look for the presence of an extra or missing copy of a chromosome. The most common case is Down syndrome or trisomy 21, caused by an extra chromosome 21. NIPT is such that an embryo with trisomy 21 will yield a positive test in 99.3% of cases, while an embryo without trisomy 21 will yield a negative test in 99.9% of cases [3, 4].

a) Recall that sensitivity α is the proportion of patients who test positive among sick patients, while specificity β is the proportion of patients who test negative among healthy patients (here, for simplicity, “healthy patients” means those pregnant with embryos without trisomy 21). Express the sensitivity α and specificity β in terms of conditional probabilities, and give their values for NIPT for trisomy 21.

b) Suppose that a NIPT is positive. Express the probability that the embryo actually has trisomy 21 as a function of α, β, and the overall probability $P(s)$ of trisomy 21 in the population. Show your intermediate calculations.

c) Consider a population where the probability of trisomy 21 is $P(s) = 1/200$: what is the numerical value of the probability that an embryo with positive NIPT actually has trisomy 21? Comment on your result.

d) Now suppose that a NIPT is negative. Express the probability that the embryo has trisomy 21 as a function of α, β and $P(s)$, showing your intermediate calculations.

e) Calculate the value of the probability that the embryo with negative NIPT has trisomy 21 for $P(s) = 1/200$. Comment.

f) Imagine that you are a medical doctor. Depending on the NIPT result, would you recommend that your patient takes an additional test (amniocentesis), which is extremely accurate, but invasive, and has some risks of complications?

g) Another possibility would be to ask your patient to take a second NIPT. Why could this be useful?

h) In practice, the main source of a false positive NIPT is the possible existence of genetic anomalies in the placenta or in some maternal tissues that do not affect the embryo. Given this, what do you think of the usefulness of recommending a second NIPT for a patient who has already taken one?

i) There are many other genetic anomalies, but they are generally much rarer than trisomy 21. For instance, Prader-Willi syndrome has an overall risk in the population of $P(s) = 5.0 \times 10^{-5}$. NIPT can also be used for Prader-Willi syndrome, and the sensitivity and specificity are similar in this case as for trisomy 21. Taking the same sensitivity and specificity values as above, calculate the value of the probability that an embryo with positive NIPT for Prader-Willi syndrome actually has Prader-Willi syndrome. Comment on the result.

j) What do you think of the usefulness of NIPT for anomalies like Prader-Willi syndrome?

k) Imagine that you are working for a laboratory that performs NIPT. Is it scientifically valid to say that these tests are "correct in more than 99% of the cases"? Is it a statement that you would recommend to include as the main advertisement of the tests (including those for Prader-Willi syndrome), destined to patients? Explain your answers, using the different conditional probabilities considered above, and comment.

7 Additional problem: Searching for fluorescent cells

Suppose that you are looking for cells tagged by expressing a fluorescent protein in a liquid sample. You spread a drop of the sample on a microscope slide marked with a tiny grid containing N boxes. Under the microscope, you can then look at whether there is any fluorescent cell in each box. Assume that a particular sample has a total of M tagged cells.

a) What is the probability that no box on the grid has more than one tagged cell?

b) What is the probability that at least one box on the grid contains more than one of these M cells?

c) Calculate the latter probability for $N = 400$ and $M = 20$. You can use Python for this.

d) Consider a class of 20 students: what is the probability that at least two students have the same birthday?

References

[1] D. A. Mistry, J. Y. Wang, M. E. Moeser, T. Starkey, and L. Y. W. Lee. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-CoV-2. *BMC Infect Dis*, 21(1):828, Aug 2021.

[2] J. Watson, P. F. Whiting, and J. E. Brush. Interpreting a covid-19 test result. *BMJ*, 369:m1808, May 2020.

[3] T. Liehr. Non-invasive prenatal testing, what patients do not learn, may be due to lack of specialist genetic training by gynecologists and obstetricians? *Front. Genet.*, 12:682980, 2021.

[4] K. Bennett. These prenatal tests are usually wrong when warning of rare disorders. *New York Times*, 2022.