Solutions 1: Probabilities and medical tests

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Useful probability distributions, mean and variance

1.1 Bernoulli distribution

a) Probability distributions are normalized: $\sum_{x\in\Omega} P(X=x) = 1$. Here, X can only take values 0 and 1, i.e. $\Omega = \{0,1\}$, and thus, $P(X=0) + P(X=1) = 1$. Thus $P(X=0) = 1-p$.

b) The mean of X reads:

$$\langle x\rangle = \sum_{x\in\Omega} xP(X=x) = 1\times P(X=1) + 0\times P(X=0) = p\,. \tag{1}$$

c) Let us use

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

Here we have:

$$\langle x^2\rangle = \sum_{x\in\Omega} x^2P(X=x) = 1\times P(X=1) + 0\times P(X=0) = p\,. \tag{3}$$

and thus

$$\mathrm{var}(X) = p - p^2 = p(1-p)\,. \tag{4}$$

To prove that the two expressions of the variance are equivalent, let us start from the first one:

$$\mathrm{var}(X) = \left\langle \left(x - \langle x\rangle\right)^2\right\rangle = \left\langle x^2 - 2x\langle x\rangle + \langle x\rangle^2\right\rangle = \langle x^2\rangle - 2\left(\langle x\rangle\right)^2 + \left(\langle x\rangle\right)^2 = \langle x^2\rangle - \left(\langle x\rangle\right)^2\,. \tag{5}$$

We obtain the second one. Note that we have employed the fact that the mean of a sum of terms is the sum of the means of these terms, and that the mean of αX where α is a constant and X a random variable, is equal to α times the mean of X. All these properties of the mean can be proved by using the definition of the mean

$$\langle x\rangle = \sum_{x\in\Omega} xP(X=x)\,. \tag{6}$$

d) We need to study the variance given by Eq. 4 as a function of p. In other words we need to study the function $F: p\mapsto p(1-p)$. We have $F'(p) = 1-2p$, which is zero if $p=1/2$, positive if $p<1/2$ and negative if $p>1/2$. Therefore, F and thus the variance of X, is maximal for $p=1/2$. In this case where $p=1/2$, we have $P(X=0) = P(X=1) = 1/2$: the two possibilities have the same probability. This corresponds to the uniform distribution.

e) For a coin, $p=1/2$ means that the coin is fair: the two outcomes, heads or tails, are equally likely. Conversely, if $p\neq 1/2$, then the coin is biased.

1.2 Binomial distribution

a) If we flip a fair coin twice, there are four possibilities that are all equally likely:

- We obtain "heads" the first time and "heads" the second time,
- We obtain "heads" the first time and "tails" the second time,
- We obtain "tails" the first time and "heads" the second time,
- We obtain "tails" the first time and "tails" the second time.

Thus, the probability to obtain "heads" twice is 1/4 (first possibility above), and the probability to obtain "heads" once and "tails" once is 1/2 (second and third possibilities above).

b) If we consider a coin and flip it N times, the number of times that one obtains "tails" can take all integer values from 0 to N.

c) First consider the probability $P(0)$ that we never obtain "tails" out of N flips: it is $P(0) = (1-p)^N$ as each of the N flips needs to give "heads", which occurs with probability $1-p$.

The probability to obtain exactly one time "tails" is $P(1) = Np(1-p)^{N-1}$ because it is equal to the probability $p(1-p)^{N-1}$ that the first flip gives "tails" and all others "heads", plus the probability $p(1-p)^{N-1}$ that the second gives "tails" and all others "heads" etc.: there are N different ways to obtain this result, and thus, summing over them yields $P(1) = Np(1-p)^{N-1}$.

More generally, to compute $P(m)$, we need to count the number of ways we can obtain m times "tails" and $N-m$ times "heads" among N flips, which is

$$\binom{N}{m} = \frac{N!}{m!(N-m)!}\,, \text{ with } N! = N \times (N-1) \times (N-2) \times \cdots \times 3 \times 2 \times 1\,, \tag{7}$$

and each of these different ways has a probability $p^m(1-p)^{N-m}$, yielding

$$P(m) = \frac{N \times (N-1) \times (N-2) \times \cdots \times (N-m+1)}{m!} p^m (1-p)^{N-m} = \binom{N}{m} p^m (1-p)^{N-m}\,. \tag{8}$$

The probability distribution in Eq. 8 is called the binomial distribution.

d) To calculate the probability to obtain 6 times "tails" out of $N = 10$ flips if $p = 1/2$, we use Eq. 8 with $N = 10$, $m = 6$ and $p = 1/2$, which gives $P(6) = 0.21$.

e) To calculate the probability to obtain 600 times "tails" out of $N = 1000$ flips if $p = 1/2$, we use Eq. 8 with $N = 1000$, $m = 600$ and $p = 1/2$, which gives $P(600) = 4.6 \times 10^{-11}$.

This probability is extremely small and really much smaller than in the previous question. Even though the proportion of "tails" is the same, such a deviation from the mean is far less likely in a large sample (large number of flips here) than in a small one.

If we flip a coin 1000 times and obtain 600 times "tails", it becomes credible to assume that this coin may be biased.

f) Since the probability to get "tails" at each flip is p, for N flips the mean number of "tails" should be Np. Let us prove this by a calculation. The mean of M reads:

$$\langle m \rangle = \sum_{m=0}^{N} m P(M=m) = \sum_{m=0}^{N} m \binom{N}{m} p^m (1-p)^{N-m}\,. \tag{9}$$

Consider the function

$$F : (p,q) \mapsto \sum_{m=0}^{N} \binom{N}{m} p^m q^{N-m} = (p+q)^N\,, \tag{10}$$

where we have used the binomial theorem. Differentiating it with respect to p gives:

$$\frac{\partial F}{\partial p} = \sum_{m=0}^{N} m \binom{N}{m} p^{m-1} q^{N-m} = N(p+q)^{N-1}\,. \tag{11}$$

In particular, for $q = 1 - p$, this yields

$$\sum_{m=0}^{N} m \binom{N}{m} p^{m-1}(1-p)^{N-m} = N\,. \tag{12}$$

Combining Eq. 9 with Eq. 12 yields

$$\langle m \rangle = \sum_{m=0}^{N} m \binom{N}{m} p^{m}(1-p)^{N-m} = Np\,. \tag{13}$$

1.3 Poisson distribution

a) We have

$$\sum_{m=0}^{\infty} P(m) = \sum_{m=0}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!} = e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!}\,. \tag{14}$$

But the series expansion of the exponential function reads

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}\,. \tag{15}$$

Thus

$$\sum_{m=0}^{\infty} P(m) = e^{-\lambda} e^{\lambda} = 1\,. \tag{16}$$

b) Based on the previous part about the binomial distribution, we expect $\langle m \rangle = \lambda$.

The mean of M reads:

$$\langle m \rangle = \sum_{m=0}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} \frac{\lambda^m e^{-\lambda}}{(m-1)!} = \lambda e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!}\,. \tag{17}$$

Introducing $k = m - 1$, we obtain

$$\langle m \rangle = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda\,, \tag{18}$$

where we have again used the series expansion of the exponential function

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}\,. \tag{19}$$

c) The mean and variance found in Python are close to 5. In fact the mean and the variance of the Poisson distribution are both equal to λ.

d) In Python, with these parameters, we find that the mean and variance are again close to 5. In the case of rare events, specifically if $p \ll 1$ while $N \gg 1$ such that $\lambda = Np$ is finite, the binomial distribution simplifies to the Poisson distribution: here this is what we can see for $\lambda = 5$.

2 HIV evolution and treatment

a) To calculate the probability p_1 that at least one mutation occurs at the reverse transcription step, we first calculate the probability p_0 that no mutation occurs at this step. As all sites are independent and have the same probability for a mutation to occur, namely $p_m = 1/(3 \times 10^4)$, we can use the binomial distribution, and $p_0 = (1 - p_m)^M$ where M is the number of sites, namely 10^4. Thus, $p_1 = 1 - p_0 = 1 - (1 - p_m)^M = 0.28$, which is quite high.

b) The probability p that the particular mutation that confers resistance to the drug will occur upon a given T cell infection by an HIV virus is the probability for any mutation to occur at the specified site (e.g. at site 220), namely $p_m = 1/(3 \times 10^4)$, times the probability that the mutation (given that it occurs) is the specified one (e.g. yields C), namely 1/3. Thus, it reads

$$p = \frac{1}{3 \times 10^4} \times \frac{1}{3} = 1.1 \times 10^{-5} \,. \tag{20}$$

c) To estimate the average number n of T cells that possess this particular mutation in an infected patient, we multiply p by the number N of infected T cells in the patient's blood:

$$n = p \times N = 1.1 \times 10^{-5} \times 10^7 = 1.1 \times 10^2 \,. \tag{21}$$

If the patient is treated with this drug, the treatment is expected to fail because the patient is extremely likely to have preexisting resistant HIV viruses in their blood. Indeed, on average, they are expected to have about 110 such viruses in their blood.

d) The two mutations giving resistance to each of the drugs are different and can occur independently from one another. Hence the probability that both of these particular mutations occur upon a given T cell infection by an HIV virus is $p^2 = 1.2 \times 10^{-10}$.

e) The number n' of T cells that possess these two particular mutations in an infected patient is on average $n' = p^2 \times N = 1.2 \times 10^{-10} \times 10^7 = 1.2 \times 10^{-3}$.

If the patient is treated with these two drugs simultaneously, we can be quite optimistic, as it is unlikely that they possess preexisting resistant HIV mutants in their blood. This shows the interest of bitherapy and tritherapy: even though the virus mutates frequently, it remains unlikely to have specific combinations of mutations that provide simultaneous resistance to 2 or 3 drugs. Note that in practice, real drugs don't completely stop virus replication, and there can be more than one mutation that can confer resistance, hence the fact that tritherapy and not bitherapy has become standard.

3 Cancer screening

The population can be partitioned in two different ways:

- people who are sick (s) and people who are healthy (h),
- people who are tested positive (p) and people who are tested negative (n).

a) We know $P(s)$, $P(p|s)$ and $P(p|h)$, and we want to calculate $P(s|p)$. Using Bayes' theorem, we can write

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p)} \,. \tag{22}$$

But, because an individual tested positive can be either sick or healthy (no third choice), we can write

$$P(p) = P(p,s) + P(p,h) = P(p|s)P(s) + P(p|h)P(h) = P(p|s)P(s) + P(p|h)\,[1 - P(s)] \,. \tag{23}$$

Using Eq. 23, Eq. 22 becomes

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + P(p|h)\,[1 - P(s)]} \,. \tag{24}$$

Here, with $P(s) = 0.003$, $P(p|s) = 0.5$ and $P(p|h) = 0.03$, we obtain $P(s|p) = 0.048$.

Because the disease is very rare, a person who is tested positive still has a rather low chance of having the disease, but this can motivate further tests, e.g. more invasive and costly ones. Thus, this test is still very useful, as colorectal cancer is a very serious disease.

b) Now we want to calculate $P(s|n)$. Using Bayes' theorem, we can write

$$P(s|n) = \frac{P(n|s)\,P(s)}{P(n)}\,. \tag{25}$$

Because an individual that is sick is either tested positive or negative (no third choice),

$$P(n|s) = 1 - P(p|s)\,, \tag{26}$$

and thus,

$$P(s|n) = \frac{[1 - P(p|s)]\;P(s)}{P(n)}\,. \tag{27}$$

Now we still need to calculate $P(n)$. Given the data, $P(s) = 0.003$, $P(p|s) = 0.5$ and $P(p|h) = 0.03$, we can have the intuition that $P(n)$ is very close to one. More rigorously,

$$\begin{aligned} P(n) &= 1 - P(p) = 1 - [P(p,s) + P(p,h)] = 1 - [P(p|s)P(s) + P(p|h)P(h)] \\ &= 1 - \{P(p|s)P(s) + P(p|h)\,[1 - P(s)]\}\,, \end{aligned} \tag{28}$$

which is 0.97 for $P(s) = 0.003$, $P(p|s) = 0.5$ and $P(p|h) = 0.03$, confirming our intuition. Finally,

$$P(s|n) = \frac{[1 - P(p|s)]\;P(s)}{1 - \{P(p|s)P(s) + P(p|h)\,[1 - P(s)]\}}\,, \tag{29}$$

which gives $P(s|n) = 0.0015$ for $P(s) = 0.003$, $P(p|s) = 0.5$ and $P(p|h) = 0.03$.

4 COVID-19 testing

Here we keep the same notations as in the previous section.

a) The test has a high specificity (close to 1) but a moderate sensitivity (less close to 1). More precisely, among sick patients, still 21.3% of them test negative, i.e. are false negatives, while among healthy patients, only 0.3% of them test positive. Because of this, it can be said that a positive test for COVID-19 test has more weight than a negative test. However, we will see that our interpretation of these tests strongly depends on the *a priori* estimate of the probability that a patient is sick.

b) The sensitivity α is the probability that a person is tested positive given that they are really sick. Thus, we can write

$$\alpha = P(p|s)\,, \tag{30}$$

which is the probability that one is tested p knowing that one is s.

Similarly, the specificity β is the probability that a person is tested negative given that they are really healthy, and we have

$$\beta = P(n|h)\,, \tag{31}$$

which is the probability of being tested n conditioned being h.

c) We know $P(s)$, $\alpha = P(p|s)$ and $\beta = P(n|h)$, and we want to calculate $P(s|p)$. As in the previous section (see the proof of Eq. 24 above), we can write

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + P(p|h)\,[1 - P(s)]}\,. \tag{32}$$

Because an individual that is healthy is either tested positive or negative (no third choice), we have

$$P(p|h) = 1 - P(n|h)\,, \tag{33}$$

and thus

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + [1 - P(n|h)]\,[1 - P(s)]}\,. \tag{34}$$

Using $\alpha = P(p|s)$ and $\beta = P(n|h)$, this yields

$$P(s|p) = \frac{\alpha\,P(s)}{\alpha P(s) + [1 - \beta]\,[1 - P(s)]}\,. \tag{35}$$

d) In the previous section (see the proof of Eq. 29 above) we showed that

$$P(s|n) = \frac{[1 - P(p|s)]\ P(s)}{1 - \{P(p|s)P(s) + P(p|h)\,[1 - P(s)]\}}\,, \tag{36}$$

which gives, using Eq. 33 and $\alpha = P(p|s)$ and $\beta = P(n|h)$:

$$P(s|n) = \frac{[1 - \alpha]\ P(s)}{1 - \alpha P(s) - [1 - \beta]\,[1 - P(s)]}\,. \tag{37}$$

e) See Jupyter notebook. Comparing the curves obtained to the $y = x$ diagonal shows that the curve for $P(s|p)$ deviates more from it than the curve from $P(s|n)$, which shows that a positive test changes the probability that a patient is sick from the *a priori* estimate $P(s)$ more than a negative test. This comes from the high specificity (close to 1) but the moderate sensitivity (less close to 1) of the test, and is consistent with the advice given to doctors that we analyzed in the first question above.

f) With $\alpha = 0.787$ and $\beta = 0.997$, and for $P(s) = 300/100000 = 3\times10^{-3}$, Eq. 35 yields $P(s|p) = 0.44$. This value is quite small, and this is mainly due to the low value of $P(s)$. A positive test is likely to be a false positive at such incidence rates.

g) With $P(s) = 4000/100000 = 4 \times 10^{-2}$, Eq. 35 yields $P(s|p) = 0.92$. At this higher incidence rate, a positive test is very likely to be a true positive.

h) With $P(s) = 0.9$, Eq. 37 yields $P(s|n) = 0.66$. This value is quite large, due to the large value of $P(s)$ and the moderate sensitivity of the test. Therefore, it would be quite risky that the medical doctor returns to work at this point.

5 Doping

a) The false-positive rate of the test is $p = P(+|\text{innocent}) = FP/(FP + TN) = 0.02$. Thus the probability that the test is negative if the cyclist is innocent is $P(-|\text{innocent}) = 1 - p$. Therefore the probability that no test is positive out of the 8 tests is $(1 - p)^8$, and the probability that at least one of the 8 tests is positive is $1 - (1 - p)^8 = 1 - (1 - 0.02)^8 = 0.15$.

b) The probability that no test is positive if the cyclist is innocent is $(1-p)^8$. The probability that one is positive but all others are negative if the cyclist is innocent is $8(1 - p)^7 p$. Note that the number of positive tests for our innocent cyclist follows a binomial distribution. Thus the probability that at least two of the 8 tests is positive is $1 - (1 - p)^8 - 8(1 - p)^7 p = 0.010$. It is much smaller than the previous result, meaning that having two positive tests really reinforces suspicions compared to having just one.

c) Using the same reasoning as above, the probability that at least one of the 126 tests performed is positive is $1 - (1 - p)^{126} = 1 - (1 - 0.02)^{126} = 0.92$. This is high! This means that we should be cautious when interpreting the results of these tests, especially if a single one comes out positive for a given cyclist.

d) The false-negative rate of the test is $P(-|\text{guilty}) = FN/(FN+TP)$. The probability that a cyclist is guilty given that the test is positive can be expressed using Bayes' theorem as

$$P(\text{guilty}|+) = \frac{P(+|\text{guilty})\,P(\text{guilty})}{P(+)}\,. \tag{38}$$

But we have

$$P(+) = P(+, \text{guilty}) + P(+, \text{innocent}) = P(+|\text{guilty})P(\text{guilty}) + P(+|\text{innocent})P(\text{innocent})\,, \tag{39}$$

and thus we obtain

$$P(\text{guilty}|+) = \frac{P(+|\text{guilty})P(\text{guilty})}{P(+|\text{guilty})P(\text{guilty}) + P(+|\text{innocent})P(\text{innocent})}\,. \tag{40}$$

But in addition we have

$$P(\text{innocent}) = 1 - P(\text{guilty})\,, \tag{41}$$

and thus we obtain

$$P(\text{guilty}|+) = \left[1 + \frac{P(+|\text{innocent})}{P(+|\text{guilty})}\left(\frac{1}{P(\text{guilty})} - 1\right)\right]^{-1}\,. \tag{42}$$

Using the fact that

$$P(+|\text{guilty}) = 1 - P(-|\text{guilty})\,, \tag{43}$$

we finally obtain

$$P(\text{guilty}|+) = \left[1 + \frac{P(+|\text{innocent})}{1 - P(-|\text{guilty})}\left(\frac{1}{P(\text{guilty})} - 1\right)\right]^{-1}\,. \tag{44}$$

This expression of the probability that a cyclist is guilty given that the test is positive depends on the false-positive rate $P(+|\text{innocent})$, the false-negative rate $P(-|\text{guilty})$ and the probability that a given person is guilty $P(\text{guilty}$.

If the false-negative rate $P(-|\text{guilty})$ is zero, the last formula becomes

$$P(\text{guilty}|+) = \left[1 + P(+|\text{innocent})\left(\frac{1}{P(\text{guilty})} - 1\right)\right]^{-1}\,. \tag{45}$$

This expression of the probability that a cyclist is guilty given that the test is positive depends on the false-positive rate $P(+|\text{innocent})$, which was given in the text of the problem, but also on the probability $P(\text{guilty})$ that a given person is guilty. Therefore we need the latter quantity in order to compute the probability that a cyclist is guilty given that the test is positive. In practice, it difficult to know it because for this we need to know who is really guilty.

6 Additional problem: Non-invasive prenatal testing

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.

a) The sensitivity α is the probability that a person is tested positive given that they are really sick. Thus, we can write

$$\alpha = P(p|s)\,, \tag{46}$$

which is the probability that one is tested p knowing that one is s. Similarly, the specificity β is the probability that a person is tested negative given that they are really healthy, and we have

$$\beta = P(n|h)\,, \tag{47}$$

which is the probability of being tested n conditioned being h.
For NIPT, their values are $\alpha = 0.993$ and $\beta = 0.999$.

b) We know $P(s)$, $\alpha = P(p|s)$ and $\beta = P(n|h)$, and we want to calculate $P(s|p)$. Using Bayes' theorem, we can write Eq. 22 above. But, because an individual tested positive can be either sick or healthy (no third choice), we can write Eq. 23 above. Using Eq. 23, Eq. 22 becomes Eq. 24. Because an individual that is healthy is either tested positive or negative (no third choice), we have Eq. 33, and thus

$$P(s|p) = \frac{P(p|s)\,P(s)}{P(p|s)P(s) + [1 - P(n|h)]\,[1 - P(s)]}\,. \tag{48}$$

Using $\alpha = P(p|s)$ and $\beta = P(n|h)$, this yields Eq. 35 above.

c) With $\alpha = 0.993$, $\beta = 0.999$ and $P(s) = 1/200$, it gives $P(s|p) = 0.83$. This value is high, but not very high – the possibility of mis-diagnosing is not negligible.

d) Now we want to calculate $P(s|n)$. Using Bayes' theorem, we can write

$$P(s|n) = \frac{P(n|s)\,P(s)}{P(n)}\,. \tag{49}$$

Because an individual that is sick is either tested positive or negative (no third choice), we obtain Eq. 26 above, and thus,

$$P(s|n) = \frac{[1-\alpha]\;P(s)}{P(n)}\,. \tag{50}$$

Now we need to calculate $P(n)$:

$$\begin{aligned} P(n) &= 1 - P(p) = 1 - [P(p,s) + P(p,h)] = 1 - [P(p|s)P(s) + P(p|h)P(h)] \\ &= 1 - \{P(p|s)P(s) + P(p|h)\,[1 - P(s)]\} = 1 - \{\alpha P(s) + (1-\beta)\,[1 - P(s)]\}\;, \end{aligned} \tag{51}$$

(or see above for the calculation of $P(p)$), yielding Eq. 29.

e) The formula above gives $P(s|n) = 3.5 \times 10^{-5}$ for $P(s) = 1/200$ and the values of α and β for NIPT. This is very small (even in this rather high-risk population): a negative NIPT essentially excludes the possibility of trisomy 21.

f) If the NIPT result is negative, the possibility of trisomy 21 is extremely small, and thus the risks associated with amniocentesis outweigh the potential gains. Conversely, if the NIPT result is positive, the risk of trisomy 21 is large, and thus it is a good idea to recommend amniocentesis (and to perform it if the patient agrees) to make the diagnosis clearer. Thus, NIPT is a screening test that allows to restrict amniocentesis to the cases where it is most useful.

g) If the two tests can be considered independent, then the probability to get 2 positive tests given that a patient is healthy is smaller than that to get just 1 positive tests. Thus, this strategy can reduce the risk of a false positive. More precisely, for independent tests, $P(t_1, t_2|s) = P(t_1|s)P(t_2|s)$.

h) Because 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, a given patient who got a false positive is likely to get another one due to the same anomaly. In other words, the two tests will not be independent at all. Thus, recommending to take multiple successive NIPT is not useful.

i) For Prader-Willi syndrome, $P(s|p) = 0.05$. Most of the positive tests (95%) correspond to false positives in this case. This is despite the high specificity and sensitivity of the test. It is due to the fact that Prader-Willi syndrome is rare, much rarer than trisomy 21.

j) Given the very large proportion of positives that are in fact false positives, the usefulness of NIPT for rare genetic anomalies like Prader-Willi syndrome is very limited and disputable. A positive result could possibly be used to recommend amniocentesis, but it is not clear at all in this case that the risks of the procedure would be worth it, given that the risk of an anomaly remains small.

k) It is scientifically valid to say that these tests as "correct in more than 99% of the cases". Indeed, both sensitivity and specificity are above 0.99. So, whether a patient is sick or healthy, the result will reflect this (i.e. be correct) in more than 99% of the cases. However, if this statement is the main advertisement of the tests destined to patients, there is a huge risks that these tests will be mis-interpreted in the (rare, but important) case of a positive test. Indeed, a patient who tests positive for a rare genetic anomaly of the embryo, e.g. Prader-Willi syndrome, will think that their positive result has more than 99% chances to be correct (i.e. to be a true positive), which is not the case at all. This is because $P(s|p)$ can be much smaller than $P(p|s)$ when $P(s)$ is very small. The dangers of such a correct but incomplete and misleading advertisement are that patients may worry for nothing, take unnecessary invasive tests, and even plan to terminate a pregnancy that is in fact healthy. This illustrates the importance of interpreting test results correctly, but also of communicating about their accuracy in a precise way.

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) To compute the probability that no box on the grid has more than one tagged cell, we will consider that each tagged cell ends up in a box chosen uniformly at random when we spread the sample. The total number of ways tagged cells can be arranged on the grid is N^M. Indeed, the first tagged cell can be in any of the N boxes, the second one too, etc., and there are M tagged cells in the sample. The number of ways can be arranged on the grid without any box containing more than one tagged cell is $N \times (N-1) \times \cdots \times (N-M+1) = N!/(N-M)!$. Indeed, the first tagged cell can be in any of the N boxes, but then the second one should be in one of the $N-1$ remaining empty boxes, etc., and there are M tagged cells in the sample. Thus, the probability p that no box on the grid has more than one tagged cell is th ratio of the number of cases satisfying this constraint to the total number of cases:

$$p = \frac{N!}{(N-M)!} \times \frac{1}{N^M} \,. \tag{52}$$

b) The probability that at least one box on the grid contains more than one of the M tagged cells is $1-p$. Indeed, either no box on the grid has more than one tagged cell or at least one box on the grid contains more than one of the M tagged cells, these two events are exclusive and there is no third possibility, so their probabilities sum to 1.

c) For $N = 400$ and $M = 20$, p is 0.62 and thus $1-p$ is 0.38. There is a 38% chance that at least one box on the grid contains more than one of the tagged cells.

d) Similarly, for $N = 365$ and $M = 20$, p is 0.59 and $1-p$ is 0.41. There is a 41% chance that at least two students in the class have the same birthday.