Exercise 2

A student test consists of N questions, each one of which has k possible answers, among which the student is requested to select the only correct one ($N, k > 0$).

Jack tries his luck by making blind, independent guesses at each question.

1) Find the expression for the PMF P_n of the number n of correct answers in the whole test
2) Discuss possible approximations of P_n when $N = 80$, $k = 5$, and compute an approximated value for the probability that Jack answers 40 or more questions correctly.
3) (For generic *N* and *k* values): assume that each correct answer earns Jack +1 points, and each wrong answer earns it -1 points. What is the mean value of Jack's score at the test s?
4) What is the score that one should get for a wrong answer, in order for the mean value of s to be equal to zero?

Now, assume that Jack has some knowledge in the subject matter of the test. He (thinks he) *knows* the answer to M questions ($0 \le M \le N$), which he sets apart in advance and answers to independently, and he *guesses* the remaining $N - M$ ones. When he *knows* the answer, he has 90% probability of actually getting the correct answer.

5) Compute the expression of the probability that Jack answers correctly to exactly *M* questions.

Solution

1) This is a repeated trial problem, with a probability of success equal to $p = 1/k$. The PMF is thus the well-known binomial PMF:

$$P_n = \binom{N}{n} p^n \cdot (1-p)^{N-n}$$

2) A binomial can be approximated with:
   a. A Poisson variable, if N is large and p is small. In this case, $p = 0.2$, hence this approximation incurs large errors.
   b. A Gaussian variable, if $N \cdot p \cdot (1-p) > 10$. With $N = 80$, $k = 5$, it is $N \cdot p \cdot (1-p) = 12.8$, hence the required condition holds.

In this case, it is $P_n \approx P\left(n-0.5 \le X \le n+0.5\right)$, with $X \sim N\left(N \cdot p, N \cdot p \cdot \left(1-p\right)\right)$, i.e. $X \sim N\left(16, 12.8\right)$. Therefore, we have $P_n \approx P\left(\frac{n-0.5-16}{\sqrt{12.8}} \le \frac{X-16}{\sqrt{12.8}} \le \frac{n+0.5-16}{\sqrt{12.8}}\right)$.

With $n=40$, it is

$$\begin{aligned} P\{n \ge 40\} &= 1-\Phi\left(\frac{23.5}{\sqrt{12.8}}\right) \\ &\approx 1-\Phi\left(6.57\right) \\ &\approx 0 \end{aligned}$$

Since notoriously $\Phi\left(x\right) \approx 0$ when $x \ge 3$. This also makes sense intuitively. Note that the answer obtained without approximations is $P\{n \ge 40\} \cong 2.06 \cdot 10^{-9}$.

3) The mean score for a question is $1 \cdot p + \left(-1\right) \cdot \left(1-p\right) = 2p-1$, and the sum of the means is equal to the mean of the sum. Therefore, the mean score for the whole test is $E\left[s\right] = N \cdot \left(2p-1\right)$. This tells us that the mean test score is negative if the probability of success is less than 50% at each question, which makes sense intuitively.

The same result could be obtained via a considerably longer route as follows: with n correct answers out of N, the test score will be $n-\left(N-n\right)$. Thus, the mean value of the score is $E\left[s\right] = \sum_{n=0}^{N} P_n \cdot \left[n-\left(N-n\right)\right] = 2\sum_{n=0}^{N} n \cdot P_n - N \cdot \sum_{n=0}^{N} P_n$. This can be further developed as:

$$\begin{aligned} E\left[s\right] &= 2\sum_{n=0}^{N} n \cdot P_n - N \cdot \sum_{n=0}^{N} P_n \\ &= 2\sum_{n=1}^{N} n \cdot \binom{N}{n} p^n \cdot \left(1-p\right)^{N-n} \quad - N \\ &= 2\sum_{n=1}^{N} N \cdot \binom{N-1}{n-1} p^n \cdot \left(1-p\right)^{N-n} \quad - N \\ &= 2N \cdot p \cdot \sum_{j=0}^{N-1} \binom{N-1}{j} p^j \cdot \left(1-p\right)^{(N-1)-j} \quad - N \\ &= 2N \cdot p - N \\ &= N \cdot \left(2p-1\right) \end{aligned}$$

4) In order for the mean test score to be null, *some* test scores must be negative, and this can only be obtained if the score for a wrong answer is negative. Assume that a wrong answer gets $-\delta$, with $\delta > 0$. The mean test score is null if and only if the mean score of each question is null, and this happens if $1 \cdot p + \left(-\delta\right) \cdot \left(1-p\right) = 0$, i.e. $\delta = p/\left(1-p\right) = 1/\left(k-1\right)$ The same result could be obtained via the longer route, by observing that the mean test score is:

$E\left[s\right] = \sum_{n=0}^{N} P_n \cdot \left[n-\left(N-n\right) \cdot \delta\right] = \left(1+\delta\right)\sum_{n=0}^{N} n \cdot P_n - \delta N \cdot \sum_{n=0}^{N} P_n$

Based on the computations of the previous point, one clearly sees that the mean value for the test score is:

$$E[s] = (1+\delta)\cdot N \cdot p - \delta N = N \cdot \left[(1+\delta)\cdot p - \delta\right].$$

Now, in order to be $E[s] = 0$, we need $\delta = p/(1-p) = 1/(k-1)$.

5) The probability that Jack gets *n* correct answers can be written as follows:

$$P_M' = \sum_{j=0}^{M} P\left\{(M-j) \text{ correct guesses, } j \text{ correct "known" answers}\right\}$$
$$= \sum_{j=0}^{M} \left(P_{M-j} \mid Q_j\right) \cdot Q_j$$

Where Q_j is the probability that Jack gets *j* correct answers among the *M* that he thinks he knows. Now, it is correct to assume that P_{M-j} and Q_j are independent of each other, hence the formula is:

$$P_M' = \sum_{j=0}^{M} P_{M-j} \cdot Q_j$$
$$= \sum_{j=0}^{M} \left[\binom{N-M}{M-j} p^{(M-j)} \cdot (1-p)^{(N-M)-(M-j)} \right] \cdot \left[\binom{M}{j} q^j \cdot (1-q)^{M-j} \right]$$
$$= p^M \cdot (1-p)^{(N-M)} \cdot \sum_{j=0}^{M} \binom{N-M}{M-j} \cdot \binom{M}{j} \cdot \left[\left(\frac{q}{p}\right)^j \cdot \left(\frac{1-q}{1-p}\right)^{M-j} \right]$$

Where $q = 0.9$ is the probability that Jack answers correctly to a question whose answer he thinks he knows. Note that the above summation can start from $j = \max(0, 2M-N)$, since the first binomial coefficient is null if $j < 2M - N$.