Test 2 – Answers

The following question is taken from the exam in the year 2014.

> The incidence rate of a rare characteristic in a large population is estimated to be one in 100,000. Consider a population of 200,000 people.

☞ Before we start answering questions on probabilities of observations from a given population, we need to come up with a probability model. But we don't need to determine an underlying probability distribution from scratch! Luckily, in most cases the distribution has a known special form. For this experiment, the underlying distribution must be of discrete type and we thus need to decide whether it is the *Bernoulli*, *binomial*, *geometric*, *hypergeometric*, *Poisson*, or *discrete uniform* distribution.

Setting: We can use a Bernoulli process as an adequate probability model: Let the possession of a rare characteristic be labeled as "success" and the lack of this characteristic as "failure". We conduct 200,000 consecutive, independent Bernoulli trials, with the probability of success $\frac{1}{100{,}000}$ in each trial. This process can be regarded as a sampling with replacement, done theoretically – without engaging in any real experiments – as if we would select a person from a population of 200,000 people, one after another, observe whether a given characteristic is present or absent, and then put this person back into the population. We are interested in probabilities to receive a certain number of successes.

Characteristics of a Bernoulli process:

1. In each Bernoulli trial there are two distinct possible outcomes: possession of a rare characteristic, labeled as "success", or lack of it, labeled as "failure".
2. The trials are statistically independent; that is, the outcome of any trial is not affected by the outcomes of earlier trials and it does not affect the outcomes of later trials.
3. The probability of success is the same in each trial.

Let X be a random variable that represents the number of occurrences of the rare characteristic in the population of size 200,000; that is, X is the number of successes that occur in 200,000 consecutive trials of the Bernoulli process. Obviously, X is of the discrete type. Our objective is to identify a probability distribution of X.

A learning tip (mnemonic tool): The *bi* part in the word "binomial distribution" refers to *two* possible outcomes in each of finitely many consecutive trials.

Model: $X \sim \text{BIN}(200{,}000, \frac{1}{100{,}000})$. The formula to compute the probability $P[X = r]$ of observing exactly r occurrences of the rare characteristic in the given population is

$$P[X = r] = \frac{200{,}000!}{r!(200{,}000 - r)!} \left(\frac{1}{100{,}000}\right)^r \left(1 - \frac{1}{100{,}000}\right)^{200{,}000-r}.$$

But this formula is not well suited for computations! Fortunately, for a Bernoulli process, in which the number of trials is large and the probability of success on any given trial is very small, the Poisson distribution can be used as an approximation to the binomial distribution. So, here $200{,}000 \times \frac{1}{100{,}000} = 2$ and we can use the Poisson distribution POI(2) instead of BIN(200,000, $\frac{1}{100{,}000}$).

An approximative model: $X \sim \text{POI}(2)$. The formula to compute $P[X = r]$ is

$$P[X = r] = \frac{2^r e^{-2}}{r!}.$$

Note that the codomain of the random variable X is the set $\{0, 1, 2, 3, \ldots\}$ of non-negative integers.

(i) Find the probability that at least one person has the rare characteristic.

Solution: We need to find the following "greater than or equal to" probability: the probability that X is greater than or equal to 1, i.e. $P[X \geq 1]$.

$$\begin{aligned} P[X \geq 1] &= 1 - P[X < 1] = 1 - P[X = 0] \\ &= 1 - \frac{2^0 e^{-2}}{0!} = 1 - \frac{1 \cdot e^{-2}}{1} \\ &= 1 - e^{-2}. \end{aligned}$$

(ii) Suppose that one person with the rare characteristic has been found in the population. Find the probability that there is at least one other person with the rare characteristic.

An alternative formulation of (ii): Given that one person with the rare characteristic has been found in the population, find the probability that there are at least two people with the rare characteristic.

Solution: We need to find the following conditional "greater than or equal to" probability: the probability that X is greater than or equal to 2, given that X takes on values in the subset $\{1, 2, 3, \ldots\}$ of the set of all non-negative integers. That is, we need to calculate $P[X \geq 2 | X \geq 1]$.

$$\begin{aligned} P[X \geq 2 | X \geq 1] &= \frac{P[X \geq 2 \text{ and } X \geq 1]}{P[X \geq 1]} = \frac{P[X \geq 2]}{P[X \geq 1]} = \frac{1 - P[X < 2]}{1 - P[X < 1]} \\ &= \frac{1 - (P[X = 0] + P[X = 1])}{1 - P[X = 0]} \\ &= \frac{1 - \left(\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!}\right)}{1 - \frac{2^0 e^{-2}}{0!}} = \frac{1 - (e^{-2} + 2e^{-2})}{1 - e^{-2}} \\ &= \frac{1 - 3e^{-2}}{1 - e^{-2}} \approx 0.69. \end{aligned}$$

(iii) Suppose that two people with the rare characteristic have been found. Find the probability that there are no others with the rare characteristic in the population.

An alternative formulation of (iii): Given that two people with the rare characteristic have been found in the population, find the probability that there are exactly two people with the rare characteristic in the population.

Solution: We need to find the following conditional "equal to" probability: the probability that X is equal to 2, given that X takes on values in the subset $\{2, 3, \ldots\}$ of the set of all non-negative integers. That is, we need to calculate $P[X = 2|X \geq 2]$.

$$\begin{aligned} P[X = 2|X \geq 2] &= \frac{P[X = 2 \text{ and } X \geq 2]}{P[X \geq 2]} = \frac{P[X = 2]}{P[X \geq 2]} = \frac{P[X = 2]}{1 - P[X < 2]} \\ &= \frac{P[X = 2]}{1 - (P[X = 0] + P[X = 1])} \\ &= \frac{\frac{2^2 e^{-2}}{2!}}{1 - \left(\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!}\right)} \\ &= \frac{2e^{-2}}{1 - 3e^{-2}} \approx 0.46. \end{aligned}$$