

HWA 2 – Probability & Statistics I – 2018: Models for real life situations.

Hand in on A4 (hard copy with name and number) **at the start of the lecture on Thursday, March 29.**

Motivate your choice of the model briefly (type of distribution) if one of the common distributions applies!

- a. It is known that the success rate of an application for grants of a research council is (about) 15%. If we assume that this success rate applies to a researcher, who is applying for a grant every year, then X is the number of trials (years) until he receives a grant.

Determine the **distribution of X** , $E(X)$ and $P(X > 5)$.

- b. When a search for a word or concept is executed on the web by a search machine, it will search for the word in a limited number of sites. Suppose that the concept “Big data analysis” occurs in about 0.1% of the websites and that the search machine selects (arbitrarily) 4000 websites to search for this concept.

X is the number of websites, that the search machine found, containing the concept “Big data analysis”.

Determine the **distribution of X** and compute (or approximate) $P(X > 7)$.

- c. A company wants to hire a new employee. Their advertisement resulted in many reactions, but after selection of the letters 9 equally suitable candidates remained: 5 men and 4 women. It was decided that not all of them will be invited for an interview, but the management demanded that at least one woman is invited. The choice of the invited persons will be made as follows.

Firstly 3 persons are chosen at random. If at least one woman is included, these 3 will be invited. But if no woman is chosen, another person is chosen at random from the remaining (6) candidates. And if this 4th person is not a woman a fifth person is chosen, etc., until a woman is among the invited candidates.

Determine the **distribution of X** = “the number of invited candidates” and compute $E(X)$ and $var(X)$.

a	b	c	Tot
1+1+1	1+2	2+1+1	10

Solutions:

- a. The yearly trials are Bernoulli-experiments with success probability $p = 0.15$, so the required number X of trials to succeed has a **geometric distribution with parameter $p = 0.15$** . (or give $P(X = k) = 0.85^{k-1} \cdot 0.15$, $k = 1, 2, \dots$)

$$E(X) = \frac{1}{p} = \frac{1}{0.15} \approx 6.67 \quad \text{and}$$

$$P(X > 5) = (1 - p)^5 = (0.85)^5 = 44.4\% \quad (\text{"}X > 5\text{" is equivalent to "first 5 trials no success"}!)$$

(If you cannot reproduce the property above you might derive the property, using the geometric series:

$$\begin{aligned} P(X > 5) &= \sum_{k=6}^{\infty} (0.85)^{k-1} \cdot 0.15 = 0.15 \sum_{k=6}^{\infty} (0.85)^{k-1} = 0.15 \cdot 0.85^5 \sum_{k=6}^{\infty} (0.85)^{k-6} \\ &= 0.15 \cdot 0.85^5 \cdot \frac{1}{1-0.85} = 0.85^5 \end{aligned}$$

or (but preferably not) via the complement rule: $P(X > 5) = 1 - \sum_{k=1}^5 0.85^{k-1} \cdot 0.15 = \dots$)

- b. We assume that, for the 4000 sites, containing the concept “Big data analysis” is an event that occurs independently, with probability 0.1%: Counting the number of sites containing the concept, X has a $B(4000, 0.001)$ -distribution.

But for the computation of the required probability we can use the Poisson approximation for large n and

small p , since $n = 4000 > 25$ and $np = 4 < 10$, so X is appr. Poisson with $\mu = np = 4000 \times 0.001 = 4$
 $P(X > 7) = 1 - P(X \leq 7) = 1 - 0.949 = 5.1\%$.

(Exact computation with the binomial distribution is possible, the result is 0.051044, but not advisable:
 directly opting for the Poisson distribution without mentioning the binomial distribution: -1/2 point)

c. If we denote M = “man” and W= “woman” in each draw, then the possible outcomes are: “at least one woman in the first 3 draws” ($X = 3$), MMMW ($X = 4$), MMMMW ($X = 5$) and MMMMMW ($X = 6$):

$$P(X = 3) = 1 - P(MMM) = 1 - \frac{\binom{5}{3}}{\binom{9}{3}} = 1 - \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} = 1 - \frac{15}{126} = \frac{111}{126} (\approx 0.8810)$$

$$P(X = 4) = P(MMMV) = P(MMM)P(V|MMM) = \frac{\binom{5}{3}}{\binom{9}{3}} \cdot \frac{4}{6} = \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{4}{6} = \frac{10}{126} (\approx 0.0794)$$

$$P(X = 5) = P(MMMMMV) = \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{4}{5} = \frac{4}{126} (\approx 0.0317)$$

$$P(X = 6) = P(MMMMMMV) = \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{1}{5} \cdot \frac{4}{4} = \frac{1}{126} (\approx 0.0079)$$

$$\text{total of the probabilities} = \frac{\quad}{\quad} + 1 (\approx 1.0000).$$

$$E(X) = \sum_{x=3}^6 x P(X = x) = \dots = \frac{399}{126} = \frac{19}{6} (\approx 3.167)$$

$$E(X^2) = \sum_{x=3}^6 x^2 \cdot P(X = x) = \dots = \frac{1295}{126} (\approx 10.278)$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{1295}{126} - \left(\frac{399}{126}\right)^2 = \frac{63}{252} = \frac{1}{4}$$

(Note that both the geometric and hypergeometric distributions cannot apply, e.g. since the range of X , $S_X = \{3, 4, 5, 6\}$, does not “fit” these distributions, or: the trials are dependent (not geometric) and it is not about the counting of successes (not hypergeometric).)