MTH 224, Spring 2024

Instructor: Bahman Angoshtari

Lecture 10

Section 2.5: hypergeometric distribution.
Section 2.7: Poisson distribution.

10.1. The Hypergeometric r.v.

- This distribution is best described through an example.

Example 10.1. An urn contains N balls of which D are green and $N-D$ are red. We take a sample of n balls, **without** replacement. Let X be the number of green balls in the sample. Find the pmf of X.

Solution. The sample space has $\binom{N}{n}$ equally likely outcomes. There are $\binom{D}{k}$ ways to choose k balls out of a total of D green balls, and there are $\binom{N-D}{n-k}$ ways to choose $(n-k)$ balls out of a total of $N-D$ red balls. So,

$$p_X(k)=\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}};\quad k=\max\{0,n-(N-D)\},\dots,\min\{n,D\}.$$

The support of X (that is, the possible values for k) is justified as follows:

- minimal value is the max between 0 and $\underset{\text{sample size}}{\underset{\downarrow}{n}} - \underset{\#\text{ of red balls}}{\underset{\downarrow}{(N-D)}}$
- maximal value is the min between D (number of green balls) and n (sample size).

Definition 10.2. A r.v. X has a hypergeometric distribution with parameter N, D, and n if its pmf is

$$p_X(k)=\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

for $\max\{0,n-(N-D)\}\le k\le\min\{n,D\}$. Notation: $X\sim\mathrm{HG}(N,D,n)$.

Remark 10.3. If $X\sim\mathrm{HG}(N,D,n)$, then $\mathbb{E}[X]=\frac{nD}{N}$ and $\mathrm{Var}(X)=\frac{nD}{N}\cdot\left(1-\frac{D}{N}\right)\cdot\frac{N-n}{N-1}$. This can be shown by direct calculation using the definition of mean and variance, or by using indicator random variables (try!).

10.2. Poisson distribution

- The pmf $p(k)$ of $\mathrm{Bin}(n,p)$ is sometimes difficult to compute, especially for large n. We will approximate the pmf of $\mathrm{Bin}(n,p)$ under the following assumptions:
  - n is large $(n\to\infty)$
  - p is small, or successes are rare $(p\to 0)$
  - np is of moderate size ($p=\frac{\lambda}{n}$ for a constant λ, or $np\to\lambda$).

- Under these assumptions, for a fixed k we have:

$$p\left(k\right)=\frac{n!}{k!\left(n-k\right)!}\cdot\left(\frac{\lambda}{n}\right)^{k}\cdot\left(1-\frac{\lambda}{k}\right)^{n-k}=$$
$$=\frac{n\left(n-1\right)\left(n-2\right)\cdots\left(n-k+1\right)}{k!}\cdot\frac{\lambda^{k}}{n^{k}}\cdot\left(1-\frac{\lambda}{n}\right)^{n}\cdot\left(1-\frac{\lambda}{n}\right)^{-k}.$$

Note that :
  - $\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^{k}}=\frac{n}{n}\times\frac{n-1}{n}\times\frac{n-2}{n}\times\cdots\times\frac{n-k+1}{n}\to 1$ ($n\to\infty$, and k fixed)
  - $\left(1-\frac{\lambda}{n}\right)^{n}\to e^{-\lambda}$. Recall that $\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}=\mathrm{e}$.
  - $\left(1-\frac{\lambda}{n}\right)^{-k}\to 1$ ($n\to\infty$, and k fixed)

- Combining the above, we get:

$$\lim_{n\to\infty}p\left(k\right)=\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^{k}\left(1-\frac{\lambda}{n}\right)^{n-k}=\frac{\lambda^{k}}{k!}e^{-\lambda}.$$

- This function defines the Poisson distribution:

Definition. (Poisson distribution). A r.v. X has the Poisson distribution with parameter $\lambda>0$ if X has pmf:

$$p\left(k\right)=\frac{\lambda^{k}}{k!}e^{-\lambda},k=0,1,2,3,\ldots$$

Notation: $X\sim\text{Pois}\left(\lambda\right)$.

Remark 10.4. The Poisson distribution was introduced by Poisson in 1837, along with applications in criminal lawsuits (jury decisions). However, it had not attracted much attention until a book by Bortkiewicz was published in 1898, which included a strange example: deaths by horse kicks in the Prussian army. Using the Poisson distribution, such rare events were shown to be quite regularized and predictable, which helped to popularize this distribution.

Example 10.5. In a class of 40 students, on average, 2 students have been sick in the past. We showed in Example 9.6 of Lecture 9 that X, the number of sick students, is $\text{Bin}\left(40,1/20\right)$. We then calculated $\mathbb{P}[X=4]=0.09012$.
Let us use Poisson approximation. Here $X\sim\text{Bin}\left(40,1/20\right)\approx\text{Pois}\left(2\right)$ (large n and small p). Hence, $\mathbb{P}\left(X=4\right)\approx\frac{2^{4}}{4!}e^{-2}\approx 0.09022$. This is a good approximation!

Remark 10.6. One may use Poisson approximation to the binomial distribution, that is $\text{Bin}\left(n,p\right)\approx\text{Pois}\left(np\right)$, whenever the number of independent trials is large, p is small, and $\lambda=np$ is of moderate size.

As a consequence of the fact that a Poisson r.v. is the limit of Binomial r.v.s, we get:

Example 10.7. For $X\sim\text{Pois}\left(\lambda\right)$, show that $\mathbb{E}\left[X\right]=\lambda$ and $\text{Var}\left(X\right)=\lambda$.

Solution. ***Method 1:*** The pmf of X is the limit of the pmfs of Binomial r.v. $X\left(n,\frac{\lambda}{n}\right)$, hence one can expect to have $\mathbb{E}\left[X\right]\approx np=\lambda$ and $\text{Var}\left(X\right)\approx np\left(1-p\right)=\lambda\left(1-p\right)\to\lambda$ (since $p\to 0$).
Method 2: (Optional reading, not covered during the lecture) Alternatively, we can directly calculate the mean and variance. First, recall that the Taylor series of an exponential function is $e^{\lambda}=\sum_{k=0}^{\infty}\frac{\lambda^{k}}{k!}$. Therefore,

$$\mathbb{E}[X]=\sum_{k=0}^{\infty}k\frac{\lambda^{k}}{k!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda.$$

Similarly,

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^{k-1}}{(k-1)!}. \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{(k-1+1)\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \left(\sum_{k=2}^{\infty} \frac{\lambda^{k-1}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \\ &= \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda} \right) = \lambda(1+\lambda).\end{aligned}$$

Finally, we obtain that $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$.

- Besides approximating the binomial distribution, the Poisson distribution is an appropriate model in various situations, and has numerous applications. For example,
  (1) The number of accidents on a section of a highway on a given day
  (2) The number of floods at a river over a century.
  (3) The number of α particles emitted from a radioactive source during a fixed period of time.

Poisson Process. Often, we consider the number $N(t)$ of occurrences of an event in an interval of length t, say $[0, t]$, where the average number of occurrences is proportional to the length of the interval, that is $\mathbb{E}[N(t)] = \lambda t$. Furthermore, the number of events happening over non-overlapping intervals are independent. It is possible to conclude from these two conditions (and some additional technical conditions) that, for each t, $N(t) \sim \mathrm{Pois}(\lambda t)$. The family of r.v.s $\{N(t)\}$ for different t is called a Poisson process with rate λ.