Introduction to probability, statistics and data handling

Tomasz Szumlak, Agnieszka Obłąkowska-Mucha
Faculty of Physics and Applied Computer Science
AGH UST Krakow

2021

The Binomial Distribution

- We already know this distribution – it emerges when we consider an experiment such as tossing a coin, rolling a die or choosing a marble from a box repeatedly.
- So, we considering trials. Each outcome will have constant probability assigned (that should not change in time, and is the parameter of the Bernoulli prob. model family).
  - Sometimes we are also interested in processes where the probability is not constant (out of the scope of our lecture, however)
- We then say that p is a success and q is a failure (in a Bernoulli sense) and can compose the following P.D.F.

$$\boldsymbol{f(x) = B(n,p) = p(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}}$$

- The RV denote the number of successes x in n trials, $x = 0, 1, \dots, n$

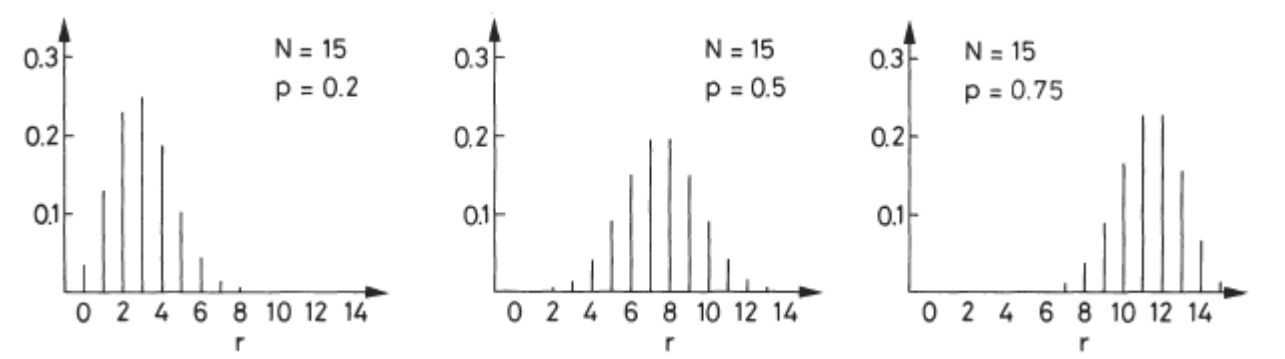
The Binomial Distribution

- The mean and variance can be fairly easy calculated:

$$\mu = \sum_x xP(x) = np$$

$$\sigma^2 = \sum_x (x-\mu)^2 P(x) = np(1-p)$$

- In the limit of „large" n and „no too small" p we can very accurately approximate the Binomial distribution with Gaussian one



The Binomial Distribution

- **Properties of the Binomial P.D.F.**

Mean	$\mu = np$
Variance	$\sigma^2 = npq$
Standard deviation	$\sigma = \sqrt{npq}$
Coefficient of skewness	$\alpha_3 = \frac{q - p}{\sqrt{npq}}$
Coefficient of kurtosis	$\alpha_4 = 3 + \frac{1 - 6pq}{npq}$

- We can mote something interesting here
- **Theorem 6**. Let X be the RV giving the number of successes in n Bernouli trials, so that $\frac{X}{n}$ is the proportion of successes. Then if p is the probability of success and ϵ is any positive number:

$$\lim_{n\to\infty} prob\left(\left|\frac{X}{n} - p\right| \geq \epsilon\right) = 0$$

Poisson distribution

- The RV of discrete type:

- the number of outcomes occurring, for instance, during a given time (e.g. number of radioactive decays in a sample of radioactive material) t: $X = X_t = 0, 1, 2, \ldots$

- number of events in a given region of space - e.g. number of typing errors per page

or:

- telephone calls arriving during a (short) period of time;
- light quanta (photons) arriving at detecting system;
- number of mutations on a strand of DNA (per unit length);
- number of customers arriving at a counter;
- number of cars arriving at a trac light;
- number of Losses/Claims;

- POISSON

Poisson distribution

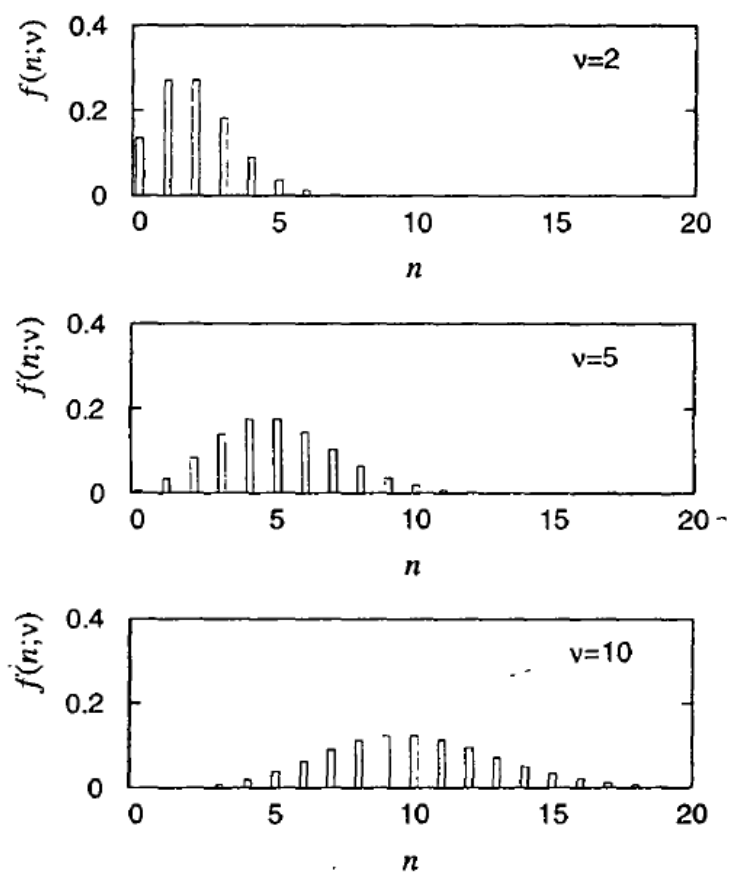
- Numbers of outcomes occurring in one time interval t are independent of each other, i.e. the number occurring in one time interval is independent of the number that occurs in any other disjoint time interval (Poisson process has no memory)
- The probability that a single outcome will occur during a very short time interval t is PROPORTIONAL to the length of interval:

$$P(X_{\Delta t} = 1) \sim \Delta t$$

- Let me to introduce the Poisson distribution:

$$P(X_t = k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Poisson distribution



- The number of decays of radioactive material in a fixed time period follows the Poisson distribution (given that the decay probability is constant over the time period)

Poisson distribution

- Let's take the binomial distribution, it can be shown that in the limit of large n and very small p (given that np is finite) we get a new distribution:

$$f(n; n \cdot p = \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

- This Poisson distribution is valid for integer variable n ($n = 0, 1, \ldots$) and has a single parameter $n \cdot p = \nu$
- Its mean value and variance

$$E[n] = \sum_{n/0}^{\infty} n \frac{\nu^n}{n!} e^{-\nu} = \nu \quad V[n] = \sum_{n/0}^{\infty} (n - \nu)^2 \frac{\nu^n}{n!} e^{-\nu} = \nu$$

- In many cases the n can be treated as a continuous variable. Also, if ν is large the Poisson random variable can be treated as a continuous variable similar to the **normal distribution**

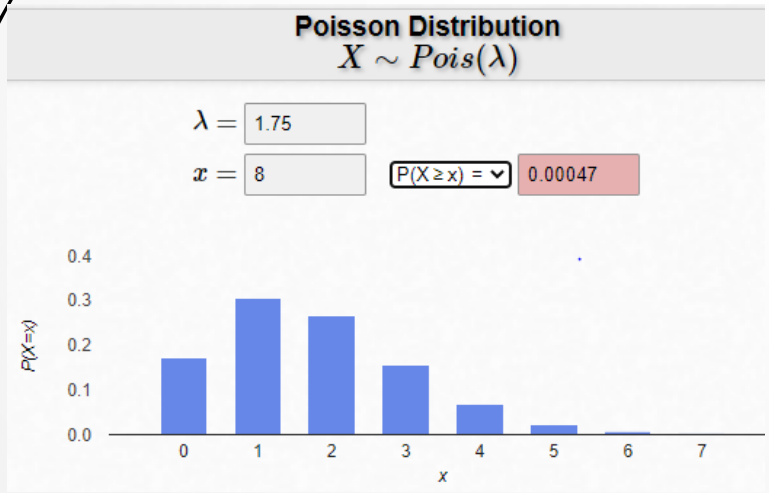
The Poisson Distribution

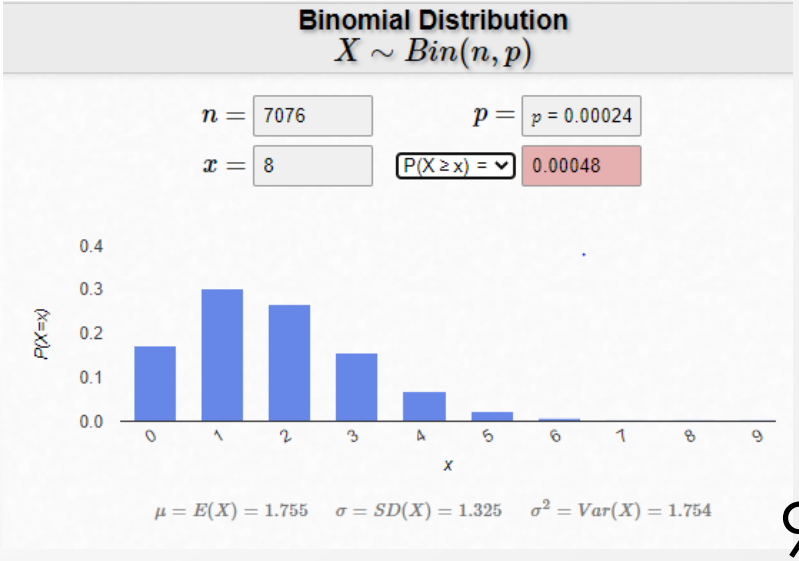
The probability of getting leukemia is $p = 0.000248$. Using the approximation Bernoulli-to-Poisson find the P of eight or more leukemia cases in a population of size $n = 7076$.

$$np = 7076 \times 0.000248 = 1.75 = \lambda.$$

$$\mathcal{P}(X \leq 7) = \sum_{0 \leq x \leq 7} e^{-1.75} \frac{1.75^x}{x!} = 0.999518.$$

hence $\mathcal{P}(X \geq 8) \approx 1 - 0.999518 = 0.000482.$

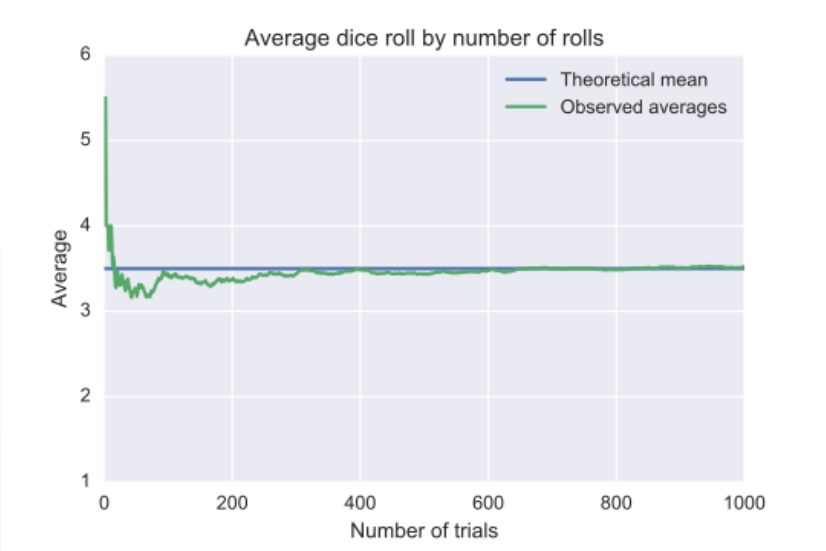




Law of large numbers

- Building on the knowledge we gained today, we can formulate a very advanced theorem, that is considered fundamental for statistics.
- **Theorem 7**. *Let $X_1, X_2, \cdots, X_n$ be mutually independent RV (discrete or continuous), each having finite mean μ and variance σ^2. Then if we take into consideration a new RV: $S_n = X_1 + X_2 + \cdots + X_n$, then:*

$$\lim_{n\to\infty} p\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0$$



The Law of Large Numbers

- Imagine that our sample space can be divided into k events (or outcomes that we wish to study): $\{A_j\}, j = 1, \ldots, k$. What are the respective probabilities of such events p_j?
- Well, in principle we should conduct an experiment, collect a data sample and then calculate the frequency f_j (we assume that n below means the number of events of type j observed):

$$f_j = \frac{1}{n}\sum_{i/1}^{i/n} X_{ij} = \frac{1}{n}X_j$$

- So, note that X_j is a binomial R.V. that takes the following values:

$$X_j = \begin{cases} 1 \; if \, A_j \; occured \\ 0 \quad otherwise \end{cases}$$

- Now, how is f_j related to the probability p_j? Remember, the probability is just a number, whilst the frequency is a R.V.

The Law of Large Numbers

- It is essential to understand, the last point – remember frequency will always depend on a particular sample! Different sample will yield a different frequency.
- Having said that, we can however write (remember that X_j is a binomial R.V.!):

$$E[f_j] = E\left[\frac{X_j}{n}\right] = \boldsymbol{p_j}, E[X_j] = np_j$$

$$\sigma^2(f_j) = \sigma^2\left(\frac{X_j}{n}\right) = \frac{1}{n^2}\sigma^2(X_j) = \frac{1}{n}p_j q_j = \boldsymbol{\frac{1}{n}p_j(1-p_j)}$$

- In words: the expectation value of the frequency (event A_j) is equal to the probability of success. The variance of the frequency about its mean value can, in turn, be made arbitrarily small – just need to collect enough data! (large n).
- **This, actually, is the law of large numbers**!

The Law of Large Numbers

- Let's just think about the variance for a second. It is a product of these two elements: $1/n$ and $p_j(1-p_j)$. The latter is always less than unity (the max value is: $max\{p_j(1-p_j) = 1/4\}$), so the „smallness" of departure **will be governed by the number of observed events**.
- Using this argument and the result from previous slide we can justify that the approach, where **respective probabilities** of events that are estimated by **frequencies measured directly** in experiments, is **the right one**!
- The square of the error we make doing so is inversely proportional to the number of measurements in an experiment – this kind of error is called a **statistical** one
- This is essence of, so called, **counting experiments** such as: number of decaying particles, number of animals with a given traits, number of defective items, …

The Law of Large Numbers

- And finally to sum up the **LoLN**

Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables (no particular P.D.F. is assumed here), each of which have finite mean, μ, and variance, σ^2. Now let's us define new R.V.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \ldots$ The probability that the arithmetic mean of $X_1, X_2, \ldots, X_n$ differs from its expected value more than ϵ approaches zero as $n \to \infty$:

$$\lim_{n\to\infty} p\left(\left|\frac{S_n}{n} - E\left[\frac{S_n}{n}\right]\right| \geq \epsilon\right) = \lim_{n\to\infty} p\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Uniform distribution

- Uniform P.D.F. is defined for the C.R.V. and given by

$$f(x;a,b)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{otherwise}\end{cases}$$

- So, we say that x is equally likely to be found inside our interval of interest (a,b). The mean and variance:

$$E[x]=\int_a^b \frac{x}{b-a}dx=\frac{1}{2}(a+b)$$

$$V[x]=\int_a^b \left(x-\frac{1}{2}(a+b)\right)^2\frac{1}{b-a}dx=\frac{1}{12}(b-a)^2$$

- There is a very important application of the uniform model related to the fact that for any R.V. x with P.D.F. $f(x)$ we can easily find transformation to a new variable that is uniform.

Uniform distribution

- If we call the new transformed variable y, the transformation rule is simply related with calculating the C.D.F. Cool!

$$x \to y: y = F(x)$$

- Remember, x – any P.D.F., y uniform P.D.F. Next, for any C.D.F. the following is true

$$\frac{dy}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(x')dx' = f(x)$$

- And using the rule for change of variables

$$g(y) = f(x)\left|\frac{dx}{dy}\right| = f(x)\left|\frac{dy}{dx}\right|^{-1} = 1, \qquad (0 \le y \le 1)$$

- This is the fundamental rule of random number generator programs that can give us a list of numbers with any distribution. We are going to look at this in more detail.

Exponential distribution

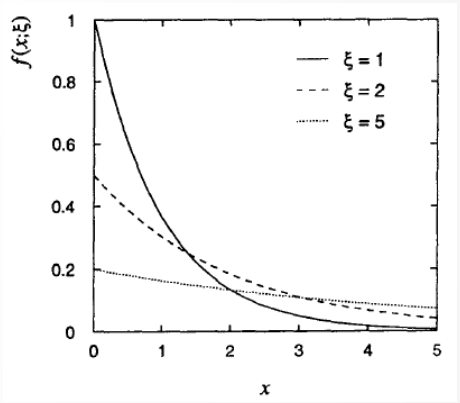
- This model is used for C.R.V. $x: 0 \leq x \leq \infty$ and is defined by

$$f(x;\xi) = \frac{1}{\xi} e^{-\frac{x}{\xi}}$$

- The model depends on a single parameter θ. The mean and variance are as follow:

$$E[x] = \frac{1}{\xi}\int_0^{\infty} x e^{-\frac{x}{\xi}} dx = \xi$$

$$V[x] = \frac{1}{\xi}\int_0^{\infty} (x-\xi)^2 e^{-\frac{x}{\xi}} dx = \xi^2$$



- The decay time of an unstable particle measured in its rest frame follows the exponential distribution. The parameter of the distribution is then interpreted as the mean lifetime.

The Normal Distribution

- This is definitely one of the most fundamental PDF with great significance in statistics

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$$

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2}\, dv$$

- We can also introduce the standardised variable corresponding to X

$$Z = \frac{X-\mu}{\sigma}, \mu_Z = 0, \sigma_Z = 1$$

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$$

$$F(z) = P(Z \leq z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-v^2/2}\, dv = \frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{z} e^{-v^2/2}\, dv$$

The Normal Distribution

- We then call the Z the standard score and the distribution function $F(Z)$ can be related to error function (tabulated) $\mathrm{erf}(z)$

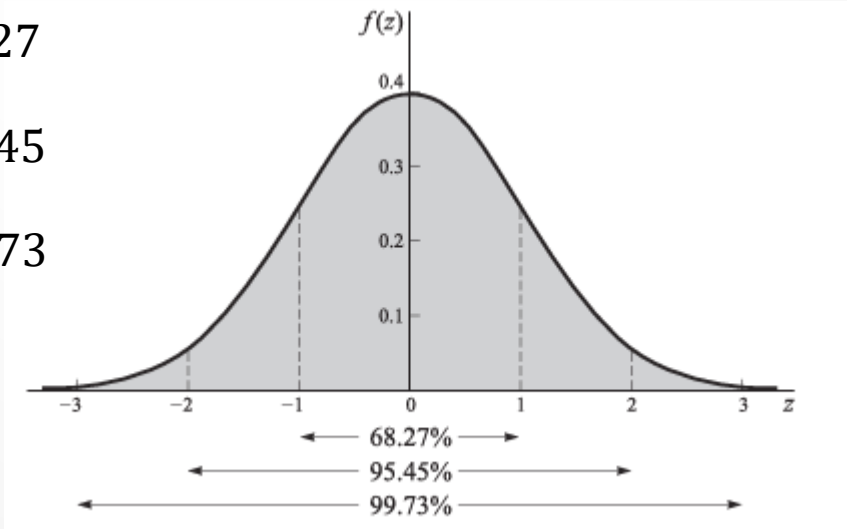
$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{z} e^{-u^2} dv$$

$$F(z) = \frac{1}{2}\left[1 + erf\left(\frac{z}{\sqrt{2}}\right)\right]$$

$$p(-1 < z < 1) = 0.6827$$

$$p(-2 < z < 2) = 0.9545$$

$$p(-3 < z < 3) = 0.9973$$



The Normal Distribution

- The properties of the Gaussian distribution

Mean	μ
Variance	σ^2
Standard deviation	σ
Coefficient of skewness	$\alpha_3 = 0$
Coefficient of kurtosis	$\alpha_4 = 3$

- Relation between binomial and normal distribution

$$Z = \frac{X - np}{\sqrt{npq}}$$

$$\lim_{n\to\infty} p\left(a \le \frac{X - np}{\sqrt{npq}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-v^2/2}\, dv$$

- We say, that the variable $\frac{X-np}{\sqrt{npq}}$ is **asymptotically normal**!