Discrete distributions

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2019-09-20

Discrete distributions of probabilities

The expression ***discrete distribution*** denotes probability distribution of variables that only take discrete values (by opposition to continuous distributions).

Notes:

- In probabilities, the observed variable (x) usually represents the number of successes of a series of tests, or the counts of some observation. In such cases, its values are natural numbers ($x \in \mathbb{N}$).
- The probability $\mathrm{P}(x)$ takes real values comprised between 0 and 1, but its distribution is said *discrete¨since it is only defined fora set of discrete values of X. It is generally represented by a step function.

Geometric distribution

Application: waiting time until the first appeearance of an event in a Bernoulli schema.

Examples:

- In a series of dices rollings, count the number rolls (x) before the first occurrence of a 6 (this occurrence itself is not taken into account).
- Length of a DNA sequence before the first occurrence of a cytosine.

Mass function of the geometric distribution

The ***Probability Mass Function*** (***PMF***) indicates the probability to observe a particular result.

For the geometric distribution, it indicates the probability to observe exactly x failures before the first success, in a series of independent trials with a probability of success p.

$$\mathrm{P}(X = x) = (1 - p)^x \cdot p$$

Justification:

- The probability of failure for the first trial is $q = 1 - p$ (complementary events).
- Bernoulli schema → the trials are independent → the probability of the series is the product of probabilities of its successive outcomes.
- One thus computes the product of probabilities of the x initial

Geometric PMF

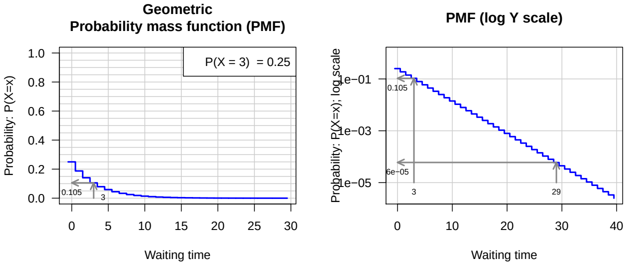


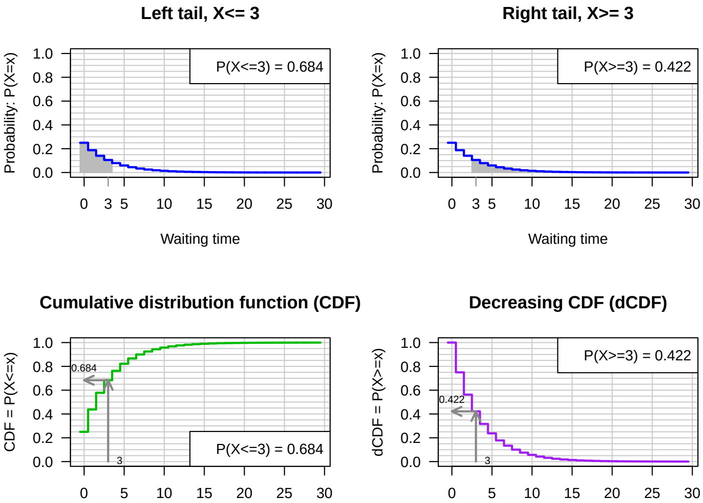
Figure 1: **Fonction de masse de la loi géométrique**. Gauche: ordonnée en échelle logarithmique.

Distribution tails and cumulative distribution function

The ***tails*** of a distribution are the areas comprised under the density curve up to a given value (***left tail***) or staring from a given value (***right queue***).

- The ***right queue*** indicates the probability to observe a result (X) **smaller than or equal to** a given value (x): $\mathrm{P}(X \leq x)$.
  - **Definition:** the ***Cumulative Density Function*** (***CDF***) $\mathrm{P}(X \leq x)$ indicates the probability for a random variable X to take a value smaller than or equal to a given value (x). It corresponds to the left tail of the distribution (including the x value).
- The **left tail** of a distribution indicates the probability to observe a result **higher than or equal to** a given value: $\mathrm{P}(X \geq x)$.
  - Note: in the next chapters we will use the [illegible] of the right tail

Distribution tails and cumulative distribution function



Binomial distribution

The **binomial distribution** indicates the probability to observe a given number of successes (x) in a series of n independent trials with constant success probability p (Bernoulli schema).

Binomial PMF

$$\mathrm{P}(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} = C_n^x p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n}$$

Binomial CDF

$$\mathrm{P}(X \geq x) = \sum_{i=x}^{n} P(X = i) = \sum_{i=x}^{n} C_n^i p^i (1-p)^{n-i}$$

Properties

i-shaped binomial distribution

The binomial distribution can take various shapes depending on the values of its parameters (success probability p, and number of trials n).

When the expectation $(p \cdot n)$ is very small, the binomial distribution is monotonously decreasing and is qualified of ***i*-shaped**.

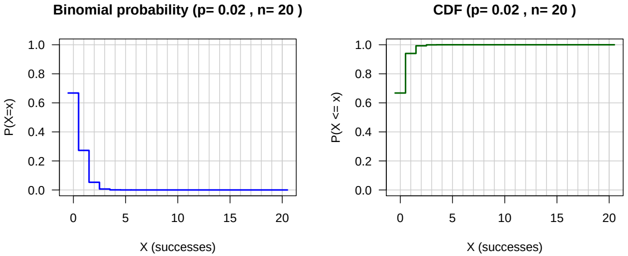


Figure 3: Distribution binomiale en forme de i.

Asymmetric bell-shaped binomial distribution

When the probability is relatively high but still lower than 0.5, the distribution takes the shape of an asymmetric bell.

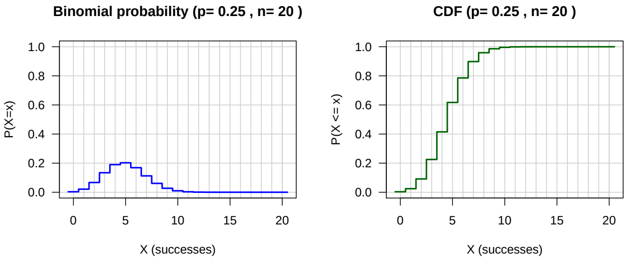


Figure 4: Distribution binomiale en forme de cloche asymétrique.

Symmetric bell-shaped binomial

When the success probability p is exactly 0.5, the binomial distribution takes the shape of a symmetrical bell.

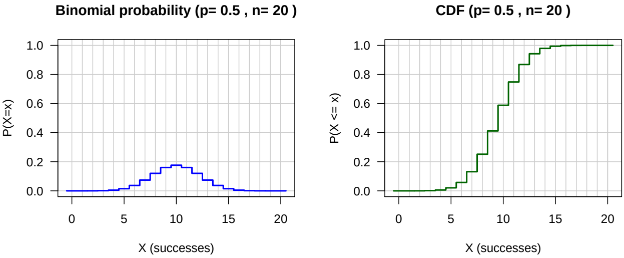


Figure 5: Distribution binomiale en forme de cloche symétrique (p=0.5).

j-shaped binomial distribution

Then the success probability is close to 1, the distirbution is monotonously increasing and is qualified of ****j*-shaped distribution.

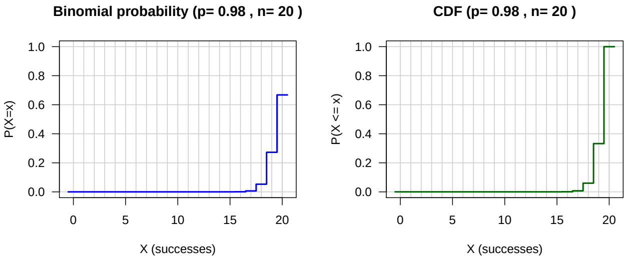


Figure 6: Distribution binomiale en forme de j.

Examples of applications of the binomial

1. **Dices**: number of 6 observed during a series of 10 dice rolls
2. **Sequence alignment**: number of identities between two sequences alignmed without gap and with an arhbitrary offset.
3. **Motif analysis**: number of occurrences of a given motif in a genome.

Note: the binomial assumes a Bernoulli schema. Forexamples 2 and 3 this amounts to consider that nucleotides are concatenated in an independent way, which is quite unrealistic.

Poisson law

La loi de Poisson décrit la probabilité du nombre d'occurrences d'un événement pendant un intervalle de temps fixé, en supposant que le nombre moyen d'événements par unité de temps est constant, et que les événements sont indépendants (les réalisations précédentes n'affectement pas la probabilité des occurrences suivantes).

Fonction de masse

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- x est le nombre d'événements observés;
- λ (lettre grècque "lambda") représente l'espérance, autrement dit la moyenne attendue pour le nombre d'événements;
- e est la base de l'exponentielle ($e = 2.718$).

Propriétés de la distribution de Poisson

- **Espérance** (nombre de succès attendus au hasard): $< X >= \lambda$ (par construction)
- **variance**: $\sigma^2 = lambda$ (***variance égale à la moyenne!***)
- **Ecart-type**: $\sigma = \sqrt{\lambda}$

Application : mutagenèse

- On soumet une population à un mutagène (agent chimique, irratiations). Chaque individu subit un certain nombre de mutations.
- En tenant compte de la dose de mutagène (temps d'exposition, intensité/concentration), on peut estimer empiriquement le nombre moyen de mutations par individu (*espérance*, λ).
- La loi de Poisson peut être utilisée pour décrire la probabilité d'observer un nombre donné de mutations ($x = 0, 1, 2, ...$).

Expérience historique de Luria-Delbruck (1943)

En 1943, Salvador Luria et Max Delbruck démontrent que les mutations ne sont pas induites par l'agent mutagène, (Luria & Delbruck, 1943, Genetics 28:491–511).

Convergence de la loi binomiale vers la Poisson

A FAIRE

Exercices

- html
- pdf
- Rmd