

Distributions discrètes

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-11-16

Éléments de théorie

Exercices

Eléments de théorie

Distribution de probabilité discrète

On parle de ***distribution discrète*** pour désigner la distribution de probabilité de variables ne pouvant prendre que des valeurs discrètes (par opposition aux distributions continues).

Notes:

- ▶ En probabilités la variable observée (x) représente généralement le nombre de succès d'une série d'observations. Elle prend donc généralement des valeurs naturelles (entières et positives).
- ▶ La probabilité $P(x)$ prend des valeurs réelles entre 0 et 1, mais sa distribution est discrète puisqu'elle n'est définie que pour des valeurs discrètes de x . On la représente généralement par une fonction en escalier.

Distribution géométrique

Application: temps d'attente jusqu'à la première réalisation d'un évènement au cours d'un schéma de Bernoulli.

Exemples:

- ▶ Comptage du nombre de jets d'un dé (x) qui précèdent la première occurrence d'un 6 (l'occurrence n'est pas incluse dans le compte).
- ▶ Longueur d'une séquence d'ADN avant la première occurrence d'une cytosine

Fonction de masse de probabilité géométrique

La **fonction de masse de probabilité** (**Probability Mass Function, PMF**) indique la probabilité d'observer une valeur particulière pour une distribution discrète.

Pour la distribution discrète, elle indique la probabilité d'observer exactement x échecs avant le premier succès, au cours d'une série d'essais indépendants avec une probabilité de succès p .

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

Justification:

- ▶ Probabilité d'échec pour un essai = $q = 1 - p$ (événements complémentaires)
- ▶ Schéma de Bernoulli → les essais sont indépendants → probabilité de la série est le produit des probabilités des résultats successifs.

Fonction de masse de probabilité géométrique

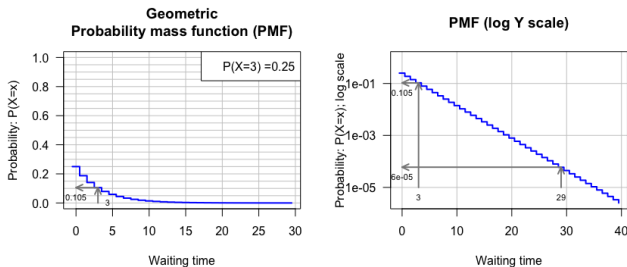


Figure 1: ****Probabilité de masse de la loi géométrique****. Gauche: ordonnée en échelle logarithmique.

Queues de distribution et fonction de répartition

Les queues de la distribution sont les aires comprises sous la courbe de densité jusqu'à une certaine valeur (**queue gauche**) ou à partir d'une certaine valeur (**queue droite**).

- ▶ La **queue droite** indique la probabilité d'obtenir un résultat (X) **inférieur ou égal** à une certaine valeur (x): $P(X \leq x)$.
 - ▶ **Définition:** la **fonction de répartition** (**Cumulative Density Function, CDF**) $P(X \leq x)$ indique la probabilité qu'une variable aléatoire X prenne une valeur inférieure ou égale à une valeur donnée (x). Elle correspond à la queue gauche (en incluant la valeur x considérée).
- ▶ La **queue gauche** d'une distribution indique la probabilité d'observer un résultat **supérieur ou égal** à une certaine valeur: $P(X \geq x)$.
 - ▶ **Note:** nous verrons ultérieurement l'utilisation de la **queue droite de différentes distributions en tant que probabilité critique** (P value***), dans le cadre de tests

Queues de distribution et fonction de répartition

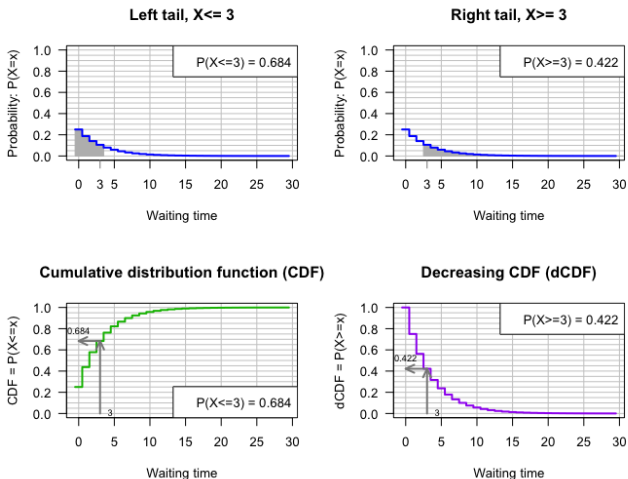


Figure 2: **Queues et fonction de répartition de probabilité

Distribution binomiale

La **distribution binomiale** indique la probabilité d'observer un certain nombre (x) de succès au cours d'une série de n essais indépendants avec une probabilité de succès p constante (schéma de Bernoulli).

Fonction de masse de probabilité binomiale

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

Fonction de répartition binomiale

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

Propriétés

Distribution binomiale en i

La distribution binomiale peut prendre différentes formes selon les valeurs des paramètres (probabilité de succès p , et nombre d'essais n).

Quand la probabilité de succès (p) est très faible par rapport au nombre d'essais (n), la distribution prend une forme de i .

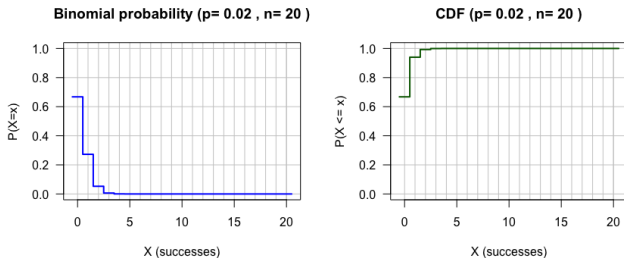


Figure 3: Distribution binomiale en forme de i .

Distribution binomiale en cloche asymétrique

Quand la probabilité de succès relativement élevée mais inférieure à 0.5, la distribution prend une forme en cloche asymétrique.

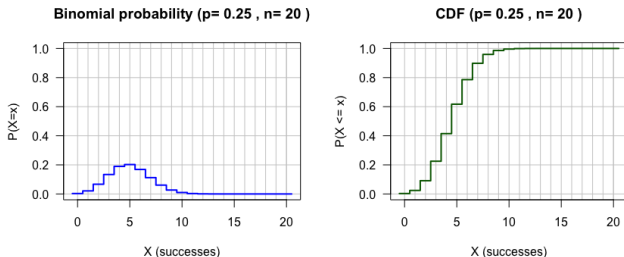


Figure 4: Distribution binomiale en forme de cloche asymétrique.

Distribution binomiale en cloche symétrique

Quand la probabilité de succès vaut 0.5, la distribution prend une forme en cloche symétrique.

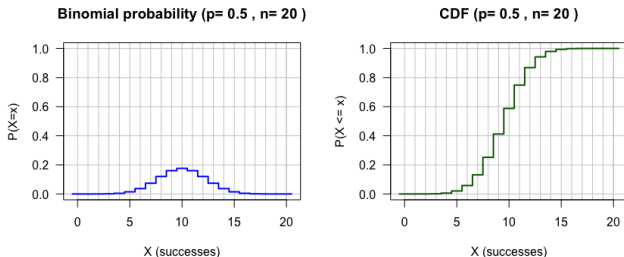


Figure 5: Distribution binomiale en forme de cloche symétrique ($p=0.5$).

Distribution binomiale en j

Quand la probabilité de succès est proche de 1, la distribution prend une forme en j .

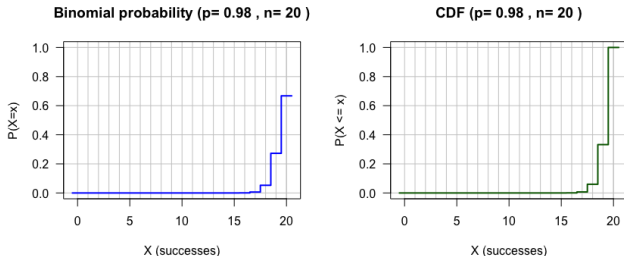


Figure 6: Distribution binomiale en forme de j .

Exemples d'applications de la binomiale

1. **Jeu de dés:** nombre de 6 observés lors d'une série de 10 tirages.
2. **Alignement de séquences:** nombre d'identités entre deux séquences alignées sans gap.
3. **Analyse de motifs:** nombre d'occurrences d'un motif dans un génome.

Note: le recours à la binomiale présuppose un modèle de Bernoulli. Pour les exemples 2 et 3 ceci revient à considérer que les nucléotides se succèdent de façon indépendante, ce qui est assez peu réaliste.

Exercices

Exercice: probabilité des longueurs d'ORF

On détecte les cadres ouverts de lecture (*open reading frames*, *ORF*) d'un génome en identifiant toutes les séquences de taille multiple de 3 comprises entre un start (ATG) et un stop (TAA, TAG ou TGA).

1. Sur base des fréquences génomiques de trinuécléotides, calculer la probabilité de trouver à une position donnée du génome un ORF d'au moins 100 codons.
2. Sachant que le génome fait 12 Mb, quel est le nombre attendu d'ORF d'au moins 100 codons ?

sequence	frequency	occurrences
AAA	0.0394	478708
ATG	0.0183	221902
TAA	0.0224	272041
TAG	0.0129	156668
TGA	0.0201	244627

Exercice: probabilité d'un motif avec erreurs

On recherche dans un génome les occurrences du motif GATAAG (où W signifie “A ou T”) en admettant un certain nombre de substitutions. En supposant que les nucléotides sont indépendants et équiprobables, quelle est la probabilité de trouver à une position du génome:

1. Une instance exacte du motif (aucune substitution) ?
2. Une séquence ne présentant aucune correspondance avec le motif (6 substitutions) ?
3. Une instance avec exactement 1 substitution ?
4. Une instance avec au plus 2 substitutions ?

Exercice: alignement de lectures NGS

Au terme d'un séquençage de type "Next Generation Sequencing" (NGS), on dispose d'une librairie de $N = 10^6$ lectures courtes. On aligne la librairie sur le génome de référence, dont la somme des chromosomes fait $G = 10^9$ paires de bases, en utilisant un algorithme d'alignement sans gap.

Calculez la distribution de probabilité du nombre de correspondances en fonction de la longueur des lectures (k).

Exercice: sites de restriction

Dans un génome bactérien de 4 Mb avec une composition de 50% de G+C, on observe 130 occurrences de l'hexanucléotide GGCGCC. On suppose un schéma de Bernoulli et une composition équiprobable de nucléotides.

1. Quelle est la probabilité d'observer une occurrence de GGCGCC à une position donnée du génome ?
2. Combien d'occurrences s'attend-on à trouver dans l'ensemble du génome ?
3. Quelle serait la probabilité d'observer un nombre aussi faible d'occurrences (130 ou moins) si l'on générerait une séquence aléatoire selon le modèle de Bernoulli avec nucléotides équiprobables ?
4. Comment peut-on interpréter cette sous-représentation de l'hexanucléotide GGCGCC du point de vue biologique ?