Université Frères Mentouri Constantine 1
Faculté des sciences de la nature et de la vie

Département de Biologie Appliquée

Licence : Biotechnologie microbienne

Année Universitaire : 2019/2020

Cours de Biostatistique

Dr. Habiba BOUHALLOUF

3. Lois de distribution

Une distribution empirique ou distribution des fréquences, est un tableau qui associe des classes de valeurs observées et obtenues lors d'une expérience. Ce tableau de valeurs est modélisé en théorie des probabilités par une loi. Les modèles de distribution les plus utilisés sont les distributions régies par la loi binomiale, la loi de Poisson et la loi normale.

3.1 Loi binomiale

On utilise la loi binomiale lorsqu'on désire connaître :

- La probabilité de k succès au bout de n tentatives sachant la probabilité P de gagner à chacune des tentatives (situation de jeu de hasard).

- La probabilité d'observer k individus possédant une caractéristique donnée dans un échantillon de n individus tirés d'une population où la probabilité P de la caractéristique est connue.

La difficulté de la loi binomiale n'est pas d'effectuer les calculs, mais de savoir poser le problème, il faut donc bien connaître la définition de ses termes.

En pratique biologique et médicale, on utilise la loi binomiale pour étudier la distribution d'une variable qualitative binaire dans un échantillon de sujet. La plus utilisée en épidémiologie est la variable *(malade / non malade)*.

3.1.1 Cas d'utilisation de la loi binomiale

La loi binomiale s'applique donc quand il y a un nombre défini de répétitions d'une même expérience dans les mêmes conditions :

- Les variables discrètes.
- Deux résultats possibles.
- Essais répétitifs dans les mêmes conditions.
- Essais indépendants.

3.1.2 Jeu de hasard et étude d'un échantillon

TABLE 3.1 – Différence entre le jeu de hasard et l'étude d'un échantillon

	Jeu de hasard	**Etude d'un échantillon**
r	nombre d'événements gagnants (succès)	nombre de sujets dans l'échantillon
n	nombre de tentatives	taille de l'échantillon
P	probabilité de gagner à chaque tentative	proportion de sujets dans la population

La loi binomiale a deux résultats possibles appellés *succès* et *échec*. Le succès est le résultat pour lequel nous souhaitons déterminer la distribution de probabilité.

Definition 3.1.1 Pour calculer la probabilité d'obtenir r succès en n essai il faut utiliser la formule suivante :

$$P(x=k) = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r} \tag{3.1}$$

où :

- n correspond au nombre d'objet de l'ensemble (taille de l'échabtillon).
- x est la variable qui peut prendre la valeur r.
- P est la probabilité d'observer la valeur r.
- p est la probabilité de succès à un essai (tentative).

Rappel :
On rappelle que le symbole "!" signifie factoriel et se calcule comme suit :

n !=n(n-1)(n-2)(n-3). 3.2.1

Exemple : $5! = 5.4.3.2.1 = 120$

3.1.3 Condition d'utilisation de la loi binomiale

Pour pouvoir utiliser la loi binomiale, il faut que :

- La variable étudiée soit de type binaire.

- l'échantillon soit tiré au sort (et les tentatives soient indépendantes s'il s'agit du jeu de hasard).

- Les individus de la population étudiée aient la même chance d'être tirés au sort (et chaque événement ait la même probabilité de succès au cours d'une tentative).

- la taille de l'échantillon n soit très petit par rapport à la taille de la population N, En fait elle doit vérifier la condition : $\frac{n}{N} < 10\%$.

3.1.4 Fonction de répartition de la loi binomiale

$$P(x < k) = P(0) + P(1) + ... + P(k-1) = 1 - P(x \geq k) \tag{3.2}$$
$$P(k \leq k) = P(0) + P(1) + ... + P(k-1) + P(k) = 1 - P(x > k) \tag{3.3}$$
$$P(x > k) = P(k+1) + ... + P(k_n) = 1 - P(x \leq k) \tag{3.4}$$
$$P(x \geq k) = P(k) + P(k+1) + ... + P(k_n) = 1 - P(x < k) \tag{3.5}$$

3.2 Loi de Poisson

Parfois, il est impossible d'utiliser la loi binomiale. Par exemple si p est très petit, donc un n très grand ..., on utilise plutôt *la loi de Poisson*.
La distribution de Poisson décrit la distribution de probabilités lorsque les événements étudiés ont lieux dans une fouchette de temps délimité ou dans un lieu défini.

3.2.1 Cas d'utilisation de la loi de Poisson

La loi de Poisson peut être utilisés dans les cas suivants :

- Variables discrètes.

- Nombre d'occurence d'un événement.

- Espace ou temps défini.

Definition 3.2.1 Soit μ le nombre moyen d'événements observés dans la population pendant une période donnée, x la variable repésentant le nombre d'individus ayant subi l'événement observé et r une valeur que peut prendre la variable x. On définit la loi de poisson par la formule suivante :

$$P(x = r) = \frac{e^{-\mu}\mu^r}{r!} \tag{3.6}$$

3.2.2 Condition d'utilisation de la loi de Poisson

Pour pouvoir utiliser la loi de Poisson, on doit avoir :

- des événements dénombrables.
- des événements indépendants les uns des autres.
- des événements rares dont la probabilie de survenue est inférieur à $0,05$. (sinon on utilise la loi binomiale).

3.2.3 Fonction de répartition de la loi de Poisson

$$P(x < k) = P(0) + P(1) + ... + P(k-1) \tag{3.7}$$
$$P(k \leq k) = P(0) + P(1) + ... + P(k-1) + P(k) \tag{3.8}$$
$$P(x > k) = 1 - P(x \leq k) \tag{3.9}$$
$$P(x \geq k) = 1 - P(x < k) \tag{3.10}$$

3.3 Loi normale

Contrairement à la loi de Poisson et la loi binomiale qui étaient des distributions de probabilité discrètes, la distribution normale (la plus importante des lois utilisées en statistique) est une distribution de probabilité continue. La variable utilisée est quantitative continue, c'est à dire qu'elle peut prendre un nombre indéfini de valeurs. La coube normale a la particularité d'être symétrique, elle s'appelle *courbe en cloche* (voir Figure 3.1).

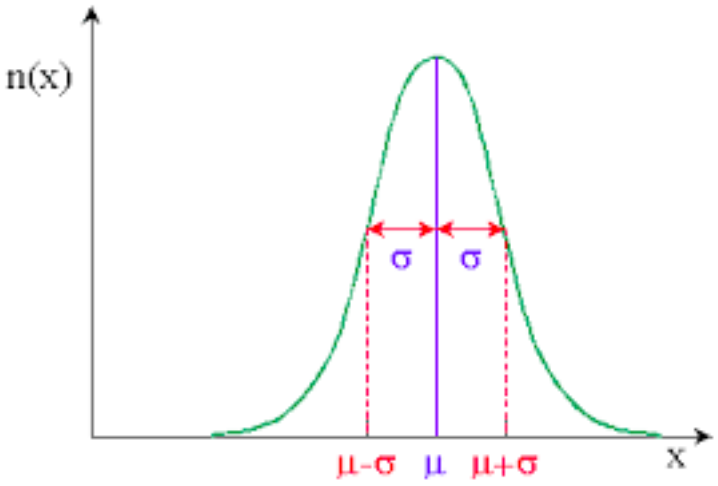

FIGURE 3.1 – Courbe en cloche

3.3.1 Propriétés

Voir la Figure 3.2.

— La loi normale est centrée autour de la moyenne (qui elle même la médiane pour la loi normale).

— L'aire contenue entre les deux points d'inflexion de la courbe mesure la probabilité pour que les valeurs de x soient comprise entre $(-\sigma)$ et $(+\sigma)$ (σ est l'écart type) autour de la moyenne. Cette probabilité est de $68,26\%$.

— L'aire comprise entre (-2σ) et $(+2\sigma)$ représente $95,44\%$ autour de l'aire totale.

— L'aire comprise entre (-3σ) et $(+3\sigma)$ représente $99,74\%$ des valeurs de x.

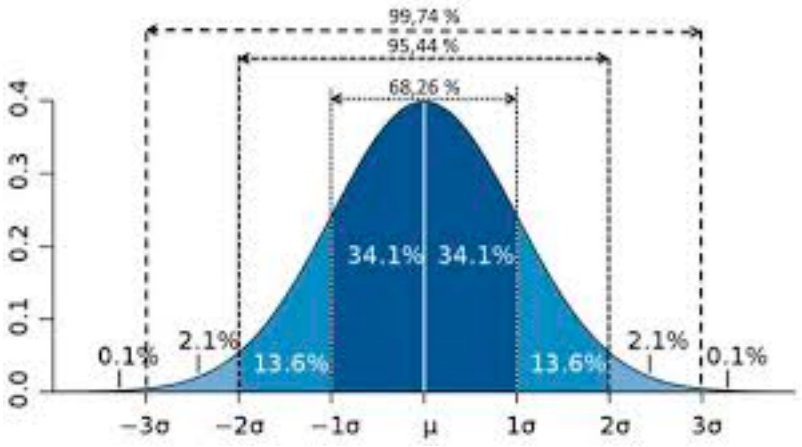

FIGURE 3.2 – La loi normale

3.3.2 Loi normale centée réduite

Toute variable suivant une loi normale est représentée par sa courbe en cloche (Figure 3.1) dont sa position et sa forme dépendent de la moyenne et l'écart type correspondant. Dans la (Figure 3.3), on ne peut pas établir une table de toutes les distributions normales possibles, mais elles peuvent se ramener en une seule distribution en effectuant une transformation de variable.

3.3.3 variable centrée Z

Pour pouvoir ramener plusieurs distributions en une seule distribution :

1. Posons une variable $x^{'}$ telle que :

$$x^{'} = x - \mu.$$

où μ represente la moyenne. Cela aboutit à centrer la distribution autour de la valeur 0 quelque soit la variable x, (Figure 3.4).

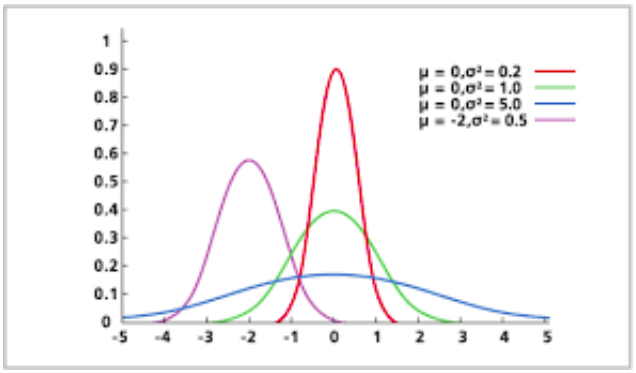

FIGURE 3.3 – Plusieurs courbes en cloches de loi normale

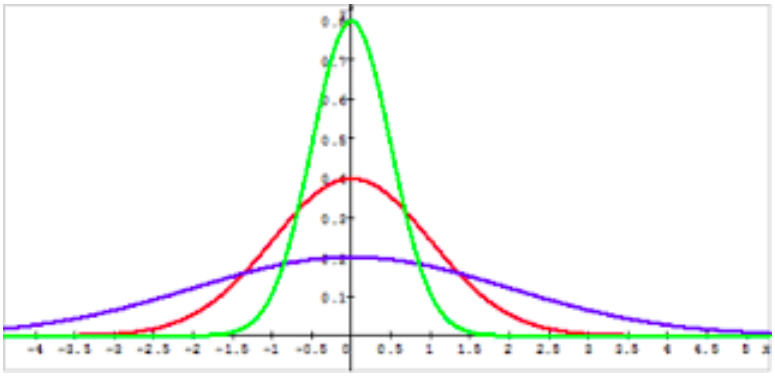

FIGURE 3.4 – Plusieurs courbes de loi normale centrée

2. Posons maintenant une nouvelle variable qu'on appelle Z, telle que :

$$Z = \frac{x - \mu}{\sigma}$$

où σ est l'écart type de la variable x, ce qui donne ainsi un écart type égale à 1 pour la variable Z, (Figure 3.5).

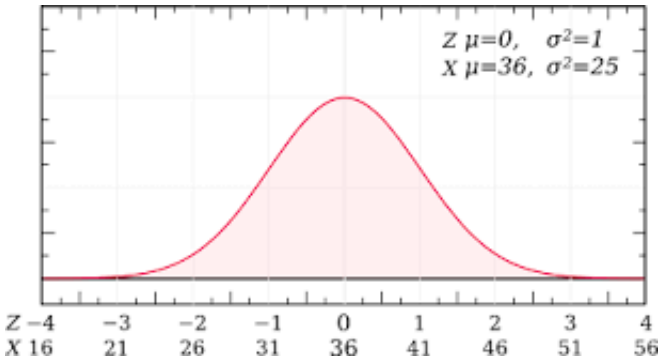

FIGURE 3.5 – Loi normale centrée réduite Z

Fiallement, on aboutit une seule courbe dont la variable est Z, la distribution est centrée autour de 0 et pour un écart type vaut 1.

3.3.4 Densité de probabilité

1. **Pour la loi normale**

Definition 3.3.1 s'appelle aussi *la loi de GAUSS*. La densité de probabilité de chaque valeur de la variable continue x est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3.11}$$

où :

— x est une valeur de la variable continue.

— μ est la moyenne de toutes les valeurs de la variable normale.

— σ est l'écart type des valeurs de la loi normale.

2. **Pour la loi normale centrée réduite** Z

Definition 3.3.2 la densité de probabilité de la valeur Z d'une variable continue est :

$$f(Z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{Z^2}{2}\right] \tag{3.12}$$

III

Estimations et Tests Statistiques

6. Théorie statistique de l'estimation

Du point de vue pratique, il est souvent très important de pouvoir obtenir de l'information sur la population à partir des échantillons. De tel problème se trouve dans la décision statistique, qui utilise le principe de la théorie d'échantillonnage comme le problème d'estimation des paramètres d'une population (moyenne, variance, pourcentage, ...) à partir des statistiques d'échantillonnage correspondantes.
Faire une estimamtion, c'est tenter de définir les paramètres d'une population à partir des paramètres observés sur un échantillon.
Lorsqu'on observe un paramètre sur un échantillon, on pressent :

1. que la valeur observée a fort peu de chances d'être exactement la valeur inconnue de la population.

2. que cette valeur est néanmoins assez proche de la valeur inconnue si notre échantillon est représentatif.

3. qu'en répétant l'échantillonnage, on trouverait d'autres valeurs, toutes assez proches les unes des autres.

Ces trois hypothèses sont une sorte de pari. Nous parions que la valeur observée est proche de la valeur exacte. Mais il faut préciser ce que l'on entend par "proche".
Le but de l'estimation en statistique est de calculer les bornes qui permettent de situer avec une confiance suffisamment grande où se trouve la valeur inconnue du paramètre dans la population. Une estimation aboutit donc à calculer ce qu'on nomme "intervalle de confiance". Ce terme est parfois appelé trivialement "fourchette d'estimation".
Le statisticien se sait donc incapable de connaître la vraie valeur, mais il en fournit modestement une estimation à l'aide de deux bornes.

Dr. Habiba BOUHALLOUF
Département de biologie appliquée, SNV, UFMC1

6.1 Définition

Soit une variable x à étudier : il s'agit d'otenir une approximation d'un certain paramètre θ de sa distribution (médiane, moyenne, variance, ...) à partir de n valeurs : x_1, x_2, ..., x_n de x.
En considérant x_1 : la réalisation d'une variable aléatoire X_1, x_2 : la réalisation d'une variable aléatoire X_2, ..., x_n : la réalisation d'une variable aléatoire X_n.
On dit que X_1, X_2, ..., X_n forment un échantillon de la variable X ayant la taille (effectif) n.

6.2 Estimation ponctuelle

Le terme estimation désigne aussi le résultat de procédé : on dira donc que t (la valeur calculée sur l'échantillon) est l'estimation *ponctuelle* de θ (la valeur théorique de la distribution), mais on dira aussi que t est un paramètre d'échantillon (estimant un paramètre de distribution).

6.2.1 Médiane d'échantillon

Une première estimation simple concerne la médiane. La médiane théorique d'une variable étudiée dans une population de N individus est située au milieu de la liste des valeurs individuelles classées par ordre croissant.
Donc, sur un échantillon de n valeurs classées par ordre croissant ($x_1 \leq x_2 \leq ..., \leq x_k \leq ..., \leq x_n$), la grandeur t est, par définition, la valeur centrale si le nombre des observations est impair, ou la demi-somme des deux valeurs centrale si le nombre des observation est pair :

$$t = x_{k+1} \qquad t = \frac{x_k + x_{k+1}}{2} \tag{6.1}$$

6.2.2 Moyenne d'échantillon

La moyenne théorique d'une variable étudiée dans une population de N individus s'obtenant par la formule $\mu = \frac{x_1 + x_2 + ... + x_N}{N}$. Sachant que la moyenne d'échantillon est : $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$. Ici $t = \bar{x}$, est une estimation de μ.

6.2.3 Variance d'échantillon

La variance théorique d'une variable étudiée dans une population de N individus :

$$\sigma_p^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ ... \ + (x_N - \mu)^2}{N}$$

mène à utiliser $t = \sigma_e^2$, comme estimation de la variance σ_p^2.

6.2.4 Estimateurs non biaisés

Si la moyenne d'une statistique d'échantillonnage est égale au paramètre correspondant de la population, on dit que la statistique est un *estimateur non biaisé* de ce paramètre.

Dans le cas contraire, on dit que l'on a un *estmateur biaisé*.

la moyenne : $moy(\bar{x}) = \mu \Longrightarrow \bar{x}$ est une estimation non biaisée.

la variance : $moy(\sigma_e^2) = \frac{N-1}{N}\sigma_p^2 \Longrightarrow \sigma_e^2$est une estimation biaisée.

où σ_p^2 est la variance de la population.

Remarque
En terme de propbabilité, on dira qu'une statistique est non biaisée si son espérance mathématique est égale à la valeur du paramètre de la population correspondant :

$$E(\bar{x}) = \mu \tag{6.2}$$
$$E(\sigma_e^2) = \sigma_p^2 \tag{6.3}$$

6.2.5 Estimateurs efficaces

Quand on désire estimer la moyenne, la distribution d'échantillonnage de deux statistiques ont la même espérance, la statistique qui a la variance la plus faible est appelée *"estimateur efficace"* de la moyenne, et l'autre statistique sera donc *"l'estimateur inefficace"*. Parfois, l'estimateur efficace est nomé *"meilleur estimateur"*.

6.3 Estimation par intervalle

Quand, dans une population, l'estimation d'un paramètre est donnée par un seul nombre, on dit que c'est une *"estimation ponctuelle"* du paramètre.
Quand on estime un paramètre d'une population donnée par deux nombres entre lesquels celui-ci peut varier, on dit que l'on a une *"estimation par intervalle"* de ce paramètre. Et on appelle l'erreur de précision d'un estimateur : *"confiance"* ou *"fiabilité"*.

6.3.1 Estimation d'une moyenne inconnue

a. Estimation d'une moyenne d'un échantillon

On considère que la population est nombreuse ($n \geq 30$) de moyenne μ et de l'écart-type σ_p relatif à un caractère quatitatif.
On désigne par $\bar{x}$, la moyenne d'un échantillon prélevé au hasard de la population.
D'après le théorème central limite, on démontre que $\bar{x}$ suit une loi normale d'esperance mathématique μ et de variance $\sigma^2 = \frac{{\sigma_p}^2}{n}$ lorsque la taille de l'échantillon est $n \geq 30$.

Definition 6.3.1 On peut exprimer $\bar{x}$ dans un intervalle comme suit :

$$\mu - t_\alpha \frac{\sigma_p}{\sqrt{n}} \leq \bar{x} \leq \mu + t_\alpha \frac{\sigma_p}{\sqrt{n}} \tag{6.4}$$

La probabilité pour que la moyenne $\bar{x}$ soit dans l'intervalle $I = \left[\mu - t_\alpha \frac{\sigma_p}{\sqrt{n}} \ , \ \mu + t_\alpha \frac{\sigma_p}{\sqrt{n}}\right]$ est :

$$P(I) = 1 - \alpha \tag{6.5}$$

Risque d'erreur α

Ici on appelle l'intervalle I, *intervalle de confiance*, $(1-\alpha)$ s'appelle *Seuil de confiance* et α, *risque d'erreur*.
t_α est une valeur donnée par **la table de la loi normale centrée réduite**.
D'après les propriétés de la loi normale, on choisit on général, le risque d'erreur $(\alpha = 5\%)$, et dans certain cas, on donne $(\alpha = 1\%)$:

1. pour $\alpha = 5\%$, on choisit $t_\alpha = 1.96$, et dans ce cas $P(I) = 0.95$.

2. pour $\alpha = 1\%$, on choisit $t_\alpha = 2.6$, et on donne $P(I) = 0.99$.

b. Estimation d'une moyenne d'une population

Le problème qui se pose généralement est d'estimer la moyenne μ de la population à partir des paramètres observés dans l'échantillon choisit au hasard, c-à-d : en fonction de $(\bar{x}, n, \sigma_e)$, où σ_e est l'écart-type de l'échantillon.

Definition 6.3.2 L'intervalle de confiance dans lequel on estime trouver la moyenne associée à la population est donnée par :

$$\bar{x} - t_\alpha \frac{\sigma_e}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_\alpha \frac{\sigma_e}{\sqrt{n-1}} \tag{6.6}$$

On donne :

$$\sigma_p{}^2 \approx \frac{n}{n-1}\,\sigma_e^2 \tag{6.7}$$

La quantité : $h = t_\alpha \frac{\sigma_e}{\sqrt{n-1}}$, s'appelle *"la précision de l'estimation"*.

6.3.2 Estimation d'un pourcentage inconnu

Lorsquòn a un pourcentage sur un échantillon, le problème est d'estimer le véritable pourcentage P inconnu de la population d'où est extrait l'échantillon.

a. Intervalle de confiance d'un pourcentage

Estimer la valeur du pourcentage inconnu de la population à partir d'une observation sur un seul échantillon, c'est estimer un intervalle dans lequel le pourcentage inconnu P à la plus grande probabilité de se trouver.

Definition 6.3.3 D'après le théorème central limite, il y a 95% de chances que le pourcentage P de la population se trouve compris dans l'intervalle :

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq P \leq p + 1.96\sqrt{\frac{p(1-p)}{n}} \tag{6.8}$$

$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}}\ ,\ p + 1.96\sqrt{\frac{p(1-p)}{n}}\right]$ est l'intervalle de confiance à 95% du pourcentage P de la population, où : p est le pourcentage calculé sur l'échantillon.