# Formules de probabilités et statistique

Probabilités et statistique pour la biologie (STAT1)

*Jacques van Helden*

*2019-01-11*

## Combinatoire

| Nom | Conditions | Formule |
|---|---|---|
| Permutations (factorielle) | | $\begin{cases} 0! = 1 \\ n! = n \cdot (n-1)! & \forall\, n \geq 1 \end{cases}$ |
| Arrangements | Sans remise, ordonné | $A_n^x = \frac{n!}{(n-x)!} = n \cdot (n-1) \cdot \ldots \cdot (n-x+1)$ |
| Combinaison (*choose*, *coefficient binomial*) | Sans remise, sans ordre | $\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$ |

## Concepts de probabilité

| Description | Conditions | Formule |
|---|---|---|
| Définition fréquentielle de la probabilité | | $P(A) = \lim_{n\to\infty} \frac{n_A}{n}$ |
| Probabilité de non-réalisation | | $P(\neg A) = 1 - P(A)$ |
| Probabilités conditionnelles | | $P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$<br>$P(B \mid A) = \frac{P(A \wedge B)}{P(A)}$ |
| Probabilité de $A$ ou $B$ | En général | $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ |
| Probabilité de $A$ ou $B$ | Evénements mutuellement exclusifs | $P(A \vee B) = P(A) + P(B)$ |
| Probabilité de $A$ et $B$ | En général | $P(A \wedge B) = P(A) \cdot P(B \mid A)$ |
| Probabilité de $A$ et $B$ | Evénements indépendants | $P(A \wedge B) = P(A) \cdot P(B)$ |
| Règle de Bayes | | $P(A \wedge B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$<br>$\implies P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$<br>$\implies P(B \mid A) = \frac{P(B) \cdot P(A \mid B)}{P(A)}$ |

# Distributions de probabilité discrètes

## Géométrique

- Conditions : nombre d'échecs avant le premier succès dans un schéma de Bernoulli
- Densité :
$$P(X = x) = (1 - p)^x p$$
- Répartition :
$$P(X \leq x) = \sum_{i=0}^{x} (1 - p)^i p$$
- Moyenne : $\mu_G = (1 - p)/p$
- Variance : $\sigma_G^2 = \frac{(1-p)}{p^2}$

## Binomiale

- Conditions : Nombre de succès au cours d'une série d'essais indépendants avec probabilité constante de succès (Schéma de Bernoulli)
- Densité :
$$P(X = x) = C_n^x p^x (1 - p)^{n-x}$$
- Répartition :
$$P(X \leq x = \sum_{i=0}^{x} C_n^i p^i (1 - p)^{n-i}$$
- Moyenne : $\mu_B = np$
- Variance : $\sigma_B^2 = np(1 - p)$
- Rapport moyenne/variance: $\sigma_B^2 < \mu_B$

## Poisson

- Conditions : nombre de succès observés au cours d'un intervalle de temps, en fonction du nombre attendu ($\lambda$)
- Application : approximation de la binomiale quand $n \to \infty, p \to 0$ et $\mu = np$ faible ($\mu_B \to \lambda$)
- Densité :
$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$
- Répartition :
$$P(X \leq x) = \sum_{i=0}^{x} \frac{e^{-\lambda} \lambda^i}{i!}$$
- Moyenne : $\mu_P = \lambda$
- Variance : $\sigma_P^2 = \lambda$
- Rapport moyenne/variance: $\sigma_P^2 = \mu_P$

### Hypergéométrique

- Conditions : Tirage non ordonné, sans remise dans un ensemble fini avec deux catégories.
- Exemple-type: urne avec boules de deux couleurs
- Densité :

$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

- Répartition :

$$P(X \leq x = \sum_{i=x}^{\min(k,m)} \frac{C_m^i C_n^{k-i}}{C_{m+n}^k}$$

- Moyenne : $\mu_H = k \cdot \frac{m}{m+n}$
- Variance : $\sigma_H^2 = \frac{k\frac{m}{N}(1-\frac{m}{N})(N-k)}{(N-1)}; N = m+n$

## Echantillonnage et estimation

- Les symboles grecs ($\mu$, $\sigma$) correspondent aux statistiques de population, les symboles romains ($\bar{x}$, $s$) aux statistiques d'échantillon.
- L'accent circonflexe ($\hat{}$) indique les estimateurs de paramètres de population calculés à partir de paramètres d'échantillons.

| Symbole | Description |
|---|---|
| $N$ | Taille (nombre d'individus) de la population. |
| $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ | Moyenne de la population. |
| $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$ | Variance de la population |
| $\sigma = \sqrt{\sigma^2}$ | Écart-type de la population |
| $n$ | Effectif (nombre d'individus) de l'échantillon. |
| $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ | Moyenne d'échantillon. |
| $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ | Variance de l'échantillon |
| $s = \sqrt{s^2}$ | Écart-type de l'échantillon |
| $\hat{\sigma^2} = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ | Estimateur non-biaisé de la variance de la population. |
| $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ | Estimateur non-biaisé de l'écart-type de la population. |
| $< \sigma_{\bar{X}} >= \frac{\hat{\sigma}}{\sqrt{n}}$ | Erreur standard: écart-type attendu sur la moyenne d'échantillon. |
| $\bar{x} \pm \frac{\hat{\sigma}}{\sqrt{(n)}} \cdot t_{1-\alpha/2}^{n-1}$ | Intervalle de confiance autour de la moyenne. |

## Test de comparaison de moyennes

| Symbole | Description |
|---|---|
| $\mu_1, \mu_2$ | Moyennes respectives des populations 1 et 2. |
| $\sigma_1, \sigma_2$ | Écarts-types respectifs des populations 1 et 2. |
| $n_1$, $n_2$ | Effectifs (nombre d'individus) des échantillons prélevés sur les populations 1 et 2. |
| $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ | Formule générale de la moyenne d'échantillon |
| $\bar{x}_1, \bar{x}_2$ | Moyennes d'échantillons. |
| $\delta = \mu_2 - \mu 1$ | Différence entre les moyennes des populations. |
| $d = \hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$ | $d =$ **Taille d'effet**: dans un test de comparaison de moyennes, il s'agit de la différence entre les moyennes d'échantillons, utilisée comme estimateur de $\delta$. |
| $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ | Formule générale de la variance d'échantillon |
| $s_1^2, s_2^2$ | Variances mesurées sur les échantillons. |
| $\hat{\sigma}_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$ | Écart-type groupé (*pooled standard deviation*), utilisé comme estimateur de l'écart-type commun des deux populations, en supposant leurs variances égales (hypothèse de travail d'homoscédasticité). |
| $\hat{\sigma}_\delta = \hat{\sigma}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ | Erreur standard sur la différence entre moyennes, en supposant que les populations ont la même variance (test de Student). |
| $t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ | Statistique $t$ de Student, $\nu = n_1 + n_2 - 2$ d.l. |
| $t_W = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$ | Statistique $t$ de Welch, $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$ d.l. |