Distributions et calculs de probabilités

1er bachelier de médecine vétérinaire

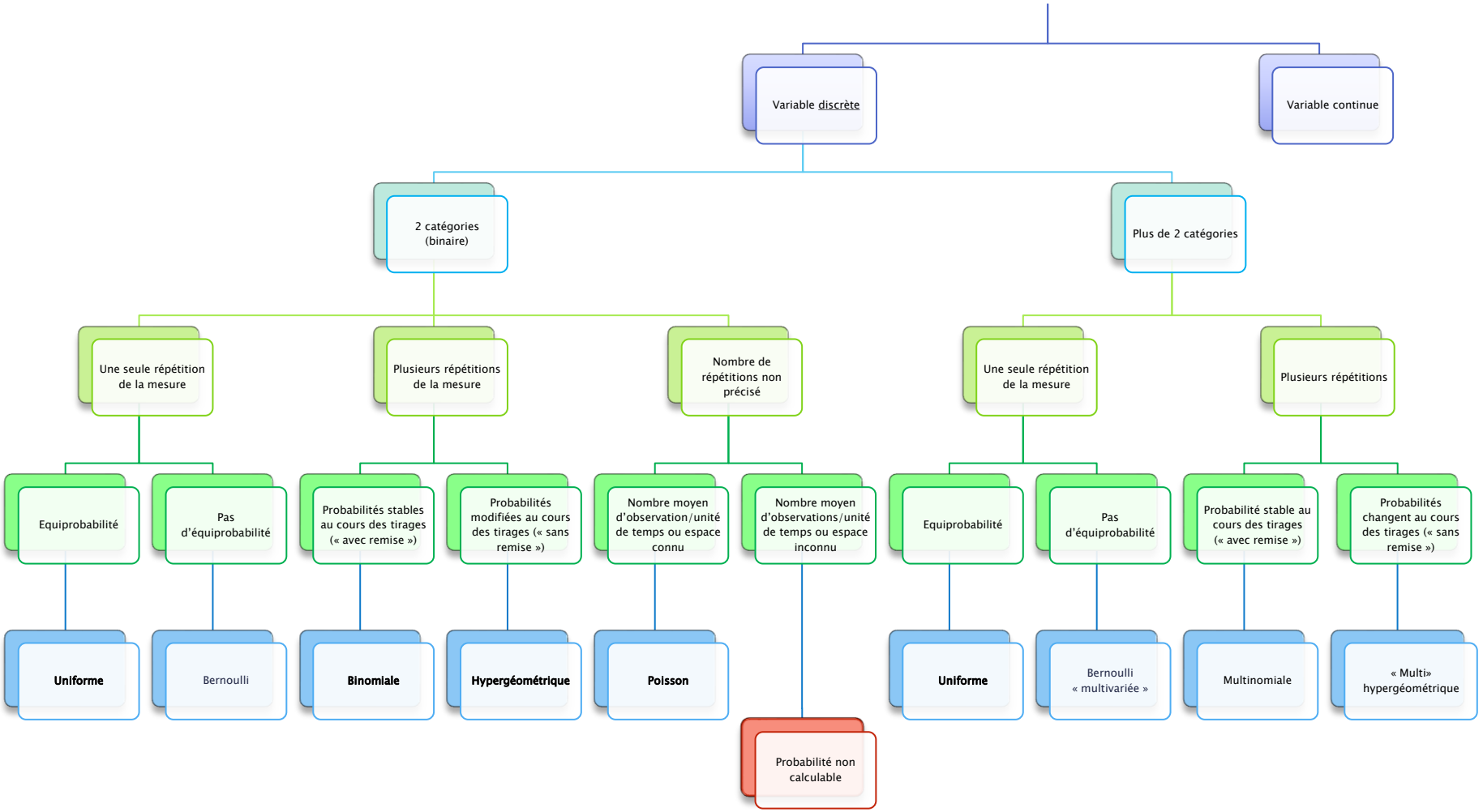
Année 2021-2022

Calcul de probabilité

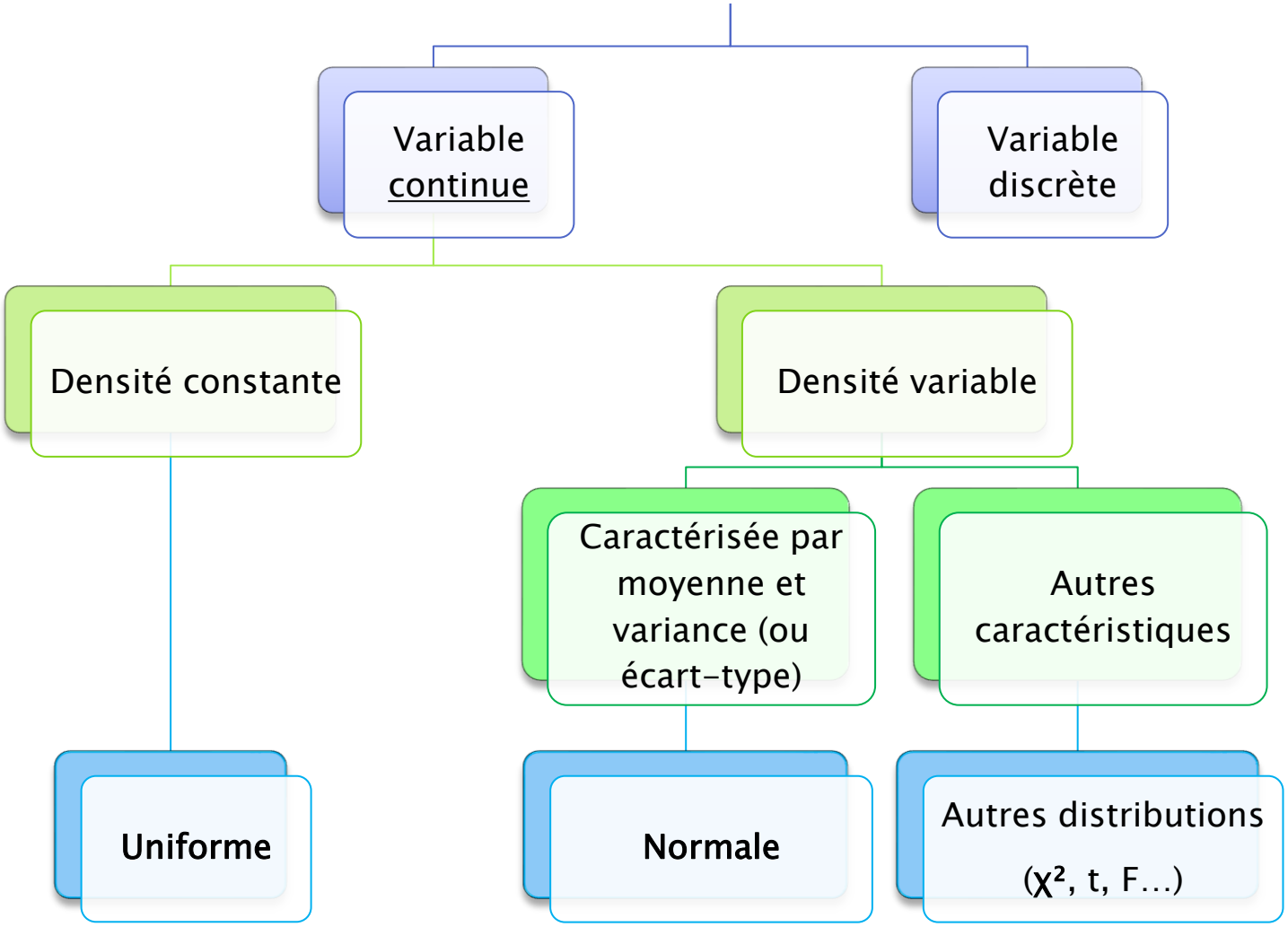
- Utilités:
  - Pour connaitre la probabilité P d'une situation donnée
  - Dans les tests d'hypothèse, où le choix d'accepter ou de refuser une hypothèse dépend la probabilité P associée à ce qui a été observé *(voir dias suivantes)*

- Comment calculer cette probabilité P?
  - Calcul dépendant de la variable aléatoire pour laquelle on calcule P
    - Un « arbre décisionnel » peut aider à identifier à identifier le type de variable puis sa distribution afin d'effectuer le calcul

Arbre décisionnel – Distributions discrètes



Arbre décisionnel – Distributions continues

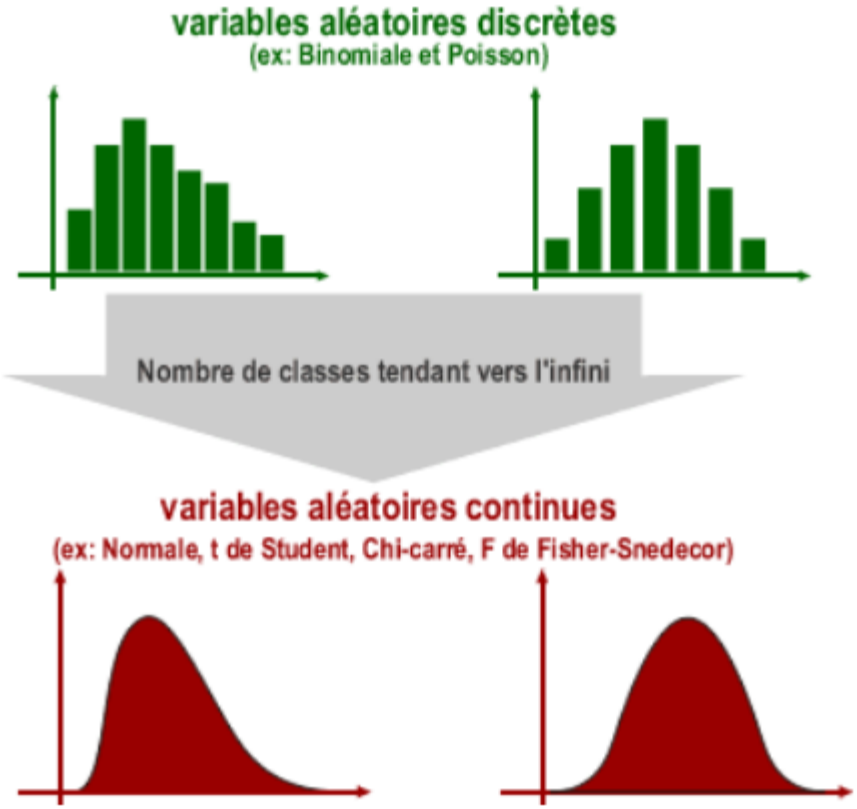


Arbre décisionnel

- 1er niveau : variable discrète ou continue?
  - Exemples de variables discrètes : statut sanitaire (malade – sain), statut vaccinal (vacciné – non vacciné), sexe (femelle – mâle), robe (chocolat, noir, doré), catégorie d'âge (chiot non sevré, chiot sevré, chien adulte entre 1 et 10 ans, chien âgé de plus de 10 ans), …
  - Exemples de variables continues : poids, taille, concentration d'un métabolite sanguin, pression artérielle, âge, temps de réaction, …

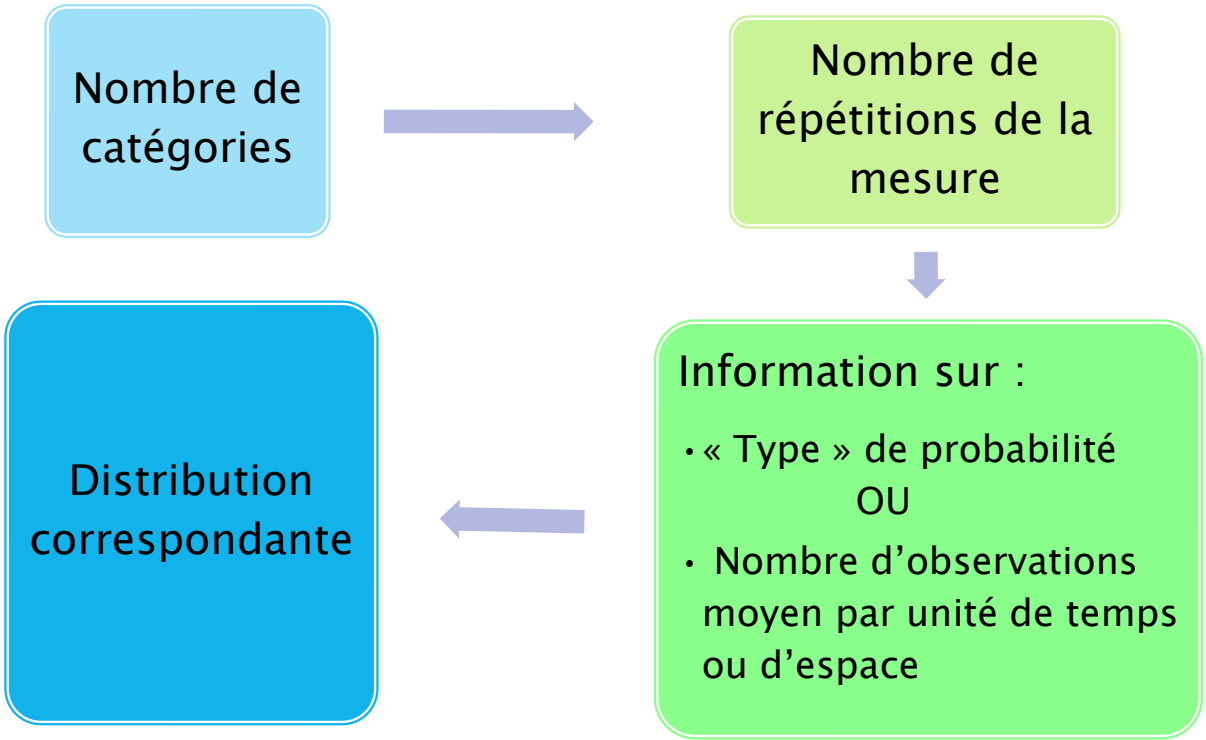
Variable discrète ou continue – remarque

- Si n est « grand » (tend vers l'infini), les distributions discrètes tendent vers des distributions continues, souvent normales.

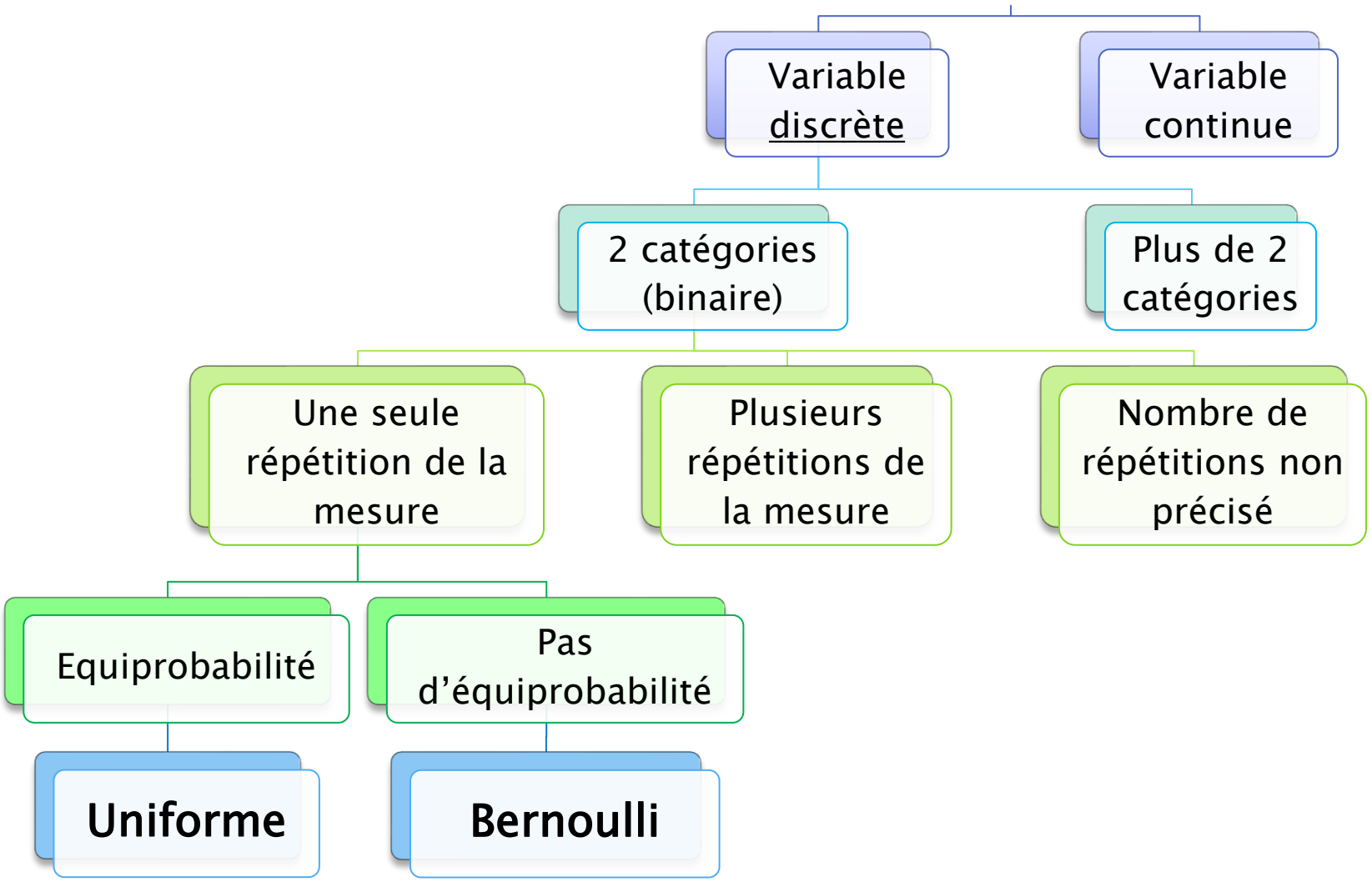


Les distributions discrètes

Niveaux de décision des distributions discrètes



Variables discrètes binaires à une répétition

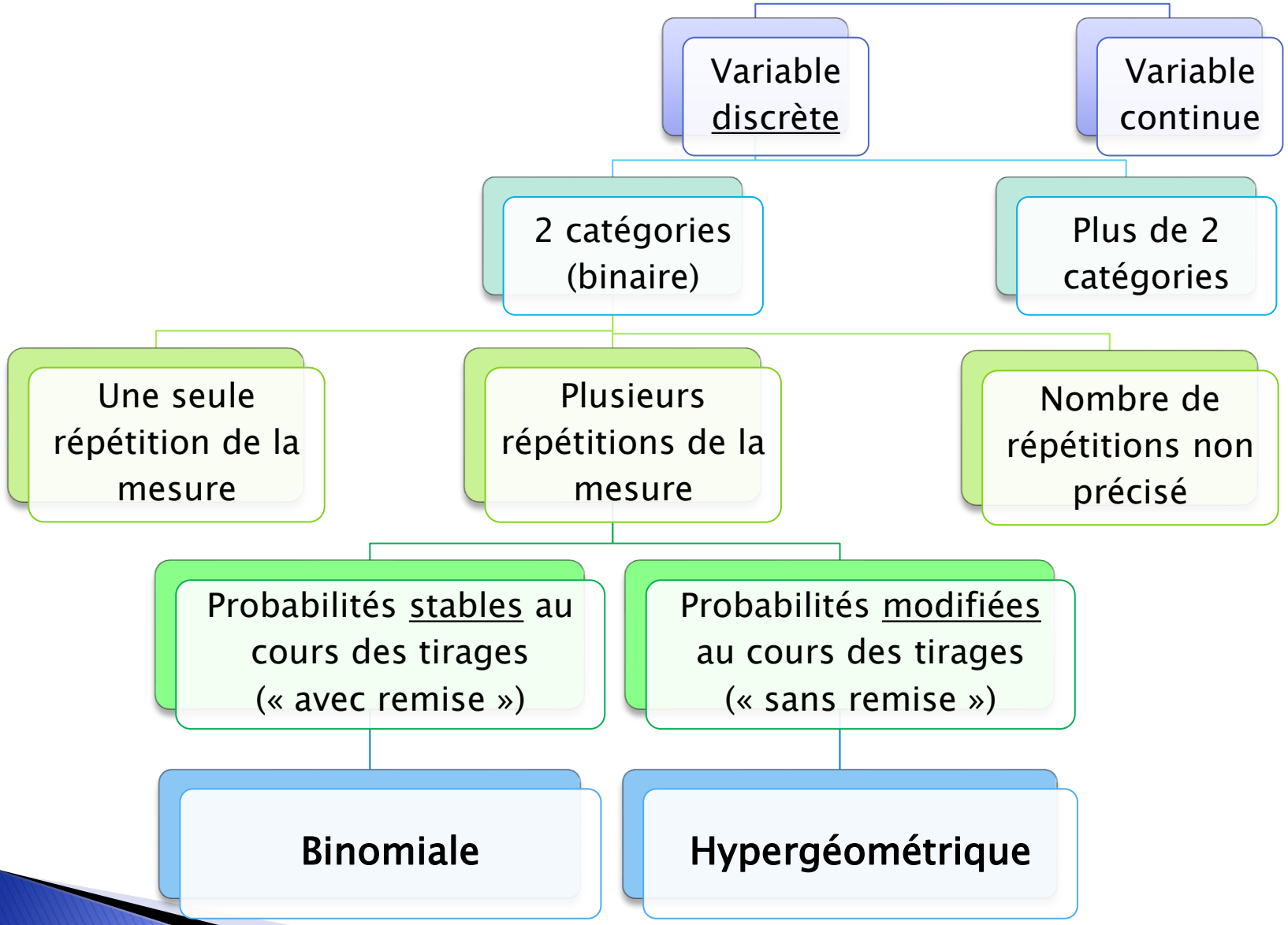


Variables discrètes binaires à une répétition

- **Bernoulli :**
  - Probabilités associées aux deux valeurs possibles de la variable :
    - p_1
    - et $p_2 = 1 - p_1$
  - Exemple : si une maladie a une prévalence de $\pi = 20\%$, l'état sanitaire d'un individu prélevé au hasard dans cette population est une variable discrète, avec 2 valeurs possibles (sain – malade), de probabilités respectives $p_1 = 0.2$ et $p_2 = 0.8$

- **Uniforme** :
  - On parlera de distribution uniforme discrète si $p_1 = p_2$, c'est-à-dire s'il y a équiprobabilité d'apparition des valeurs.
  - Ici, $p_1 = p_2 = 0{,}5$ puisqu'il n'y a que deux possibilités.
  - Exemple : s'il y a une proportion équivalente de mâles et de femelles dans une population, le sexe d'un individu prélevé au hasard dans cette population est une variable discrète, avec 2 valeurs possibles (femelle – mâle), de probabilités respectives $p_1 = 0.5$ et $p_2 = 0.5$

Variables discrètes binaires à plusieurs répétitions



Variables discrètes binaires à plusieurs répétitions

- **Binomiale – Bi (n, p_1)**
  - Probabilité d'observer la valeur est stable au cours des répétitions (« avec remise »)
  - n répétitions, fixées et indépendantes
  - Probabilité d'observer r fois la valeur qui a une probabilité p_1 (probabilité de « succès ») et $(n-r)$ fois celle qui a la probabilité $1-p_1$ (probabilité « d'échec ») lors de n tirages :

$$P = C_n^r * {p_1}^r * (1-p_1)^{(n-r)}$$

  - Exemple: si j'échantillonne 10 personnes au hasard dans une population où la prévalence d'une infection est de $\pi = 0.20$, quelle est la probabilité que 3 des 10 personnes soient infectées ?

    → si la population est très grande, prélever 10 personnes n'affecte quasiment pas la prévalence dans la population restante $=>$ on considère π constante de tirage en tirage.

    → Par conséquent:

$$P = C_{10}^3 * \pi^3 * (1-\pi)^{(10-3)} = 120 * 0.2^3 * 0.8^7 = 0.201$$

Variables discrètes binaires à plusieurs répétitions

- **Hypergéométrique – H (N, n)**
  - Probabilité d'observer la valeur modifiée à chaque répétition (« sans remise »)
  - Probabilité d'observer r fois la valeur qui nous intéresse dans un échantillon de taille n, si cet échantillon est extrait d'une population de taille N (≥ n) contenant R (≥ r) fois la valeur d'intérêt :

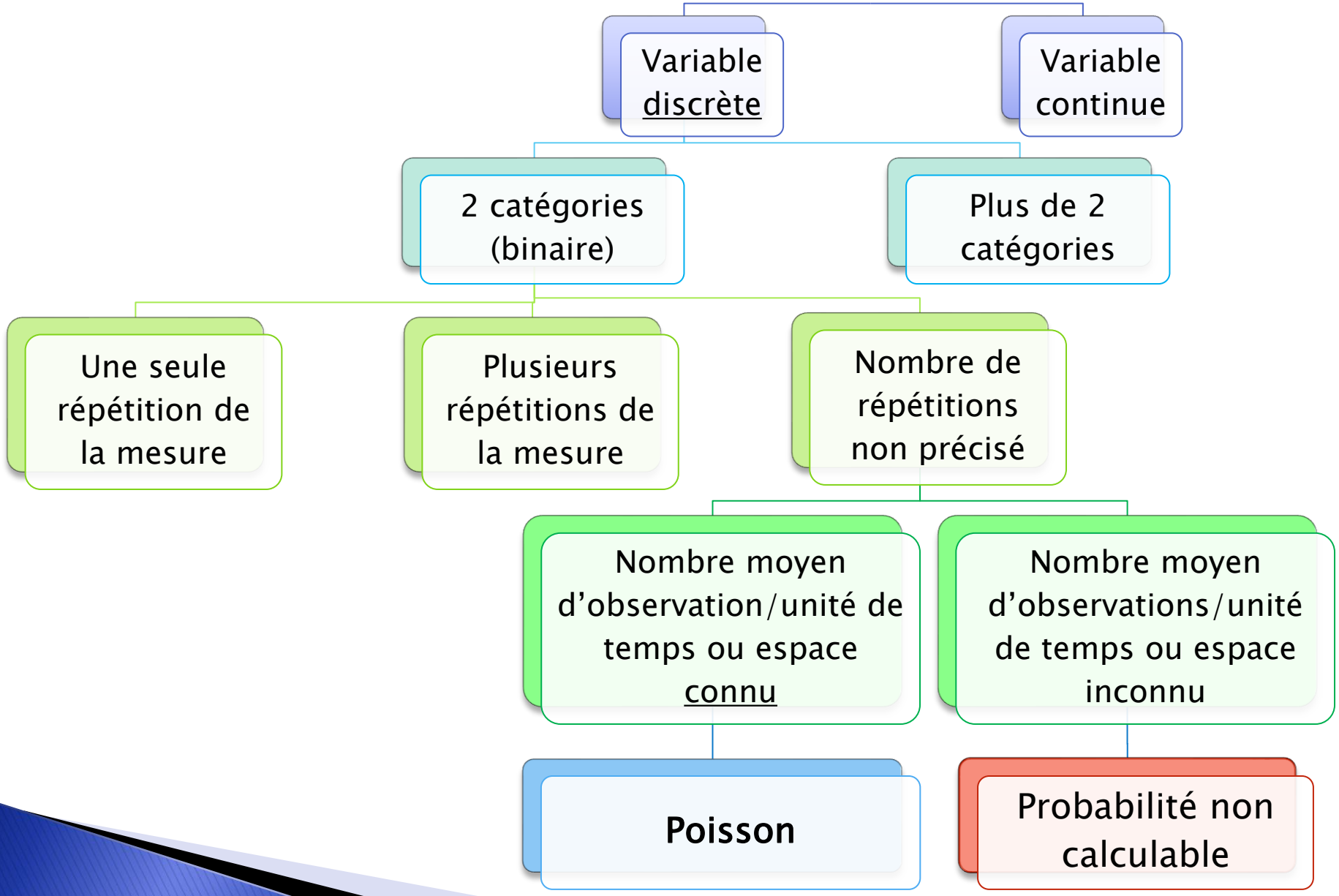
$$P = C_n^r * \frac{C_{N-n}^{R-r}}{C_N^R}$$

  - Exemple: si 5 vaches d'une ferme de 50 vaches sont parasitées, quelle est la probabilité d'avoir 1 vache parasitée dans un échantillon aléatoire de 5 vaches prélevées dans cette ferme ?

    → La probabilité que la première vache prélevée soit parasitée est 5/50. Si cette vache est parasitée, la probabilité que la seconde le soit aussi est 4/49. Et si la première vache est non parasitée, la probabilité que la seconde le soit est 5/49 => la probabilité change à chaque tirage !

$$P = C_5^1 * \frac{C_{45}^9}{C_{50}^{10}} = 0.4313$$

Variables discrètes binaires à nombre de répétitions inconnu



Variables discrètes binaires à nombre de répétitions inconnu

- **Poisson – Po (m)**
  - Probabilité d'observer r fois l'événement qui se produit en moyenne m fois vaut:

$$P = e^{-m} * m^r / r!$$

  - Exemple: si, en moyenne, 2 cas d'échinococcose sont observés par mois dans la population de renards d'une raison, quelle est la probabilité d'observer 5 cas un mois particulier ?

    → il s'agit bien d'un événement binaire (événement observé – événement non observé), sans limite sur le nombre de répétition de l'observation, mais dont le nombre moyen d'observations par mois est connu. On est bien dans un tirage de type Poisson

$$P = e^{-2} * 2^5 / 5! = 0.036$$

Choix de la distribution : éléments supplémentaires

- Binomiale VS Poisson

Binomiale	Poisson
Etude de la probabilité de répéter le nombre d'essais	Indique le nombre d'événements indépendants se produisant de manière aléatoire par unité de temps ou espace
Biparamétrique : caractérisée par n et p_1	Uniparamétrique : caractérisée par m
Nombre n de répétitions fixé	Nombre n « illimité », non fixé
Moyenne > Variance	Moyenne = Variance

- Distribution de Poisson peut être utilisée pour approximer distribution binomiale si :
  - Le nombre de tentatives (n) tend vers l'infini
  - Et que la probabilité de réussite (p) tend vers 0 de sorte que $m = np$

Choix de la distribution : éléments supplémentaires

- Binomiale VS Poisson :
  - Intéressant d'approximer avec la loi de Poisson quand N est grand → simplification des calculs manuels

- Exemple : Si la probabilité qu'un individu soit victime d'un choc sérique après injection d'un sérum donné est de 0,001, déterminez la probabilité que, parmi 5000 individus traités, 3 individus soient victimes de ce choc.

Poisson, avec k=3 et $\mu = n^*p = 5$

$$p(k) = \frac{e^{-\mu} * \mu^k}{k!}$$

$$p(k=3) = \frac{e^{-5} * 5^3}{3!} = \frac{e^{-5} * 5^3}{1 * 2 * 3}$$

$= 0{,}1403739$

Binomiale, avec r=3, n=5000, $p = 0{,}001$ et $q = 1-p = 0{,}999$

$$P(r) = C_n^r * p^r * q^{n-r} = \frac{n!}{r!(n-r)!}$$

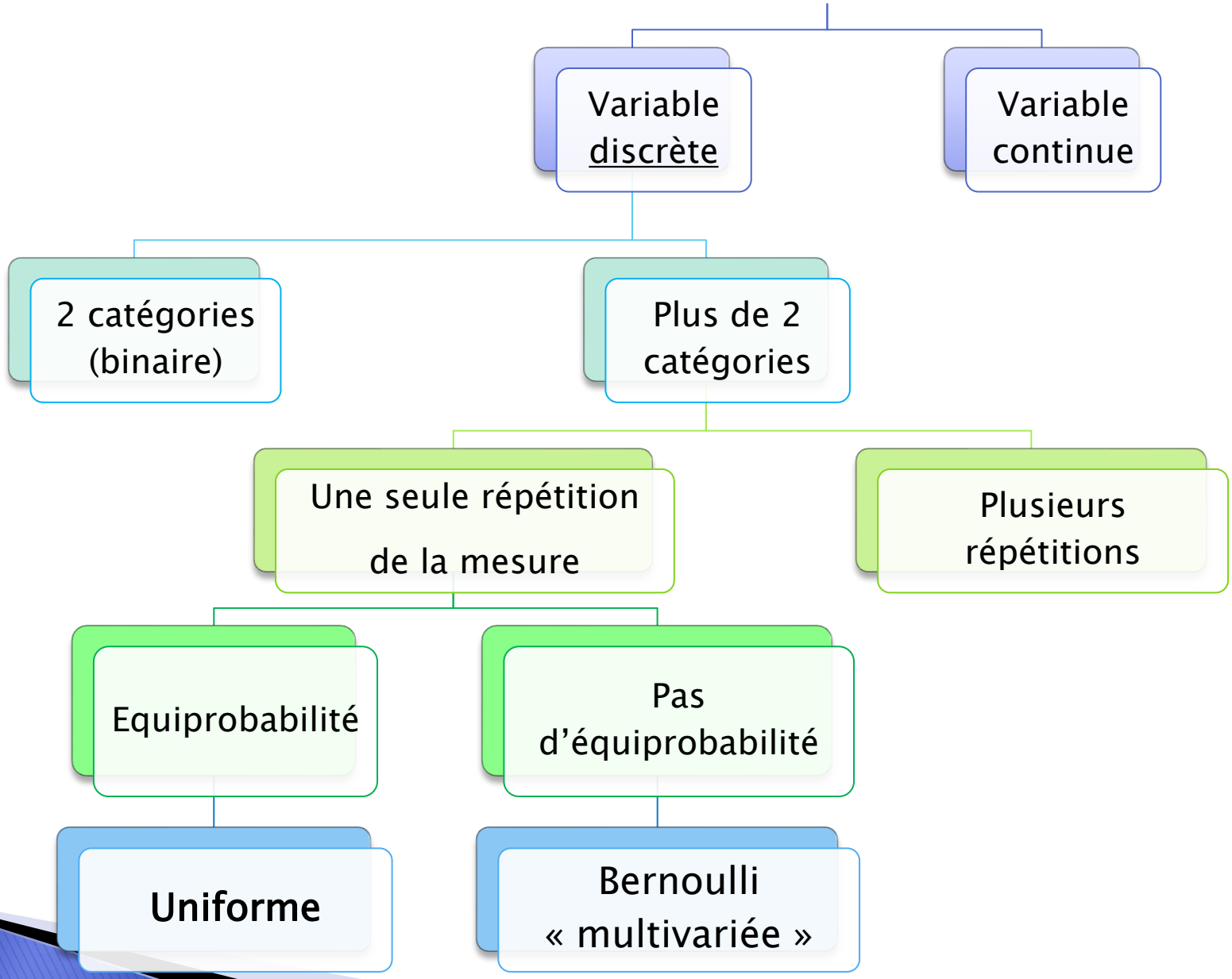
$$P(r=3) = C_{5000}^3 * 0{,}001^3 * 0{,}999^{4997}$$

$$= \frac{5000!}{3!\,4997!} 0{,}001^3 * 0{,}999^{4997}$$

$$= \frac{1 * 2 * 3 * 4 * 5 * \cdots * 4998 * 4999 * 5000}{1 * 2 * 3 * 1 * 2 * 3 * 4 * \cdots * 4996 * 4997} * 0{,}001^3 * 0{,}999^{4997}$$

$= 0{,}1403598$

Variables discrètes à plus de 2 catégories et à une répétition



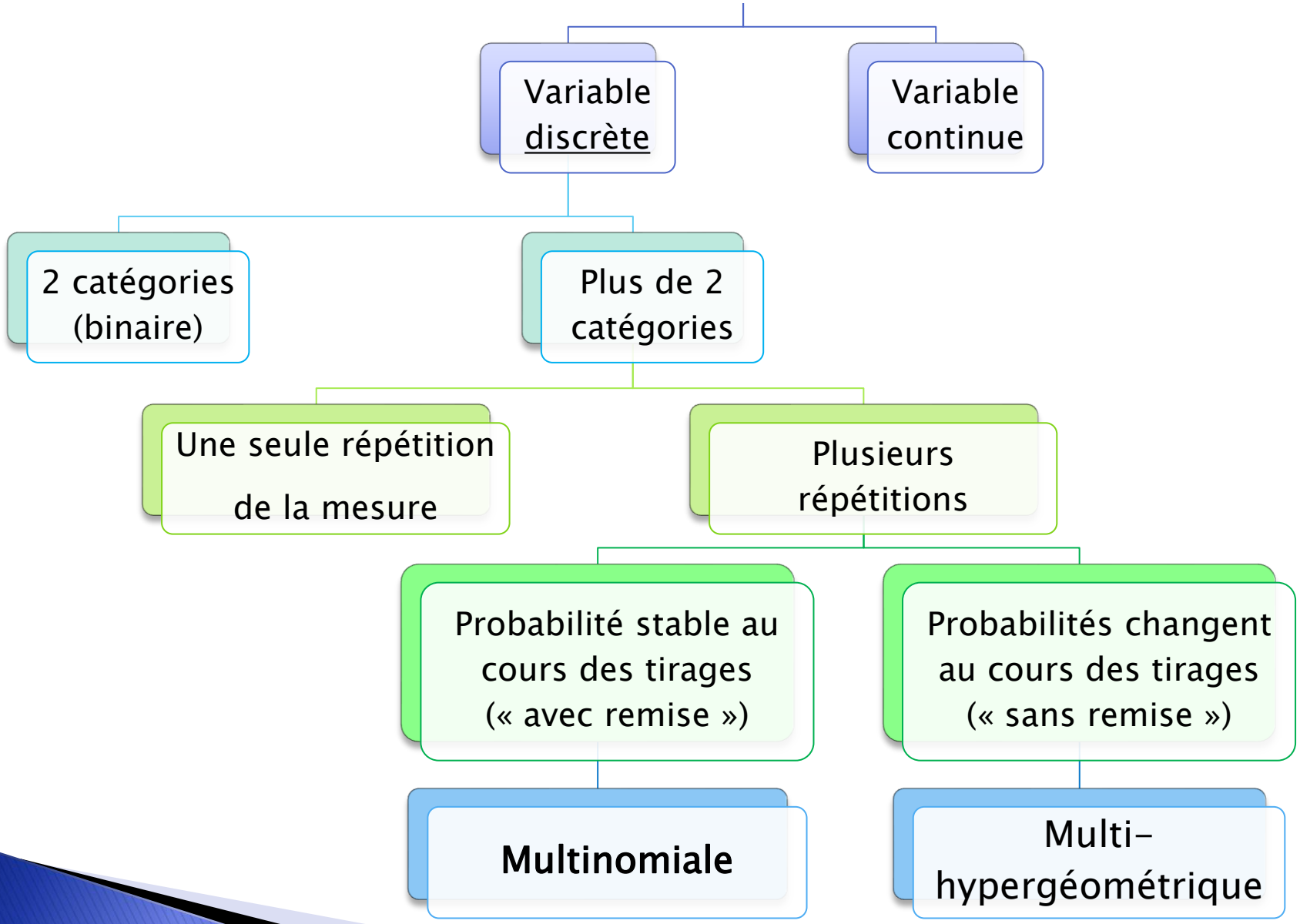
Variables discrètes à plus de 2 catégories et à une répétition

- **Bernoulli « multivariée »** :
  - Probabilités associées aux k valeurs possibles de la variable sont $p_1, p_2, \cdots, p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}$.
  - Exemple : si on croise 2 hétérozygotes AB, les génotypes AA, AB et BB chez les descendants ont des probabilités respectives valant $p_{AA} = 0.25$, $p_{AB} = 0.50$ et $p_{BB} = 0.25$.

- **Uniforme :**
  - On parlera de distribution uniforme discrète si $p_1 = p_2 - p_3 = \ldots = p_k$, c'est-à-dire s'il y a équiprobabilité d'apparition des valeurs.
  - Exemple 2: si on jette un dé correct, les probabilités des 6 faces sont identiques et valent :

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = {}^1/_6$$

Variables discrètes à plus de 2 catégories et à plusieurs répétitions



Variables discrètes à plus de 2 catégories et à une répétition

- **Multinomiale** :
  - Probabilité d'observer la valeur est stable au cours des répétitions (« avec remise »)
  - Probabilité d'observer r_1 fois la valeur qui a une probabilité p_1, r_2 fois celle qui a une probabilité p_2, …, r_k fois celle qui a une probabilité p_k lors de $n = r_1 + r_2 + \cdots + r_k$ tirages (avec $p_1 + p_2 + \cdots + p_k = 1$) :

$$P = \frac{n!}{r_1! * r_2! * \cdots * r_k!} * {p_1}^{r_1} * {p_2}^{r_2} * \cdots * {p_k}^{r_k}$$

  - Exemple: si je prends 10 labradors au hasard dans une population où 60% des chiens sont dorés, 30% sont noirs et 10% sont bruns, quelle est la probabilité que mon échantillon contienne 5 chiens dorés, 2 noirs et 3 bruns ?

    → Comme la population est très grande, prélever 10 chiens n'affecte quasiment pas la fréquence des robes dans la population restante => on considère les probabilités constantes de tirage en tirage. Par conséquent:

$$P = \frac{10!}{5! * 2! * 3!} * 0.6^5 * 0.3^2 * 0.1^3 = 0.0176$$

Variables discrètes à plus de 2 catégories et à une répétition

- Multi-hypergéométrique :
  - Probabilité d'observer chaque valeur modifiée à chaque tirage.
  - Généralisation de l'approche développée pour le tirage hypergéométrique: la population de taille N décomposée en sous-populations de tailles n_1, n_2, …, n_k, et échantillon, de taille R, formé de sous-échantillons de talles r_1, r_2, …, r_k provenant des sous-populations correspondantes.
  - Evidemment, $n_1 + n_2 + \cdots + n_k = N$ et $r_1 + r_2 + \cdots + r_k = R$ et :

$$P = \left[C_{n_1}^{r_1} * C_{n_2}^{r_2} * \cdots * C_{n_k}^{r_k}\right] / C_N^R$$

  - Exemple: une exploitation ovine compte 10 moutons Texel, 8 Bleu du Maine et 6 Ile de France. Si je prends 5 moutons au hasard dans cette exploitation, quelle est la probabilité d'avoir 3 Texel, 1 Bleu du Maine et 1 Ile de France ?

    → les probabilités de tirer un mouton d'une race particulière changent à chaque tirage, suite aux tirages.

$$P = [C_{10}^3 * C_8^1 * C_6^1] / C_{24}^5 = 0.1355$$

Les distributions continues

Les distributions continues

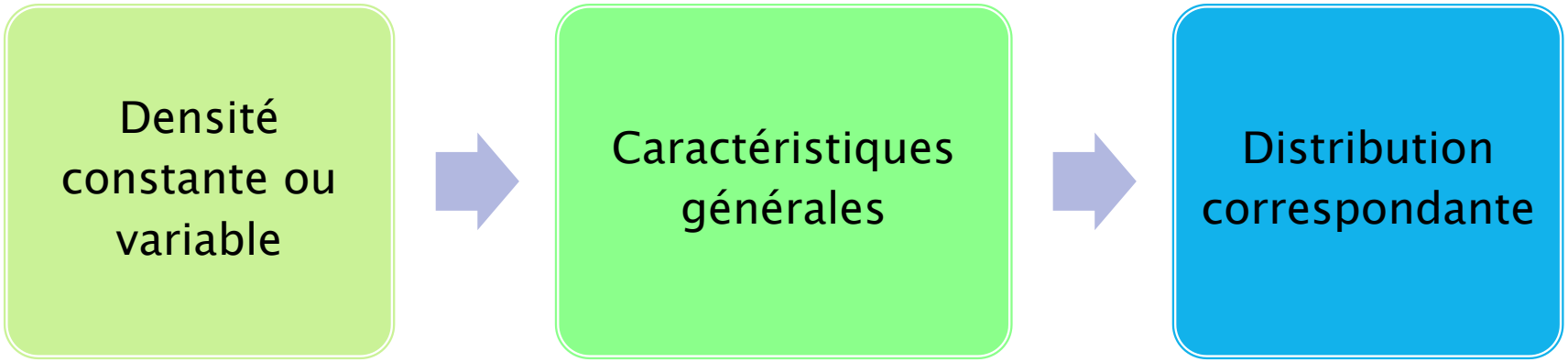
- Généralités
  - Dans le cas des variables continues, on calcule la probabilité que la variable tombe dans un intervalle [a,b] (et non la probabilité d'une valeur particulière, qui vaut 0) par la surface comprise entre X = a, X = b, l'axe horizontal et f(X).

  - Le calcul de probabilité se fait donc via une intégrale:

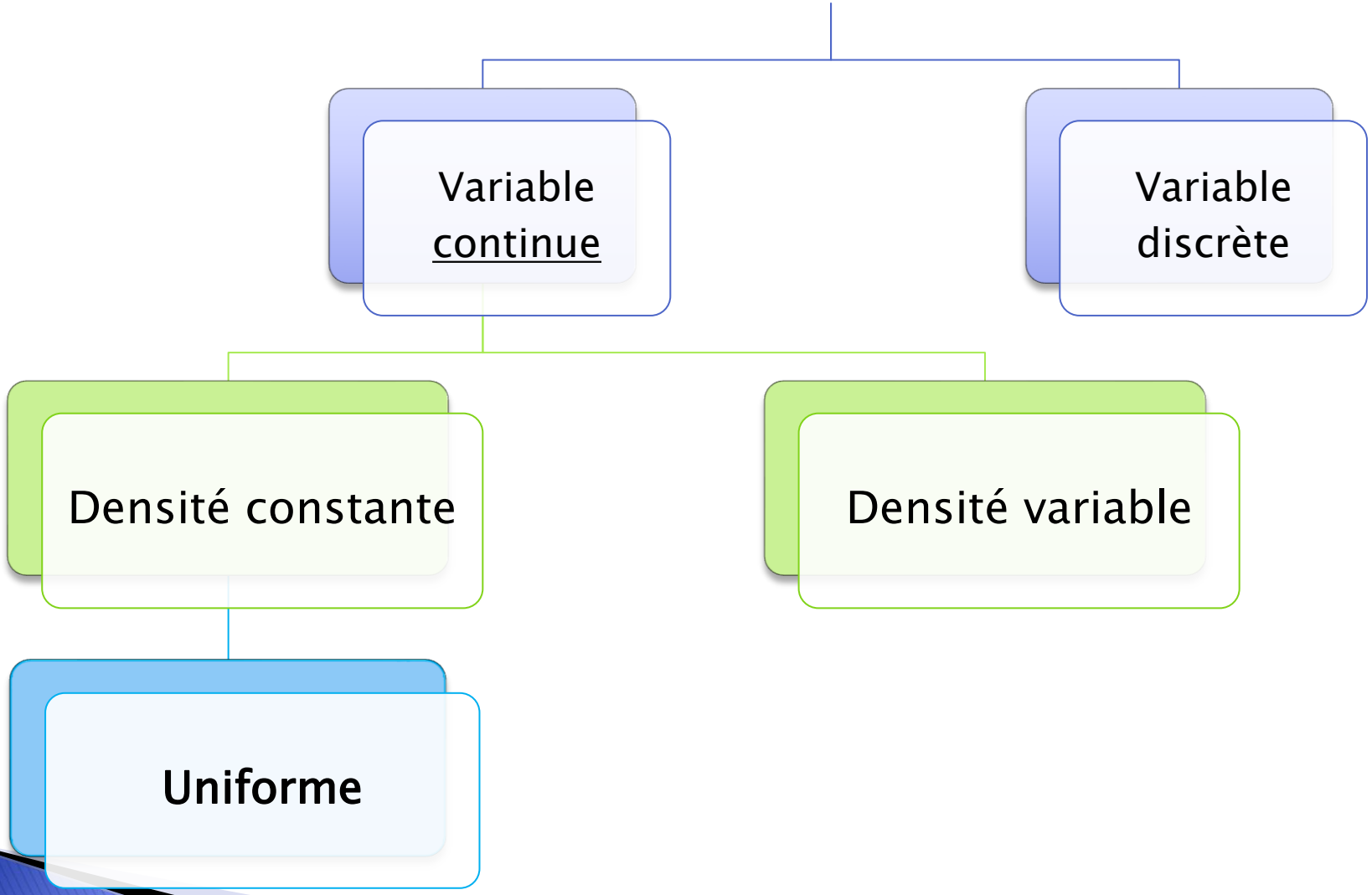
$$P = \int_a^b f(X) * dX$$

    où $f(X)$ est une fonction appelée « densité de probabilité » ($\neq$ probabilité).
  - Evidemment : $\int_{-\infty}^{+\infty} f(X) * dX = 1$

Niveaux de décision des distributions continues

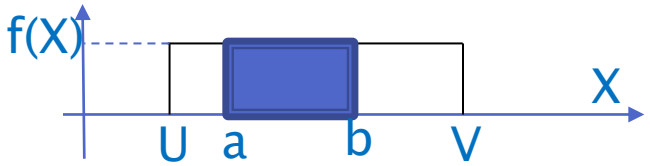


Variables continues à densité constante



Variables continues à densité constante

- **Uniforme** :
  - Densité de probabilité constante dans un intervalle [U,V], et nulle ailleurs.
  - Comme la surface du rectangle formé vaut 1, la valeur de la densité de probabilité entre U et V vaut 1/(V–U)
  - Probabilité d'être entre a et b (où $U \leq a \leq b \leq V$) égale à la surface du rectangle ainsi formé, et vaut (b–a)/(V–U)

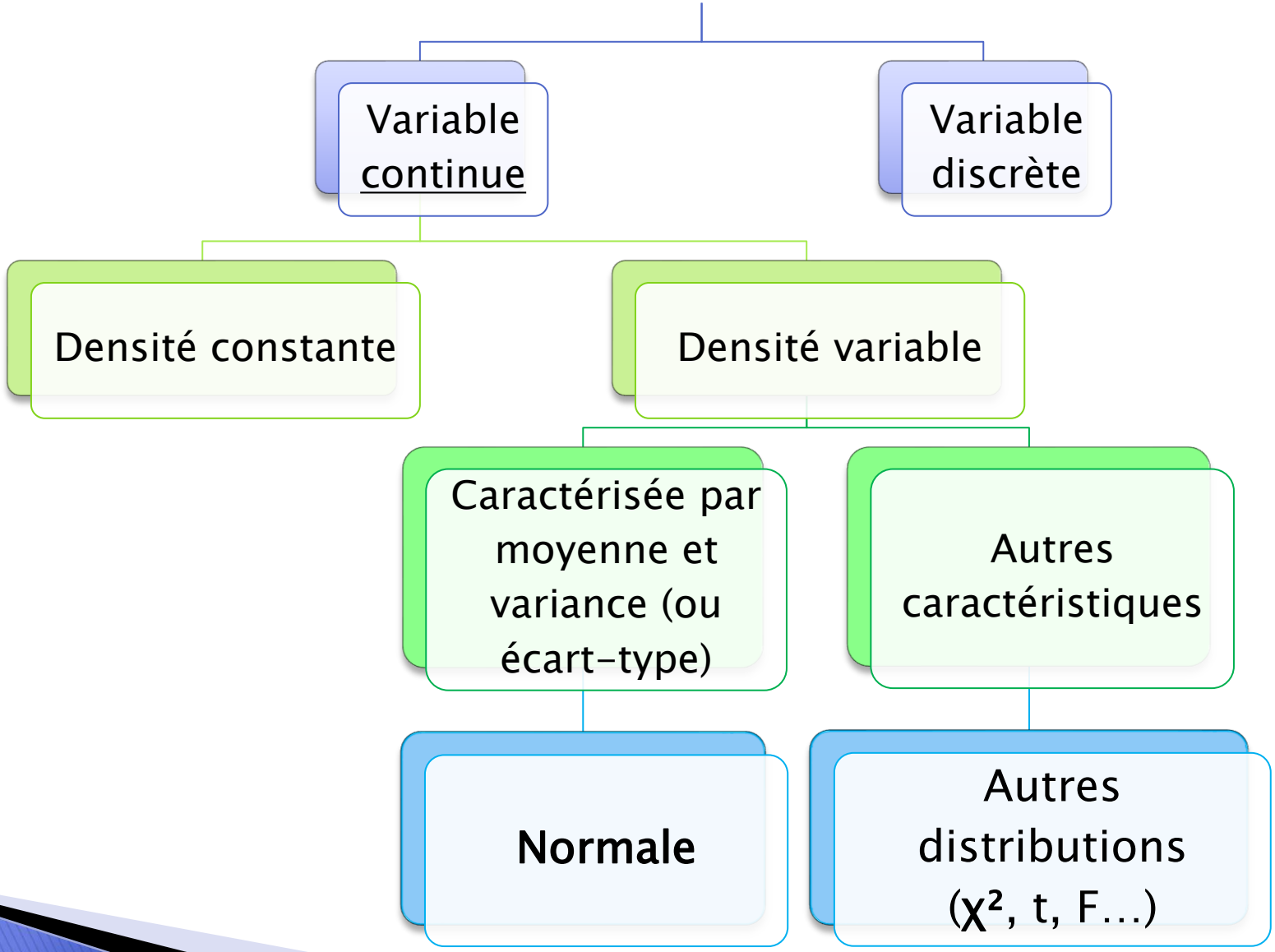


  - Exemple : les prévalences d'une infection dans les chenils varient de manière uniforme entre 10% et 70%. Si je prends un chenil au hasard, quelle est la probabilité que la prévalence dans ce chenil soit supérieure à 50% ?

    → la prévalence a une densité uniforme entre U = 0.1 et V = 0.7 => f(X) = 1/0.6. La probabilité d'une prévalence supérieure à 0.5 (donc entre 0.5 et 0.7) vaut donc:

$$P = \frac{0.7-0.5}{0.7-0.1} = 0.333$$

Variables continues à densité variable



Variables continues à densité variable

- **Normale – N(μ, σ²)**
  - Répartition des données selon laquelle la plupart des valeurs sont regroupées autour de la moyenne et les autres s'en écartent symétriquement des deux côtés.
    → Définie uniquement par la moyenne μ et variance σ²
  - Calculs sur n'importe quelle distribution normale peuvent être ramenés a des calculs sur la distribution N(0,1) par un changement de variable : c'est la distribution normale réduite
  - Exemple: les poids des veaux à la naissance ont une distribution normale de moyenne 30 kgs et de déviation standard 5 kgs. Quelle est la probabilité d'avoir un veau pesant entre 35 et 40 kgs à la naissance ?

    →L'intégrale à calculer est $P = \int_{35}^{40} f(X) * dX$ où $f(X)$ est la densité normale de moyenne $\mu = 30$ et de déviation standard $\sigma = 5$.

    → On peut utiliser un logiciel pour le calcul ou utiliser les tables de Z en transformant X en Z:

$$P(35 \leq X \leq 40) = P\left(\frac{35-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{40-\mu}{\sigma}\right)$$
$$= P(1 \leq Z \leq 2) = P(Z \leq 2) - P(Z \leq 1)$$

Variables continues à densité variable

- **Chi-carré (ou Khi-carré) χ^2 – K(ddl)**
  - Discrétisation des données : catégories regroupant des effectifs
  - Dépend du nombre de degrés de libertés
  - Approximation d'un test exact (bi ou multinomial) pour tester une hypothèse

  - Exemple : un calcul a conduit a une valeur de χ^2 (avec 3 degrés de liberté) valant 10.21. Est-il fréquent d'avoir des valeurs aussi élevées, voire plus élevées, pour une telle variable ?

    → l'intégrale à calculer est $P = \int_{10.21}^{+\infty} f\left(\chi^2_{3\ ddl}\right) * d\,\chi^2_{3\ ddl}$
    Comme cette intégrale est complexe à calculer, il est préférable d'utiliser les tables où un logiciel (voir plus loin).
    On obtient $P = 0.0169$

- **Autres distributions (t, F…) : *Cfr BMV2***