Méthodes statistiques pour Les données qualitatives

Rapport de projet

Février 2025

Soumis à:
Pr. N'deye Niang
n-deye.niang_keita@cnam.fr

Soumis par :
Krissaan Amen Allah , M2 TRIED
amenallahkrissane10@gmail.com

Bacha Amine, M2 TRIED
ia_bacha@esi.dz

Dépôt du code :
https://github.com/amenallah01/Methodes-statistiques-pour-les-donnees-qualitatives.git

Date de soumission : 19/02/2025

I. Introduction

Dans un monde où la donnée est omniprésente, l'analyse statistique joue un rôle central dans l'extraction d'informations pertinentes et exploitables. Si les données numériques sont souvent privilégiées, les données qualitatives sont tout aussi cruciales, notamment dans les domaines où les caractéristiques descriptives dominent, comme la biologie, la sociologie ou encore l'industrie agroalimentaire. L'analyse des données qualitatives permet de mieux comprendre les relations entre les différentes modalités, d'identifier des structures sous-jacentes et d'exploiter ces connaissances pour la prise de décision.

Le jeu de données **Mushroom** illustre parfaitement l'importance de cette analyse. Composé de **22 variables qualitatives** décrivant des caractéristiques morphologiques des champignons, il est utilisé pour déterminer si un spécimen est **comestible ou toxique**. Une mauvaise classification peut avoir des conséquences graves sur la santé humaine, d'où la nécessité d'une **analyse rigoureuse et approfondie** des données disponibles.

Pour mener cette analyse de manière efficace, plusieurs étapes méthodologiques sont essentielles :

1. **Statistique descriptive** : Cette première phase permet d'obtenir une vision d'ensemble des données. En résumant les fréquences des modalités et en identifiant d'éventuelles valeurs dominantes ou rares, elle aide à mieux comprendre la distribution des variables et les premiers liens entre elles.
2. **Analyse des correspondances multiples (ACM)** : Étant donné que nos variables sont exclusivement qualitatives, l'ACM est un outil puissant pour représenter ces données dans un espace à faible dimension. Elle permet d'identifier des axes factoriels expliquant au mieux les variations des modalités et de dégager les structures sous-jacentes du dataset.
3. **Classification non supervisée** : Une fois l'ACM réalisée, une classification non supervisée (comme le clustering) peut être appliquée aux composantes obtenues. L'objectif est d'identifier **des groupes homogènes de champignons** partageant des caractéristiques communes, ce qui peut être utile pour repérer des similarités entre espèces.
4. **Analyse discriminante** : Enfin, cette méthode permet d'évaluer la capacité des composantes issues de l'ACM à différencier correctement les classes des champignons (notamment entre comestibles et toxiques). Elle est particulièrement utile pour valider l'utilité des axes factoriels et pour **développer un modèle prédictif fiable** basé sur ces nouvelles représentations des données.

L'application de ces méthodes au dataset **Mushroom** apporte une **réelle valeur ajoutée** : elle permet de mieux comprendre les caractéristiques déterminantes dans la toxicité des champignons, d'identifier des regroupements naturels et de poser les bases d'un **modèle de classification robuste**. Ces analyses ne sont pas seulement intéressantes d'un point de vue statistique, elles ont également des **implications concrètes en mycologie**, en aidant à établir des critères fiables pour différencier les champignons comestibles des espèces dangereuses.

Ainsi, l'ensemble de cette démarche méthodologique met en lumière **la richesse et l'importance de l'analyse des données qualitatives**, confirmant qu'elles sont un levier puissant pour extraire de la connaissance, même dans un domaine où la sécurité est un enjeu crucial.

II. Description du dataset

Le **jeu de données Mushroom** contient les descriptions de **8 124 échantillons** de champignons appartenant à **23 espèces** des familles **Agaricus** et **Lepiota**. Chaque champignon est caractérisé par **22 variables qualitatives**, décrivant des aspects morphologiques tels que la forme du chapeau, la couleur des lamelles, la texture de la tige, et bien d'autres. La **variable cible** indique si le champignon est **comestible** ou **toxique** (en combinant les espèces toxiques et celles à comestibilité inconnue). Ce dataset est particulièrement intéressant car il n'existe **aucune règle simple** pour prédire la

comestibilité des champignons, rendant l'analyse statistique cruciale pour identifier les caractéristiques associées à la toxicité.

Voici la table avec la variable cible, prête à être insérée dans Word :

Nom de la variable	Description
Classe cible	Comestibilité : edible (comestible), poisonous (toxique)
cap-shape	Forme du chapeau : bell (en cloche), conical (conique), convex (convexe), flat (plat), knobbed (bosselé), sunken (enfoncé)
cap-surface	Surface du chapeau : fibrous (fibreuse), grooves (striée), scaly (écailleuse), smooth (lisse)
cap-color	Couleur du chapeau : brown (brun), buff (chamois), cinnamon (cannelle), gray (gris), green (vert), pink (rose), purple (violet), red (rouge), white (blanc), yellow (jaune)
bruises	Présence de meurtrissures : bruises (oui), no (non)
odor	Odeur : almond (amande), anise (anis), creosote (créosote), fishy (poisson), foul (nauséabonde), musty (moisi), none (aucune), pungent (pénétrante), spicy (épicée)
gill-attachment	Attache des lamelles : attached (attachées), descending (descendantes), free (libres), notched (échancrées)
gill-spacing	Espacement des lamelles : close (serrées), crowded (bondées), distant (éloignées)
gill-size	Taille des lamelles : broad (larges), narrow (étroites)
gill-color	Couleur des lamelles : black (noir), brown (brun), buff (chamois), chocolate (chocolat), gray (gris), green (vert), orange (orange), pink (rose), purple (violet), red (rouge), white (blanc), yellow (jaune)
stalk-shape	Forme du pied : enlarging (s'élargissant), tapering (s'effilant)
stalk-root	Type de racine du pied : bulbous (bulbeuse), club (en massue), cup (en coupe), equal (égale), rhizomorphs (rhizomorphes), rooted (avec racines), missing (inconnue)
stalk-surface-above-ring	Surface du pied au-dessus de l'anneau : fibrous (fibreuse), scaly (écailleuse), silky (soyeuse), smooth (lisse)
stalk-surface-below-ring	Surface du pied en dessous de l'anneau : fibrous (fibreuse), scaly (écailleuse), silky (soyeuse), smooth (lisse)
stalk-color-above-ring	Couleur du pied au-dessus de l'anneau : brown (brun), buff (chamois), cinnamon (cannelle), gray (gris), orange (orange), pink (rose), red (rouge), white (blanc), yellow (jaune)
stalk-color-below-ring	Couleur du pied en dessous de l'anneau : brown (brun), buff (chamois), cinnamon (cannelle), gray (gris), orange (orange), pink (rose), red (rouge), white (blanc), yellow (jaune)
veil-type	Type de voile : partial (partiel), universal (universel)
veil-color	Couleur du voile : brown (brun), orange (orange), white (blanc), yellow (jaune)
ring-number	Nombre d'anneaux : none (aucun), one (un), two (deux)
ring-type	Type d'anneau : cobwebby (toile d'araignée), evanescent (évanescent), flaring (évasé), large (large), none (aucun), pendant (pendant), sheathing (gaine), zone (zone)
spore-print-color	Couleur de l'empreinte des spores : black (noir), brown (brun), buff (chamois), chocolate (chocolat), green (vert), orange (orange), purple (violet), white (blanc), yellow (jaune)
population	Population : abundant (abondante), clustered (en grappes), numerous (nombreuse), scattered (dispersée), several (plusieurs), solitary (solitaire)
habitat	Habitat : grasses (herbes), leaves (feuilles), meadows (prairies), paths (chemins), urban (urbain), waste (déchets), woods (bois)

III. Etude dimensionnelle

Etude uni- dimensionnelle

1. Variable cible : class cible poisonous

- **Modalités** : edible (comestible), poisonous (toxique)

Remarque : L'histogramme montre la répartition des champignons en deux catégories : comestibles ("e" pour edible) et toxiques ("p" pour poisonous). Le jeu de données contient 4208 champignons comestibles (51.8%) et 3916 champignons toxiques (48.2%). Cette répartition est presque équilibrée, garantissant une analyse statistique non biaisée.

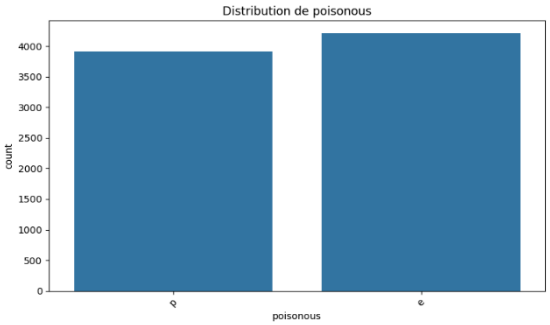


Figure 1-Distribution de poisonous

2. cap-shape (forme du chapeau)

- **Modalités** : bell, conical, convex, flat, knobbed, sunken

Remarque : L'histogramme révèle que la majorité des champignons ont une forme de chapeau convexe (45%) ou plate (38%), représentant ensemble 83% des observations. Les formes knobbed (10.19%) et bell (5.56%) sont moins fréquentes, tandis que les formes sunken et conical sont marginales (0.44%). Ces dernières sont regroupées sous une modalité unique pour simplification.

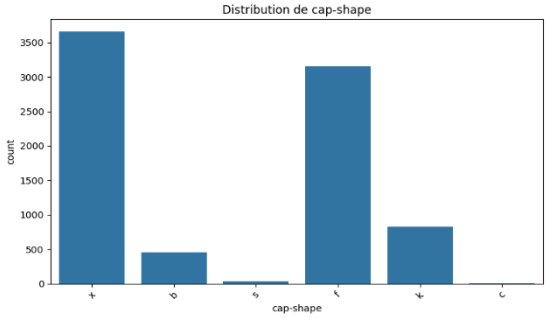


Figure 2 -Distribution de cap-shap

3. cap-surface (surface du chapeau)

- **Modalités** : fibrous, grooves, scaly, smooth

Remarque : La texture dominante est scaly (39.93%), suivie de smooth (31.46%) et fibrous (28.56%). La modalité grooves est extrêmement rare (0.05%) et est regroupée avec fibrous pour éviter une dispersion excessive des données.

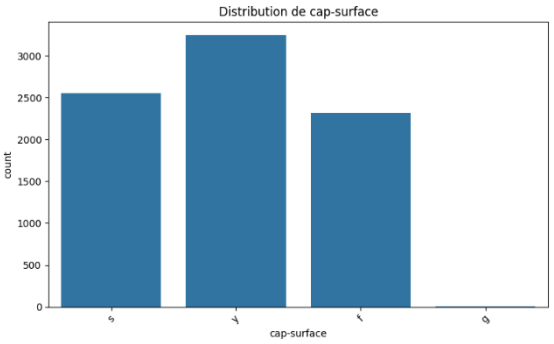


Figure 3 -Distribution de cap-surface

4. cap-color (couleur du chapeau)

- **Modalités** : brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow

Remarque : Les couleurs les plus fréquentes sont brun gris rouge , jaune et blanc .

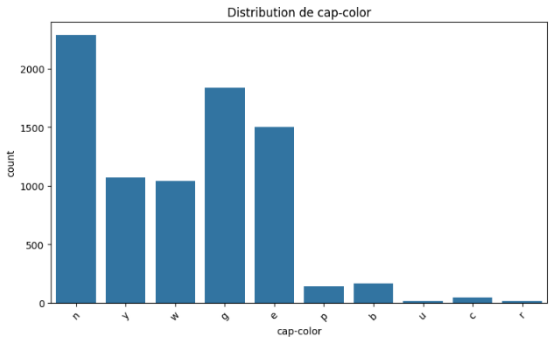


Figure 4 -Distribution de cap-Color

5. bruises (présence de meurtrissures)

- **Modalités** : bruises, no

Remarque : Les champignons sans ecchymoses sont plus nombreux (58.44%) que ceux qui en ont (41.56%), bien que la répartition reste relativement équilibrée.

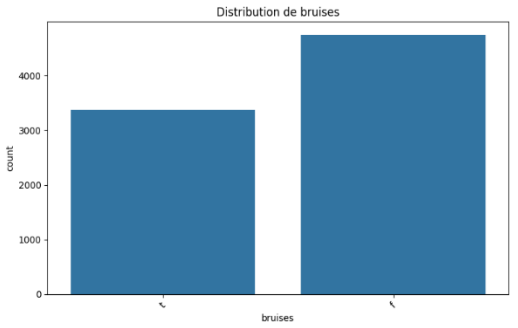


Figure 5 -Distribution de Bruises

6. odor (odeur)

- **Modalités** : almond, anise, creosote, fishy, foul, musty, none, pungent, spicy

Remarque : L'absence d'odeur est la caractéristique la plus fréquente (43.43%), suivie des odeurs foul (26.59%), fishy (7.09%), et spicy (7.09%). Les odeurs rares musty, creosote, et pungent (5.96%) sont regroupées ensemble pour simplifier l'analyse.

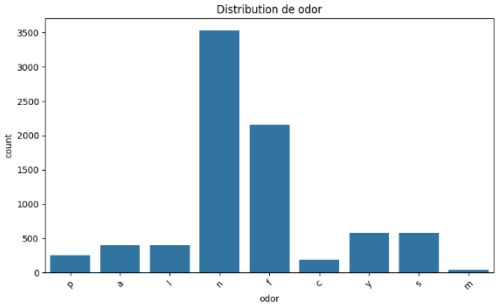


Figure 6 -Distribution de odor

7. gill-attachment (attache des lamelles)

- **Modalités** : attached, descending, free, notched

Remarque : La modalité free domine largement (97.42%), tandis que la modalité attached est rare (2.58%). Cette variable pourrait être peu informative pour l'analyse.

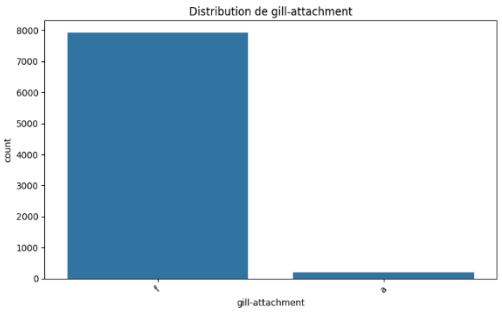


Figure 7 -Distribution de gill-attachment

8. gill-spacing (espacement des lamelles)

- **Modalités** : close, crowded, distant

Remarque : L'histogramme montre une nette prédominance des lamelles rapprochées (close, 83.85%), tandis que les lamelles crowded (16.15%) sont moins courantes.

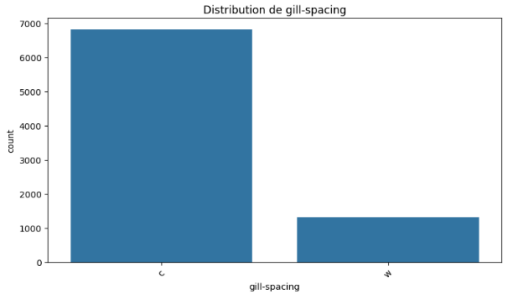


Figure 8 -Distribution de gill-spacing

9. gill-size (taille des lamelles)

- **Modalités** : broad, narrow

Remarque : Les lamelles larges (broad, 69.08%) sont plus fréquentes que les lamelles étroites (narrow, 30.92%), rendant cette variable déséquilibrée.

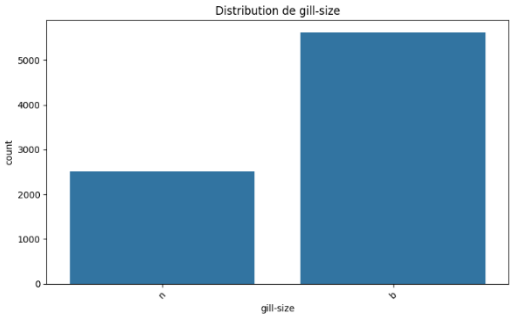


Figure 9 -Distribution de gill-size

10. gill-color (couleur des lamelles)

- **Modalités** : black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow

Remarque : Les couleurs les plus courantes sont buff (21.27%), pink (18.37%), white (14.80%), et brown (12.90%).

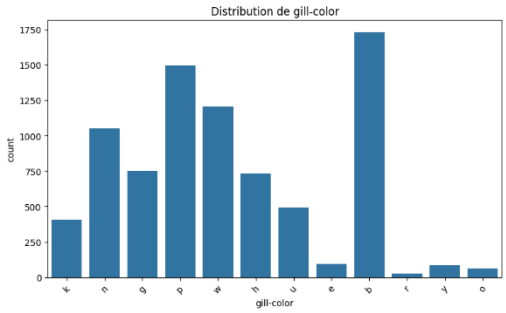


Figure 10 -Distribution de gill-color

11. stalk-shape (forme du pied)

- **Modalités** : enlarging, tapering

Remarque : Les formes sont relativement équilibrées entre tapering (56.72%) et enlarging (43.28%).

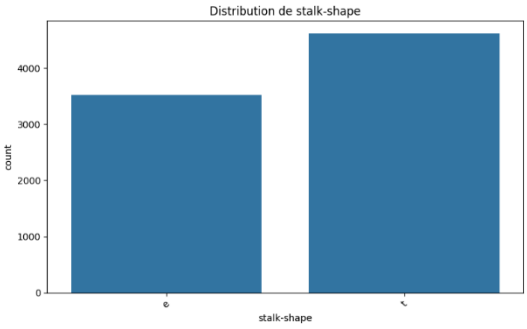


Figure 11 -Distribution de gill-shape

12. stalk-root (type de racine du pied)

- **Modalités** : bulbous, club, cup, equal, rhizomorphs, rooted, missing

Remarque : La modalité bulbous (46.48%) est majoritaire, suivie de equal, tandis que la modalité rooted est rare, faut noter aussi que cette variable contient un certain nombre de lignes vides

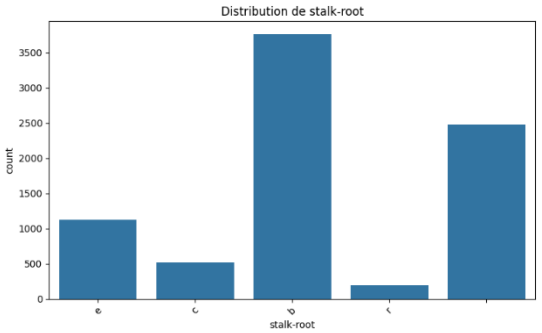


Figure 12 -Distribution de stalkl-root

13. stalk-surface-above-ring (surface du pied au-dessus de l'anneau)

- **Modalités** : fibrous, scaly, silky, smooth

  Remarque : Présence d'une modalité rare scaly

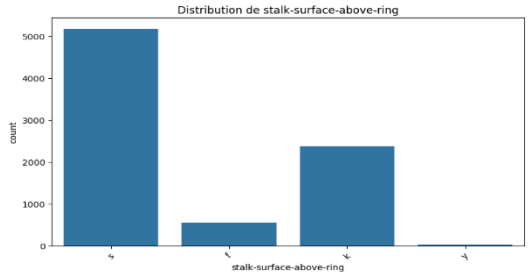


Figure 13 -Distribution de stalk-surface-above-ring

14. stalk-surface-below-ring (surface du pied en dessous de l'anneau)

- **Modalités** : fibrous, scaly, silky, smooth

Remarque : Variable déséquilibrée

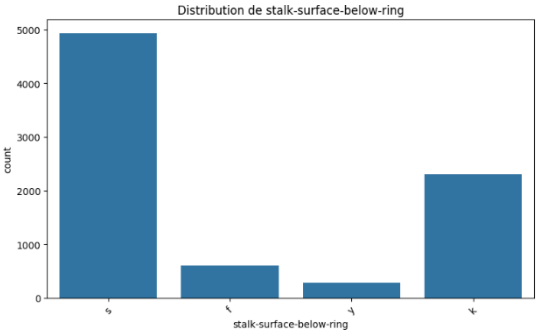


Figure 14 -Distribution de stalk-surface-below-ring

15. stalk-color-above-ring (couleur du pied au-dessus de l'anneau)

- **Modalités** : brown, buff, cinnamon, gray, orange, pink, red, white, yellow

Remarque : Présence de couleurs rares : cinnamon, yellow, et red

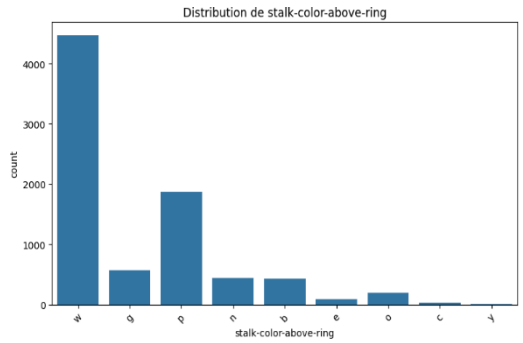


Figure 15 -Distribution de stalk-color-above-ring

16. stalk-color-below-ring (couleur du pied en dessous de l'anneau)

- **Modalités** : brown, buff, cinnamon, gray, orange, pink, red, white, yellow

Remarque : Les mêmes couleurs rares cinnamon, yellow, et red

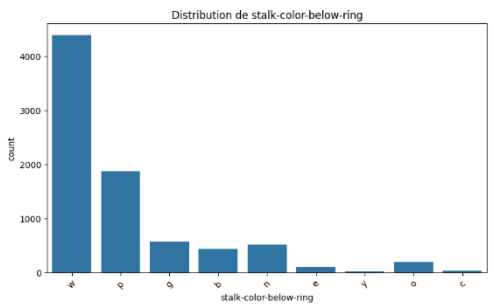


Figure 16 -Distribution de stalk-color-below-ring

17. veil-type (type de voile)

- **Modalités** : partial, universal

Remarque : Une seule modalité est présente, rendant la variable inutile et susceptible d'être supprimée

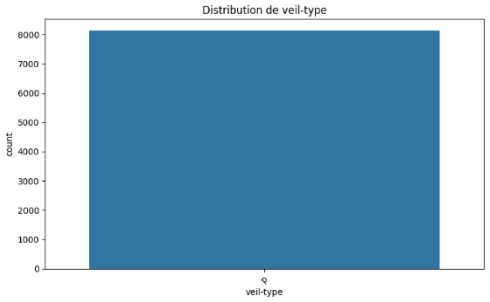


Figure 17 -Distribution de veil-type

18. veil-color (couleur du voile)

- **Modalités** : brown, orange, white, yellow

Remarque : Les couleurs rares brown, yellow, et orange

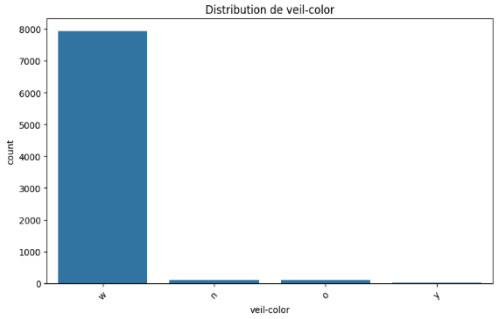


Figure 18 -Distribution de veil-color

19. ring-number (nombre d'anneaux)

- **Modalités** : none, one, two

Remarque : La modalité none est rare

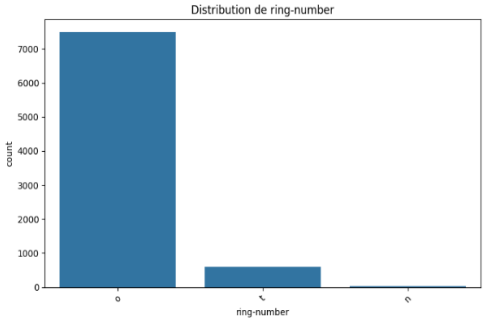


Figure 19 -Distribution de ring-number

20. ring-type (type d'anneau)

- **Modalités** : cobwebby, evanescent, flaring, large, none, pendant, sheathing, zone

Remarque : Présence de modalités rares flaring et none, difficilement regroupables avec d'autres catégories

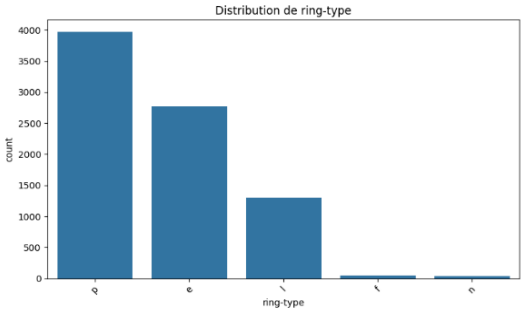


Figure 20 -Distribution de ring-type

21. spore-print-color (couleur de l'empreinte des spores)

- **Modalités** : black, brown, buff, chocolate, green, orange, purple, white, yellow*

Remarque : Les couleurs black brown chocolate et white sont très présents contrairement aux autre qui sont rares

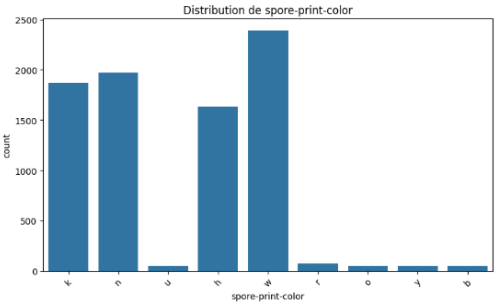


Figure 21 -Distribution de spore-print-color

22. population (population)

- **Modalités** : abundant, clustered, numerous, scattered, several, solitary

Remarque : Variable déséquilibrée mais sans modalités rares.

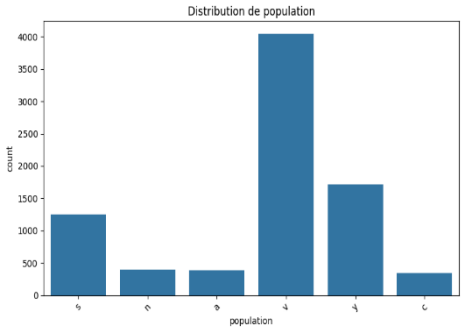


Figure 22 -Distribution de population

23. habitat (habitat)

- **Modalités** : grasses, leaves, meadows, paths, urban, waste, woods

Remarque : Variable également déséquilibrée, sans modalités rares

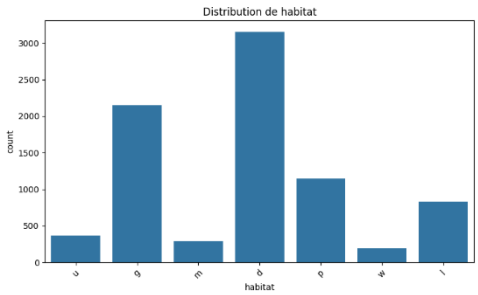


Figure 21 -Distribution de habitat

Prétraitement des données :

Certaines modalités sont très peu représentées dans les données. Un déséquilibre important peut poser problème, car le modèle risque de ne pas apprendre correctement sur ces modalités rares. Il est alors préférable de :

- Les supprimer si elles sont trop peu nombreuses comme dans cap-surface la modalité grooves.
- Les regrouper sous une catégorie "Autre" si elles existent en plusieurs occurrences mais restent minoritaires comme les couleur rares dans stalk-color.

Certaines variables n'apportent pas d'information utile à la prédiction. Par exemple, si une variable a une seule modalité dominante à plus de 95%, elle n'apporte pas de variation significative et peut être supprimée sans perte d'information.

a. Suppression de variables entières

Nous avons supprimé deux variables :

- **gill-attachment** : Cette variable a une seule modalité dans le dataset, elle est donc inutile.
- **veil-type** : Pareil, une seule modalité présente, donc sans intérêt pour la classification.

b. Regroupement ou suppression de modalités

Pour les autres variables, nous avons appliqué deux stratégies :

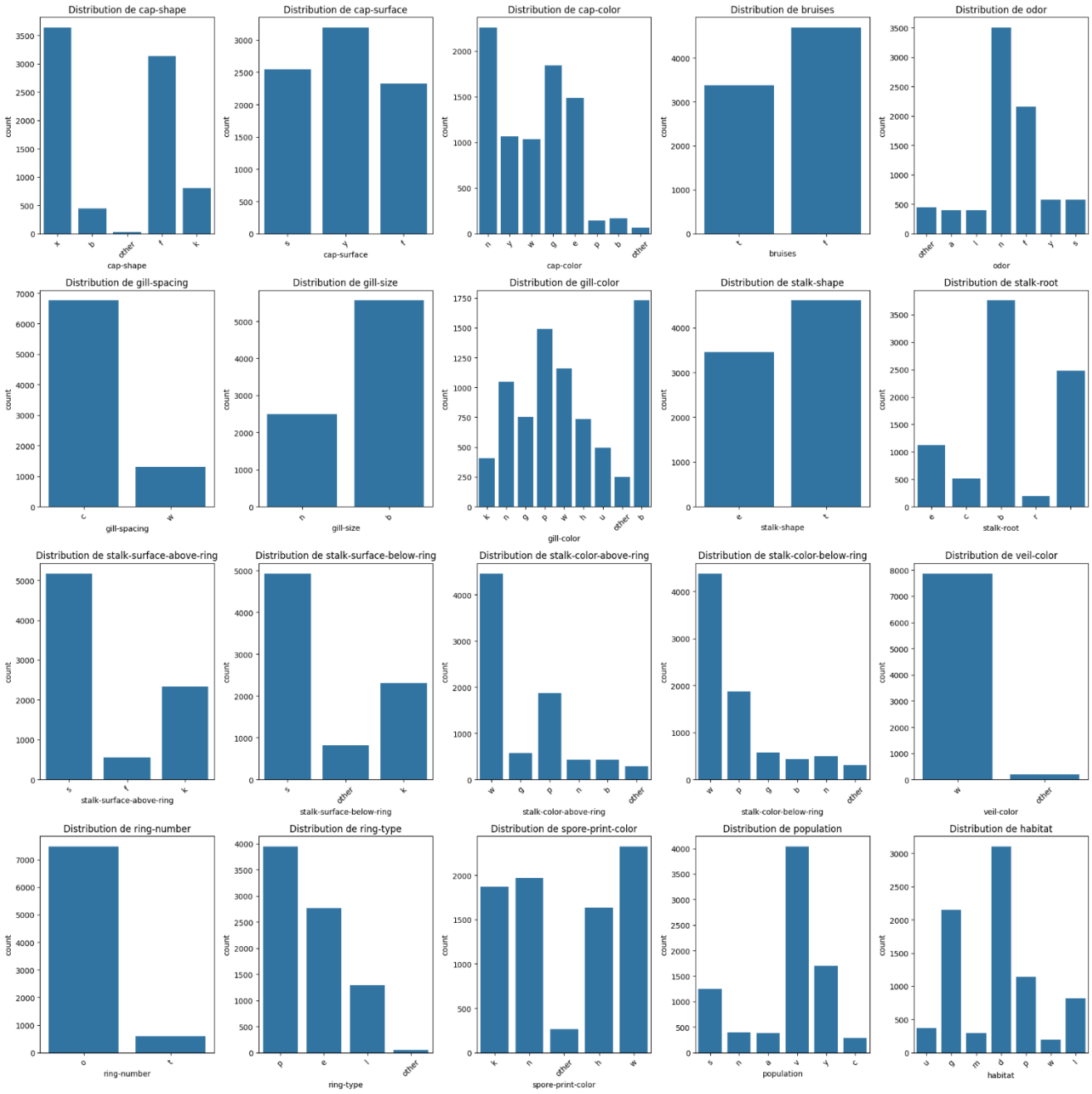
1. **Si une seule modalité était sélectionnée rares, nous l'avons supprimée** (peu représentative seule).
2. **Si plusieurs modalités étaient sélectionnées rares, nous avons regroupé les modalités restantes sous "Autre"** pour conserver de la variance.

Exemples concrets :

- **cap-shape** : Les formes (s, c) ont été regroupées sous "Autre".
- **odor** : Les odeurs "m", "p" et "c" ont été regroupées sous "Autre".
- Ring-number : Suppression de la modalité "n".

c. Résultat :

Nous nous retrouvons avec 8060 instance (après suppression des modalités "g" de cap-surface et "n" de ring-number et "y" de stalk-surface-above-ring), 20 variables (une cible) et 91 modalitées en totale



a. Etude bi- dimensionnelle

Nous allons maintenant nous intéresser à l'étude bidimensionnelle des variables. Cette analyse nous permettra d'examiner les corrélations linéaires entre notre variable cible (class) et les autres variables explicatives, deux à deux. Les dépendances les plus significatives sont présentées ci-dessous :

1. Cas 1 : Poisonous vs Cap-Shape

1 : Table de contingence pour poisonous et cap-shape :
Fréquences absolues :

cap-shape	b	f	k	other	x	Total
poisonous						
e	400	1592	224	32	1944	4192
p	45	1541	585	1	1696	3868
Total	445	3133	809	33	3640	8060

Pourcentages par colonne :

cap-shape	b	f	k	other	x
poisonous					
e	89.89	50.81	27.69	96.97	53.41
p	10.11	49.19	72.31	3.03	46.59

Pourcentages par ligne :

cap-shape	b	f	k	other	x
poisonous					
e	9.54	37.98	5.34	0.76	46.37
p	1.16	39.84	15.12	0.03	43.85

Pourcentage total :

cap-shape	b	f	k	other	x
poisonous					
e	4.96	19.75	2.78	0.40	24.12
p	0.56	19.12	7.26	0.01	21.04

Résumé :

- **Colonnes :**
  - Pour les cap-shape **b (bell)**, 89,89% des champignons non-poisonnés ont cette forme, contre 10,11% des champignons empoisonnés.
  - Les formes **k** montrent des pourcentages extrêmement élevés pour les champignons empoisonnés .

Analyse :

- Les formes **b** sont plus fréquentes parmi les champignons non-poisonnés.
- **k (conique)** sont plus susceptibles d'être associés aux champignons empoisonnés, bien que cela concerne un faible pourcentage (autour de 7%).
- L'implication : Les champignons **à forme cloche** sont nettement plus susceptibles d'être non-empoisonnés.

2. Cas 13 : Poisonous vs Stalk-Color-Above-Ring

13 : Table de contingence pour poisonous et stalk-color-above-ring :
Fréquences absolues :

stalk-color-above-ring	b	g	n	other	p	w	Total
poisonous							
e	0	576	0	288	576	2752	4192
p	432	0	432	0	1296	1708	3868
Total	432	576	432	288	1872	4460	8060

Pourcentages par colonne :

stalk-color-above-ring	b	g	n	other	p	w
poisonous						
e	0.0	100.0	0.0	100.0	30.77	61.7
p	100.0	0.0	100.0	0.0	69.23	38.3

Pourcentages par ligne :

stalk-color-above-ring	b	g	n	other	p	w
poisonous						
e	0.00	13.74	0.00	6.87	13.74	65.65
p	11.17	0.00	11.17	0.00	33.51	44.16

Pourcentage total :

stalk-color-above-ring	b	g	n	other	p	w
poisonous						
e	0.00	7.15	0.00	3.57	7.15	34.14
p	5.36	0.00	5.36	0.00	16.08	21.19

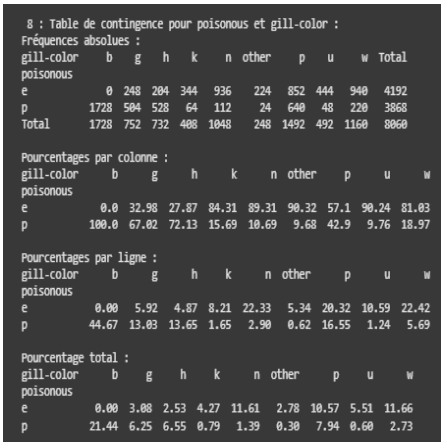
Résumé :

- **Colonnes :**
  - Pour **w (white), other et g (gray)**, ces champignons non-poisonnés ont cette couleur de pied.
  - Pour **b (buff/brun)**, les champignons non-poisonnés ont 0% de cette couleur de pied, mais les champignons empoisonnés en ont 100% (cela concerne 5,36% du total).

Analyse :

- **Les champignons empoisonnés** ont tendance à avoir des pieds **bruns** (100% de ceux avec cette couleur sont empoisonnés), ce qui pourrait être un indicateur clé.
- **Les champignons non-poisonnés** sont plus souvent associés à des couleurs comme **blanc et le gris**, et cette relation pourrait indiquer une caractéristique rassurante.

3. Cas 8 : Poisonous vs Gill-Color



```
 8 : Table de contingence pour poisonous et gill-color :
Fréquences absolues :
gill-color     b    g    h    k    n  other    p    u    w  Total
poisonous
e              0  248  204  344  936    224  852  444  940   4192
p           1728  504  528   64  112     24  640   48  220   3868
Total       1728  752  732  408 1048    248 1492  492 1160   8060

Pourcentages par colonne :
gill-color      b      g      h      k      n  other      p      u      w
poisonous
e             0.0  32.98  27.87  84.31  89.31  90.32   57.1  90.24  81.03
p           100.0  67.02  72.13  15.69  10.69   9.68   42.9   9.76  18.97

Pourcentages par ligne :
gill-color      b      g      h      k      n  other      p      u      w
poisonous
e            0.00   5.92   4.87   8.21  22.33   5.34  20.32  10.59  22.42
p           44.67  13.03  13.65   1.65   2.90   0.62  16.55   1.24   5.69

Pourcentage total :
gill-color      b      g      h      k      n  other      p      u      w
poisonous
e            0.00   3.08   2.53   4.27  11.61   2.78  10.57   5.51  11.66
p           21.44   6.25   6.55   0.79   1.39   0.30   7.94   0.60   2.73
```

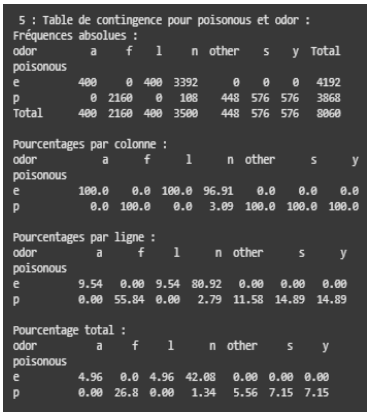
Résumé :

- **Colonnes :**
    - Pour n, k, other, u les champignons non-poisonnés ont des couleurs plus fortes que les empoisonnés.
    - Pour **b (buff)** et **h (chocolate)**, les champignons empoisonnés sont nettement beaucoup plus dominants (100% pour buff)

Analyse :

- **Les champignons avec des lamelles brunes (b)** sont significativement plus susceptibles d'être empoisonnés avec presque 22% des empoisonnés totales étant bruns.

4. Cas 5 : Poisonous vs Odor



```
 5 : Table de contingence pour poisonous et odor :
Fréquences absolues :
odor          a     f    l     n  other    s    y  Total
poisonous
e           400     0  400  3392      0    0    0   4192
p             0  2160    0   108    448  576  576   3868
Total       400  2160  400  3500    448  576  576   8060

Pourcentages par colonne :
odor          a      f      l      n  other      s      y
poisonous
e         100.0    0.0  100.0  96.91    0.0    0.0    0.0
p           0.0  100.0    0.0   3.09  100.0  100.0  100.0

Pourcentages par ligne :
odor          a      f      l      n  other      s      y
poisonous
e          9.54   0.00   9.54  80.92   0.00   0.00   0.00
p          0.00  55.84   0.00   2.79  11.58  14.89  14.89

Pourcentage total :
odor          a     f     l     n  other     s     y
poisonous
e          4.96   0.0  4.96  42.08   0.00  0.00  0.00
p          0.00  26.8  0.00   1.34   5.56  7.15  7.15
```

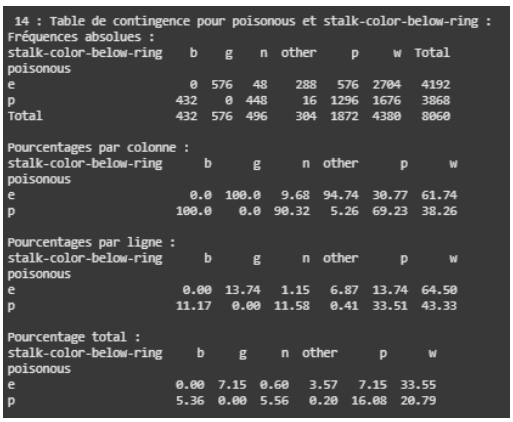
Résumé :

- **Les champignons non-poisonnés** ont un **odeur** typique "n" (naturelle) ou “a” amende ou “l” anise.
- **Les champignons empoisonnés** montrent une forte diversité, avec des odeurs "f" (foul) et "y" (fishy) et “s” spicey, chacune représentant environ 7-26% des champignons empoisonnés.

Analyse :

- Les **champignons empoisonnés** tendent à être associés à des odeurs plus variées, en particulier **foul**, tandis que **les non-poisonnés** ont principalement une odeur naturelle.
- **Les champignons non-poisonnés** peuvent généralement être distingués par leur absence d'odeurs atypiques ou des odeurs subtiles (amende, anise).

5. Cas 14 : Poisonous vs Stalk-Color-Below-Ring



```
 14 : Table de contingence pour poisonous et stalk-color-below-ring :
Fréquences absolues :
stalk-color-below-ring    b    g    n  other    p     w  Total
poisonous
e                         0  576   48    288  576  2704   4192
p                       432    0  448     16 1296  1676   3868
Total                   432  576  496    304 1872  4380   8060

Pourcentages par colonne :
stalk-color-below-ring      b      g      n  other      p      w
poisonous
e                         0.0  100.0   9.68  94.74  30.77  61.74
p                       100.0    0.0  90.32   5.26  69.23  38.26

Pourcentages par ligne :
stalk-color-below-ring      b      g      n  other      p      w
poisonous
e                       0.00  13.74   1.15   6.87  13.74  64.50
p                      11.17   0.00  11.58   0.41  33.51  43.33

Pourcentage total :
stalk-color-below-ring     b     g     n  other      p      w
poisonous
e                       0.00  7.15  0.60   3.57   7.15  33.55
p                       5.36  0.00  5.56   0.20  16.08  20.79
```

Résumé :

- **Les champignons non-poisonnés** sont plus souvent associés à des couleurs de pied **autres (other, gray)**, représentant en tout 10% des non-poisonnés.
- **Les champignons empoisonnés** ont plus de variabilité dans les couleurs de pied buff, brown, spécialement pink.

Analyse :

- **Les champignons empoisonnés** sont plus fréquemment associés à des couleurs de pied **non-grise**, en particulier **marron, brun ou rose**, tandis que les **non-poisonnés** ont tendance à être plus gris dans leurs couleurs de pied.

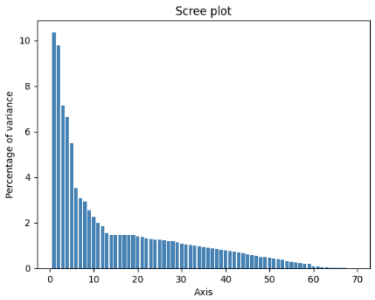
En résumé, ces données suggèrent qu'un certain nombre de caractéristiques morphologiques et olfactives peuvent être utilisées pour prédire la toxicité des champignons donc une corrélation forte entre “poisonous” et les autres variables explicatives.

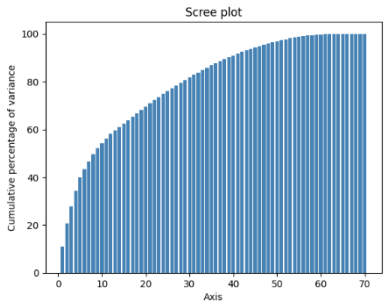
IV. ACM

Nous appliquons ici l’Analyse des Correspondances Multiples (ACM) afin de résumer les informations contenues dans un ensemble de 20 variables explicatives et de les représenter dans un espace réduit pour en faciliter l’interprétation.

Une première ACM a été réalisée avec 20 dimensions. Pour choisir le nombre d’axes à interpréter, nous nous basons sur le critère du coude appliqué à la courbe des valeurs propres cumulées. Ce critère suggère de retenir 5 axes, expliquant environ 39.42 **%** de l’inertie totale.

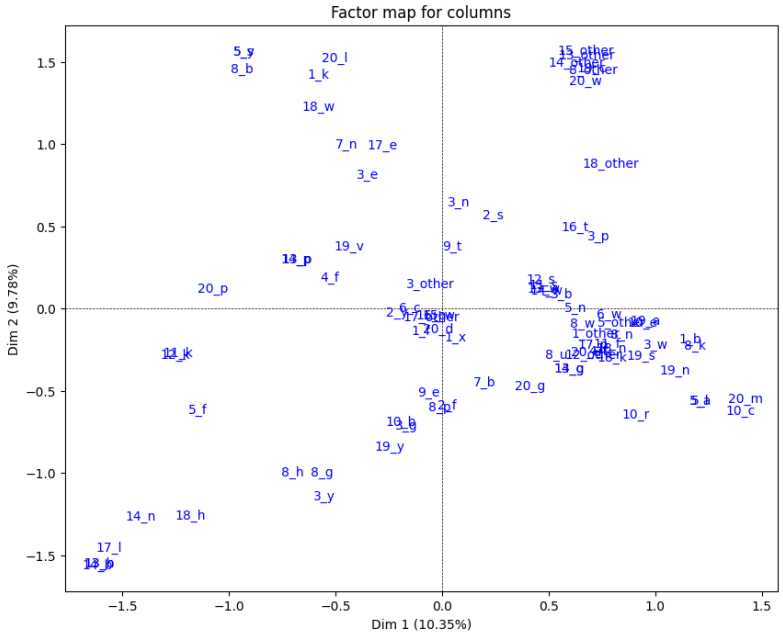
Par revanche, pour être plus précis nous avons pris les axes qui expliquent 80% de l’inertie totale qui est présenté avec 29 axes. Dans la suite du programme nous allons utiliser la configuration de 5 axes. La règle de Kaiser suggère de retenir 18 axe supérieur à 1/20 ce qui ne réduit pas suffisamment la dimensionnalité.





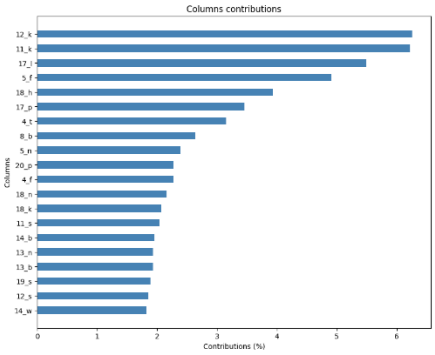
L’analyse des axes retenus met en évidence que certains plans factoriels offrent une meilleure distinction des modalités étudiées, ce qui facilite l’interprétation des corrélations sous-jacentes. Afin d’approfondir cette analyse, nous nous appuyons sur les contributions et poids des modalités, en sélectionnant uniquement celles dont l’inertie est significative par rapport à leur poids.

Nous illustrerons cette approche avec un exemple d’interprétation du plan factoriel 1-2, présenté comme suit (les numéros de colonnes sont ajoutées afin de distinguer les modalités de chaque variable) :



Contributions:

Les graphes suivants illustrent les contributions sur chaque axe des 20 meilleurs modalitées par axe:



Axe 1 :

Les modalités avec les contributions les plus importantes sont : 12_k(stalk-surface-below-ring k=**silky**), 11_k (stalk-surface-abovering k=**silky**), 17_l(ring type l=**large**),5_f (odor f=**foul**),18_h (spore-print-color h=**chocolate**),17_p (ring-type p=**pendant**),4_t (bruises t=**bruises** true), 8_b(gill color b =buff), 5_n (odor n=none)

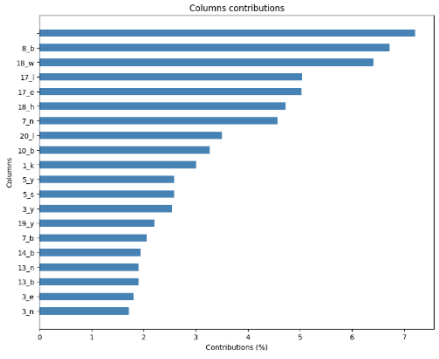
stalk-surface-above-ring_k	-1.238365
stalk-surface-above-ring_k	-1.238365
ring-type_l	-1.562710
odor_f	-1.144342
spore-print-color_h	-1.178028
ring-type_p	0.710249
bruises_t	0.733698
gill-color_b	-0.937535
odor_n	0.626634

L'axe 1 met en évidence une opposition marquée entre certaines modalités des variables analysées. En examinant les contributions des modalités à cet axe, on observe une séparation entre deux groupes de champignons :

- D'un côté, les champignons associés aux modalités stalk-surface-above-ring_k (soyeux), ring-type_l (grand), odor_f (odeur fétide), spore-print-color_h (impression de spores chocolat), et gill-color_b (lames brunâtres) possèdent des coordonnées négatives élevées.
- De l'autre côté, les champignons liés aux modalités ring-type_p (pendant), bruises_t (présence de bleuissement) et odor_n (absence d'odeur marquée) affichent des coordonnées positives sur cet axe.

Cette opposition suggère une différenciation nette entre deux types de champignons. Les champignons caractérisés par une surface soyeuse sous l'anneau, un anneau large, une odeur désagréable et une impression de spores chocolatées sont plutôt distincts de ceux qui ont un anneau pendant, présentent un bleuissement et n'ont pas d'odeur marquée.

D'un point de vue biologique, cette distinction pourrait refléter une séparation entre des espèces potentiellement toxiques ou non comestibles (souvent associées à des odeurs fortes et des spores foncées) et des espèces plus neutres ou comestibles (sans odeur marquée et avec des réactions visibles comme le bleuissement).



Axe 2 :

Les modalités avec les contributions les plus importantes sont : 8_b(gill color b =buff), 18_w (sspore-print-color w=white), 17_l(ring type l=**large**),17_e (ring type e= evanescent),18_h (spore-print-color h=**chocolate**),7_n (gill-sizee n=narrow),20_l (habitat l=**leaves**), 10_b (stalk-root b=bulbous) , 1_k (cap-shape k= knobbed)

gill-color_b	1.454643
spore-print-color_w	1.224889
ring-type_l	-1.454849
ring-type_e	0.993799
spore-print-color_h	-1.255713
gill-size_n	0.996956
habitat_l	1.524631
stalk-root_b	-0.688405
cap-shape_k	1.420911

L’axe 2 met en évidence une opposition marquée entre deux groupes de champignons selon leurs caractéristiques morphologiques et écologiques.

Cette opposition suggère une différenciation entre deux groupes de champignons.

- Le premier groupe (coordonnées positives) semble associé à des espèces ayant des lamelles de couleur buff, une empreinte de spores blanche, un anneau évanescent, des lamelles étroites et vivant sous les feuilles. Cette combinaison pourrait correspondre à des champignons plus fragiles ou à une certaine famille écologique spécifique.
- Le second groupe (coordonnées négatives) est marqué par un anneau large, une empreinte de spores chocolat et une racine bulbeuse, ce qui pourrait être indicatif d’un autre type de champignons, peut-être plus robustes et adaptés à un environnement différent.

D’un côté, les champignons associés à des modalités positives présentent des caractéristiques qui pourraient être liées à un habitat spécifique et à une structure plus fine ou délicate, comme des lamelles étroites, une empreinte de spores blanche et une croissance sous les feuilles. Ces éléments suggèrent une adaptation à des environnements plus abrités ou humides.

À l’opposé, les champignons associés aux modalités négatives se distinguent par des traits plus robustes, notamment un anneau large et une empreinte de spores chocolat, souvent observés chez des espèces plus résistantes ou mieux adaptées à des conditions plus ouvertes. Cet axe traduit ainsi une différenciation entre des champignons plus fragiles et ceux possédant des structures plus marquées, ce qui peut refléter des stratégies évolutives distinctes.

Analyse de la variable cible:

D'après les valeurs test obtenues , nous pouvons dégager une analyse approfondie des axes factoriels et leur lien avec la variable cible.

Modalité	Axe 1	Axe 2
p	-58.724412	5.516540
e	137.337689	-12.220535

Analyse du premier axe factoriel (Dim 1)

L'axe 1 oppose nettement les champignons comestibles (*e*) aux champignons toxiques (*p*), comme en témoignent leurs valeurs test respectives.

- **Champignons comestibles (*e*)** : Ils sont fortement associés aux modalités **habitat = leaves, gill color = buff, cap shape = knobbed**, qui présentent des coordonnées positives sur cet axe. Ces champignons sont donc plus fréquemment trouvés dans des environnements boisés et feuillus, avec des caractéristiques morphologiques spécifiques comme un chapeau bosselé et des lames de couleur beige.
- **Champignons toxiques (*p*)** : Ils se caractérisent par des modalités telles que **spore-print-color = white, ring-type = large, spore-print-color = chocolate**, qui sont fortement négatives sur cet axe. Ces champignons ont une impression de spores blanche ou chocolat, ce qui pourrait être un indicateur clé pour les différencier des espèces comestibles.

En résumé, l'axe 1 met en évidence une séparation entre les champignons trouvés en milieu forestier avec un chapeau bosselé et ceux possédant des spores blanches ou chocolat, souvent associés à des espèces toxiques.

Analyse du deuxième axe factoriel (Dim 2)

L'axe 2 présente une opposition plus subtile entre certaines modalités :

- **Champignons comestibles (*e*)** : Ils sont associés aux modalités **ring-type = evanescent et gill-size = narrow**, qui sont bien représentées sur cet axe. Cela signifie que les champignons comestibles ont souvent un anneau évanescent et des lames étroites, ce qui peut être une caractéristique distinctive.

- **Champignons toxiques (*p*)** : Ils sont liés aux modalités **stalk-root = bulbous**, qui montrent une corrélation négative avec cet axe. Cela signifie que les champignons toxiques ont plus souvent une base bulbeuse, un élément clé à considérer dans l'identification des espèces potentiellement dangereuses.

En conclusion, l'axe 2 permet de différencier les champignons selon des critères morphologiques plus fins, mettant en avant des différences dans la structure de l'anneau et des lames.

Synthèse globale

Les axes factoriels de l'ACM offrent une séparation claire entre les champignons comestibles et toxiques. Le premier axe est principalement lié à l'habitat et à la couleur des spores, tandis que le second axe met en avant des distinctions morphologiques comme la taille des lames et la structure de la base du pied. Ces résultats permettent d'affiner notre compréhension des caractéristiques les plus discriminantes entre espèces comestibles et toxiques, facilitant ainsi leur identification.

V. Analyse discriminante

A ce stade la , on va faire une étude approfondie visant à examiner la variable « class » en lien avec les autres facteurs explicatifs. Cette démarche repose sur une comparaison détaillée de multiples approches, s'appuyant sur l'analyse des matrices de confusion associées à la variable cible. L'objectif principale c'est d'affiner l'évaluation de la précision pour chaque classe, tout en améliorant notre perception des structures sous-jacentes des données.

Dans un premier temps, l'Analyse en Composantes Multiples (ACM) requiert une identification des dimensions les plus pertinentes pour distinguer la variable cible « class ». Après cette étape de sélection, nous avons choisi de conserver 15 dimensions parmi les 23 retenues initialement, en nous basant sur le critère de Kaiser. Les dimensions qui s'avèrent les plus efficaces pour différencier les individus selon leur classe sont les suivantes, par ordre d'importance :
{Dim1, Dim6, Dim3, Dim2, Dim5, Dim7, Dim4, Dim9, Dim20, Dim18, Dim19, Dim8, Dim17, Dim10, Dim15.}

Nous pouvons vérifier le pouvoir discriminatoire de ces dimensions à travers le trace du nuage de points des individus, avec comme couleur la classe associée

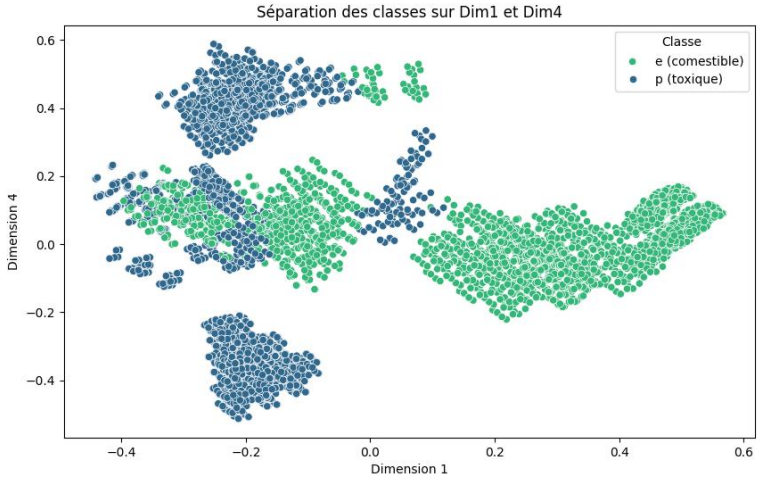


Figure : Scatter plot sur les deux dimensions Dim1 et Dim4 des individus avec la classe associ´ee selon la couleur

Les axes discriminants permettent d'aborder la classification supervisée. Pour cela, les 81 dimensions sont conservées, et une première approche basée sur la décision bayésienne est appliquée :

Nombre d'observations et pourcentage classifiés dans class			
	e	p	Totale
e	3712 100.00	0 0.00	3712 100.00
p	0 0.00	3792 100.00	3792 100.00

Totale	3712 49.47	3792 50.53	7504 100.00
A priori	0.5	0.5	

L'analyse de la table de confusion montre une précision de 100 % pour les classes comestibles et toxiques. Une analyse discriminante a ensuite été réalisée en utilisant la méthode des k plus proches voisins, avec k=5, sélectionné sur la base des variables les plus discriminantes.

⇨ Les résultats obtenus indiquent un taux d'erreur de classification de 0 %, confirmant que les deux classes sont parfaitement distinguables grâce à cette méthode, comme illustré dans la figure suivante.

Nombre d'observations et pourcentage classifiés dans class			
	e	p	Totale
e	3712 100.00	0 0.00	3712 100.00
p	0 0.00	3792 100.00	3792 100.00
Totale	3712 49.47	3792 50.53	7504 100.00
A priori	0.5	0.5	

À l'instar de la discrimination linéaire, l'algorithme des k-plus-proches-voisins (kPPV) fournit des résultats sans erreur (0 %), confirmant ainsi la bonne séparabilité des données dès les premiers axes de l'ACM.

VI. CAH

Maintenant, l'objectif est simple : distinguer les comestibles des toxiques grâce à la variable « outcome ». Cette fois, on passe à une approche de clustering pour regrouper les individus et leurs caractéristiques. Pour trouver le nombre idéal de clusters, on a utilisé une Classification Ascendante Hiérarchique (CAH), en se concentrant sur les deux premières dimensions principales.

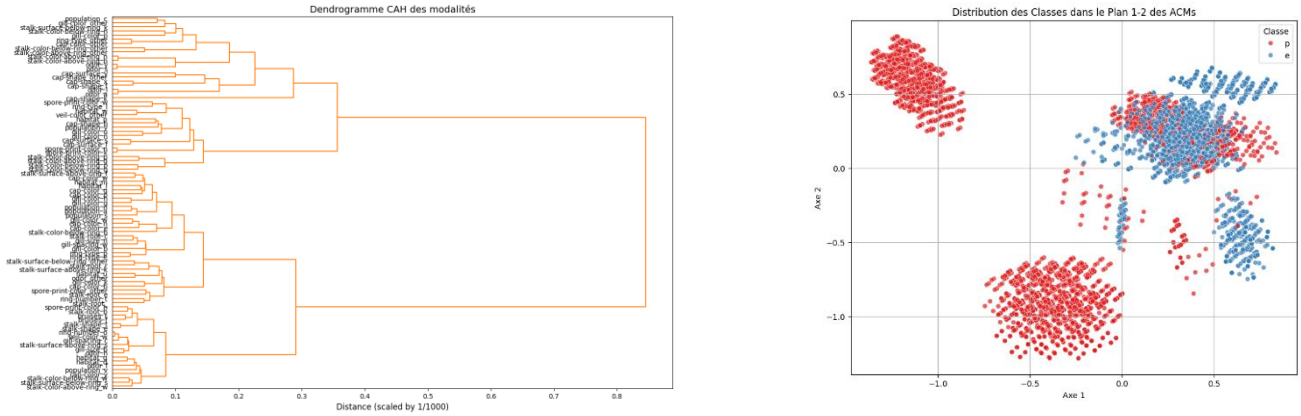


Figure -Résultats de la CAH sur les modalités

En examinant le dendrogramme et les différentes étapes d'agrégation, nous constatons que le nombre optimal de groupes est de trois et non de deux. Cette divergence s'explique par la grande dispersion des individus appartenant aux classes comestibles. Il est fort probable que ce troisième groupe corresponde aux individus situés dans une zone intermédiaire entre les deux classes principales.

Afin d'évaluer la pertinence de la classification et d'interpréter les classes résultant de ce clustering, nous représentons la distribution des classes sur le plan 1-2 des ACMs. La figure à droite illustre cette répartition.

⇨ Nous avons remarqué que la distribution des groupes suit parfaitement la distribution des classes comestibles et toxique, ce qui est prévue précédemment.

Conclusion

L'objectif de ce travail était d'approfondir la maîtrise des Concept Statistique ainsi que des procédures d'analyse des données qualitatives. Cela nous a permis de mettre en application nos connaissances théoriques sur l'analyse des correspondances multiples. Par ailleurs, l'exploitation de la base de données nous a permis d'acquérir de nouvelles connaissances, particulièrement utiles dans la vie courante, notamment pour distinguer les champignons comestibles des espèces toxiques.

Dépôt :

https://github.com/amenallah01/Methodes-statistiques-pour-les-donnees-qualitatives.git