**Université Grenoble Alpes**
**L2/STA351**

---

# TP2 : Description d'une variable quantitative et calculs de probabilités

*Objectifs : savoir produire les résumés numériques ou graphiques standards pour décrire une variable quantitative discrète ou continue et savoir utiliser les fonctions de R : densité, fonction de répartition (FdR) et fonction quantile pour les lois usuelles*

## 1 Statistiques descriptives

### Exercice 1

On étudie l'âge de mères non-fumeuses au moment de leur accouchement.

| age | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| effectifs | 7 | 8 | 9 | 10 | 12 | 3 | 2 | 5 | 4 | 5 | 2 | 4 | 2 | 1 | 1 |

1. Nature ce la variable ?
2. Créer un vecteur avec l'ensemble de ces données :
   `ages <- c(rep(21,7), rep(22,8),rep(23,9),rep(24,10),rep(25,12), rep(26,3),rep(27,2), rep(28,5),rep(29,4),rep(30,5),rep(31,2), rep(32,4),rep(33,2), 34, 35)`
   Afficher le vecteur `ages`
3. Combien d'individus y a-t-il dans la base ? Vous pouvez utiliser la commande `length`.
4. Calculer les effectifs de chaque modalité avec :
   `table(ages)`
5. Calculer les proportions (fréquences) de chaque age. Quelle est la proportion de mères de 25 ans ? de 30 ans ? de 35 ans ?
6. Combien de femmes ont moins de 30 ans ? plus de 28 ans ? Quelle proportion de femmes ont moins de 30 ans ? plus de 28 ans ? (on pourra utiliser `table()` et les vecteurs `ages<30`,`ages>28`).
7. Trouver l'âge en dessous duquel on ne trouve que 5% des observations :
   `quantile(ages,0.05)`
   Quel est l'âge au delà duquel on trouve 5% des femmes non-fumeuses qui accouchent ?
8. Quel est l'âge moyen des mères non-fumeuses au moment de leur accouchement ?
   `mean(ages)`
9. Calculer la variance empirique corrigée notée $s'^2$ ($s'^2 = n/(n-1)s^2$).
   `var(ages)`
10. Calculer l'écart-type empirique corrigé $s'$ avec :
    `sd(ages)`
    Interpréter.
11. Calculer la médiane et les quartiles de l'échantillon :
    `summary(ages) # quelques indicateurs statistiques`
    `median(ages) # mediane`
    Retrouver les résultats avec la fonction `quantile`.

12. Tracer les fréquences empiriques avec un diagramme en barres :

    `barplot(table(ages))`

    Attention : la fonction `barplot()` a été écrite pour représenter des variables qualitatives. Ainsi les modalités observées de la variable n'étant ni numériques ni ordonnées il n'y a pas d'unité sur l'axe des abscisses et les barres sont placées à égales distances les unes des autres et étiquettées selon la modalité qu'elles représentent. Dans le cas où on travaille avec une variable discrète qui a des modalités régulièrement espacées cette fonction produira un diagramme en barre juste. Par contre si ce n'est pas le cas la répartition représentée sera fausse (l'axe des abscisse n'ayant pas d'unité) et dans ce cas il est préférable d'utiliser :

    `plot(table(ages)/n,type="h",lwd=10, ylab="fréquences")`

13. Tracer la fonction de répartition empirique avec `plot(ecdf(ages))`. Superposer une ligne verticale bleue représentant la médiane, deux vertes représentant le premier et troisième quartiles. On pourra utiliser la fonction de R `abline` :

    ```
    abline(h=0.5,col="red") # ligne horizontale à 0.5
    abline(v=median(ages),col="red") # ligne verticale à la médiane
    abline(h=0.25,col="blue") # ligne horizontale à 0.25
    abline(v=quantile(ages,0.25),col="blue") # ligne verticale au premier quartile
    abline(h=0.75,col="green") # ligne horizontale à 0.75
    abline(v=quantile(ages,0.75),col="green") # ligne verticale au 3eme quartile
    ```

14. Tracer la boite à moustaches de l'échantillon.

    `boxplot(ages)`

    Que représente chaque ligne ?

    Superposer une ligne horizontale rouge représentant la moyenne, une bleue représentant la médiane, deux vertes représentant le premier et troisième quartiles.

    `boxplot(ages)`

15. Sur quel graphique peut-on lire la fréquence de l'intervalle $[22; 25]$ ?

**Exercice 2**

Le jeu de données proteine.csv provient d'une étude menée par Hunault et Jaspard sur les protéines LEA (Late Embryogenesis Abundant proteins). Ce sont des protéines qui contribuent, principalement chez les végétaux, à l'acquisition de la tolérance à la dessiccation, en particulier dans le cas de déshydratation ou de stress induit par le froid. Elles sont assez mal connues et encore moins bien classifiées.
La base contient les variables suivantes

`longseq` : longueur de la séquence (en acides aminés)
`regne` : règne associé : Viridiplantae (plante verte), Fungi (champignon), Metazoa (animal), Bacteria (bactérie)
`repliement` : indice de repliement
`isoelec` : point isoélectrique ou pHi
`poidsmol` : poids moléculaire (en g/mol)
`hydro` : hydropathie moyenne, échelle de KD
`genre` : taxonomie

Enregistrer la base dans votre répertoire de travail TP2.

1. Charger la base `proteine.csv`, et lui donner le nom `data`.
2. Calculer sa dimension, afficher les 10 premières lignes, les données de la ligne 2, 4, 5 et les colonnes 5 et 6.
3. Quelles sont les noms des colonnes ?
4. On s'intéresse à la variable `isoelec`. On appellera `I` cette variable. Extraire les données de cette variable et calculer les indicateurs statistiques standards.

5. Tracer un histogramme en effectifs puis en fréquences de `I` et commenter (utiliser la fonction `hist()`.

   Ajouter une ligne rouge avec la valeur moyenne. Ajouter une ligne bleue avec la valeur médiane, et deux lignes vertes avec les quartiles. Ajouter une courbe de densité d'une variable gaussienne de moyenne et écart-type bien choisis (on pourra utiliser les fonctions de R : `points()` ou encore `curve()`). Semble-t-il raisonnnable de modéliser cette variable avec une loi normale ?

6. Tracer un histogramme en fréquence avec 40 classes et commenter.

   ```
   hist(I, nclass=40, prob=T)
   ```

7. Représenter la distribution à l'aide d'une boite à moustaches. Comparer avec les boites à moustaches des distributions pour les Viridiplantae et les Metazoa.

   ```
   IV<-data[data$regne=="Viridiplantae",4]
   IM<-data[data$regne=="Metazoa",4]
   boxplot(IV, IM, names=c("Viridiplantae", "Metazoa"), ylab="isoelec")
   ```

8. Tracer la fonction de distribution empirique cumulée de la variable `I`.

# 2 Probabilités et tirages aléatoires

R permet de calculer exactement les fonctions de densité, les fonctions de distribution cumulée, les fonctions quantiles pour les familles de probabilités classiques. Par exemple,

- distribution normale :
  - La fonction de densité est `dnorm`,
  - sa fonction de distribution cumulée est `pnorm` (en entrée une valeur, en sortie une probabilité),
  - sa fonction quantile est `qnorm` (en entrée une probabilité, en sortie la valeur quantile).
- distribution binomiale
  - La fonction de densité est `dbinom`,
  - sa fonction de distribution cumulée est `pbinom`,
  - sa fonction quantile est `qbinom`,

Regarder dans l'aide quels sont les paramètres de ces fonctions.

## Exercice 3

Soit $X$ une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = 2$ et $\sigma = 1/2$.

1. Affecter à `mu` la valeur 2 et à `sig` la valeur 0.5. Dans la suite on écrira toutes les lignes de commandes en utilisant `mu` et `sig` plutôt que les valeurs qu'on leur a affectées dans l'exercice.
2. Tracer sur un même graphique la fonction densité de $X$ et sa FdR. On pourra utiliser la fonction `curve` avec l'option `add=T` dans la ligne de commande demandant la superposition de la seconde courbe tracée.
3. Calculer le quantile d'ordre 0.6 et le placer en rouge sur le précédent graphique avec `points` et le symbole de votre choix. Où doit-on lire la valeur 0.6 ? Tracer les deux lignes verticales et horizontales qui permettent de guider la lecture du quantile d'ordre 0.6 (on pourra représenter ces lignes en traits pointillés avec l'option `lty=2`).

R permet aussi de réaliser des tirages aléatoires, soit dans un ensemble de valeurs données (avec ou sans remise), soit à partir d'une loi connue

- `sample(x, N, replace=TRUE)` échantillonne aléatoirement `N` valeurs dans le vecteur `x` avec remise
- `sample(x)` échantillonne aléatoirement toutes les valeurs de `x` sans remise, ce qui revient à permuter aléatoirement le vecteur `x`.
- `rnorm(N)` simule un échantillon de taille `N` dans la loi normale centrée réduite
- `rbinom(N, n, p)` simule un échantillon de taille `N` dans la loi binomiale de paramètres `n` et `p`

## Exercice 4

1. Définir `x` le vecteur `c(0,1,1)`.
2. Permuter aléatoirement les trois valeurs de `x` avec `sample(x)`

   Répéter 20 fois. Combien y at-il de permutations possibles et combien de fois avez vous obtenu chacune d'entres elles ? Que peut-on déduire de cette expérimentation ?
3. Pour `n=100`, créer un vecteur `E` d'un échantillon de `n` valeurs tirées aléatoires dans `x` mais **avec remise**.

   `E <- sample(x,n,replace=TRUE)`
4. Calculer les effectifs et les proportions de 0 et de 1 dans `E`. Tracer un diagramme en barre. Quelle est selon vous la loi de variable qui a permis de simuler ces $n$ tirages (on pourra soit l'observer statistiquement en prenant $n$ grand soit l'établir par un calcul simple de probabilités). Que vaut son espérance ?
5. Calculer la moyenne de `E`. Est ce que la différence entre la moyenne de l'échantillon simulé `E` et l'espérance précédente est grande ? Que se passe-t-il lorsque $n$ augmente (on pourra refaire tourner le script pour des valeurs de $n$ de plus en plus grandes) ?
6. Reprendre l'exercice à partir de la question 3 en remplaçant la commande créant `E` par

   `E <- sample(c(0,1),n,replace=TRUE, prob=c(1/3,2/3))`

   L'échantillon ainsi produit est-il tiré avec la même loi que dans la question 3 ?