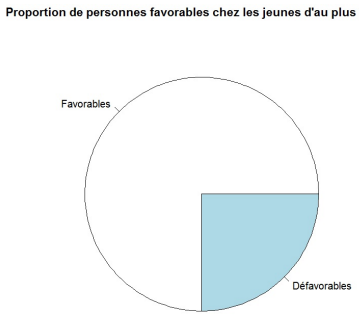
TP n° 2 : Un peu de Statistique

Exercice 1. On dispose du prénom, de la taille en centimètres, du salaire annuel net en euros, de la réponse à la question : "Êtes-vous favorable à la réforme relative aux professions réglementées actuellement discutée ?" et de la date de naissance de 12 personnes :

Bob 187 32000 non 06/02/1967
Bill 182 18000 oui 07/02/1992
Eva 167 35000 oui 5/09/1983
Ines 156 21500 non 8/02/1976
Léa 162 41500 oui 9/02/1959
Léo 177 20000 oui 17/06/1987
Lina 175 24000 non 31/01/1990
Tom 192 29000 non 6/02/1982
Théo 172 14500 oui 14/03/1974
Sarah 160 36000 oui 28/03/1987
Yann 179 28000 non 25/02/1971
Zoé 168 34500 oui 11/03/1984

1. Créer un dossier appelé "TP" sur l'ordinateur et le définir comme répertoire courant de R.
2. Copier le tableau ci-dessus dans un document texte (.txt) qui sera enregistré dans le dossier TP et créer sous R une table `Table` qui contient ces données.
3. Donner à chaque ligne le nom de l'individu correspondant et aux variables les noms "Nom", "Taille", "Salaire", "Favorable" et "Date". On définit ainsi les variables `Nom`, `Taille`, `Salaire`, `Favorable` et `Date`. Supprimer la variable `Nom`.
4. La date de naissance de Sarah est erronée : elle est née le 28/01/1987. Modifier dans `Table` la date de naissance de Sarah. On pensera, si nécessaire, à modifier les niveaux de la variable concernée.
5. Convertir la variable `Date` pour que chaque élément soit un objet de classe Date.
6. On s'aperçoit que l'on a oublié une treizième personne : Jean 174 28500 non 22 septembre 1981. Compléter la table.
7. Créer un vecteur `Nais` contenant les données de la variable `Date`, mais contenant des objets de classe POSIXlt (on utilisera `strptime`).
8. Créer, à partir du vecteur `Nais` et de la date courante (`Sys.time()` que l'on convertira en objet de la classe `POSIXlt`), un vecteur `Age1` contenant l'âge de chaque individu. Ajouter ce vecteur à `Table`, et appeler la nouvelle variable `Age`.
9. Trier `Table` pour que les noms soient dans l'ordre alphabétique.
10. Ajouter à la table la variable `Sup30` qui prend la valeur TRUE si la personne a 31 ans ou plus et FALSE si la personne a 30 ans ou moins. Convertir la variable `Favorable` au format booléen (logical).
11. Créer deux vecteurs identiques `Prop1` et `Prop2`, contenant chacun la proportion des personnes favorables à la loi chez les moins (inférieur ou égal) de 30 ans et chez les plus de 31 ans de deux façons différentes : en utilisant des produits vectoriels d'une part, et en utilisant la fonction `tapply` d'autre part.
12. Réaliser le diagramme circulaire ci-dessous. Faire sur le même modèle celui pour les plus de 30 ans.



13. Représenter le salaire en fonction de l'âge sous forme d'un nuage de points. Les axes s'appelleront "Age" et "Salaire", et ces noms seront écrits en vert foncé. Le titre "Salaire en fonction de l'âge" sera écrit en bleu foncé.
14. Créer un vecteur `IC` qui contiendra les deux bornes de l'intervalle de confiance pour la proportion de personnes favorables à la réforme en utilisant une approximation normale. Cet intervalle est-il acceptable ?
15. Calculer de trois façons différentes la corrélation entre les variables `Salaire` et `Age` : en utilisant la fonction `cor`, en utilisant les fonctions `cov` et `var`, et enfin en faisant des produits scalaires. Quelle formule est utilisée par R pour calculer la covariance et pour calculer la corrélation ?

Exercice 2.

1. Charger la table `iris`. Créer 3 tables `Seto`, `Vers` et `Virg` contenant les 4 premières variables de la table iris pour chacune des espèces. On n'utilisera pas le fait que les espèces figurent les unes à la suite des autres, avec 50 observations pour chaque espèce.
2. R dispose d'une fonction qui permet de réaliser le test de Shapiro-Wilk servant à tester la normalité d'un échantillon. Créer 3 vecteurs `Norm1`, `Norm2` et `Norm3` contenant, pour chacune des trois variétés, la p-valeur associée au test de normalité pour les 4 variables.
3. Faire un histogramme de la distribution de la quatrième variable pour la première variété : visuellement, peut-on penser que cette variable est normale ?
4. Créer trois vecteurs `Vse`, `Vve` et `Vvi` contenant, pour chaque échantillon, les variances des 4 variables.
5. Créer 3 vecteurs `Test12`, `Test23` et `Test13` de quatre éléments chacun et contenant pour chaque couple d'échantillons, la p-valeur associée au test d'égalité des variances de Fisher pour chacune des 4 variables (le premier élément de `Test12` sera la p-valeur du test d'égalité de la variance de la première variable dans l'échantillon des Setosa et dans celui des Versicolor). On réalisera ce test sans utiliser la fonction dédiée dans R : on utilisera les vecteurs `Vse`, `Vve` et `Vvi`, ainsi que la fonction quantile pour une loi de Fisher.
6. D'après les résultats obtenus dans la question précédente, peut-on considérer que chaque variable a la même variance dans chaque groupe ? Était-il légitime de réaliser un test de Fisher ?
7. En utilisant la fonction de R qui réalise le test de Fisher, vérifier que la valeur obtenue en premier élément de `Test12` est bonne.
8. En acceptant l'hypothèse d'égalité des variances de la seconde variable dans l'échantillon de la variété Versicolor et celui de la variété Virginica ainsi que l'hypothèse de normalité de cette variable dans chacun de ces échantillons, réaliser un test d'égalité des moyennes sans utiliser la fonction de test de Student de R. Comparer la p-valeur obtenue à celle renvoyée par R en utilisant la fonction de test de Student. Que devient cette dernière p-valeur si on suppose que les variances ne sont pas égales ?