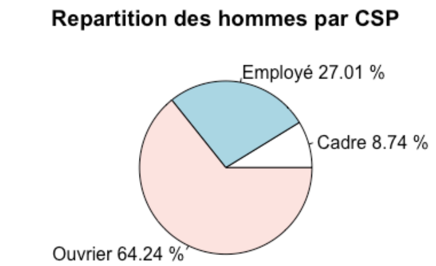
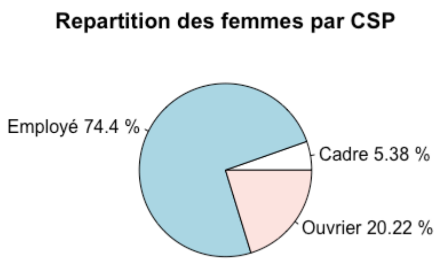
JAULGEY Thomas CCR DE2
DE CARVALHO Quentin

Question 1 - Représenter graphiquement la répartition par CSP chez les hommes et chez les femmes.





A l'aide de ces 2 graphes montrant la répartition par CSP des hommes et des femmes, on remarque qu'il y a 3 fois plus d'hommes ouvriers que de femmes ouvrières. A l'inverse, on constate qu'il y a environ 3 fois plus de femmes employées que d'hommes employés. En ce qui concerne les cadres, la proportion est similaire que ce soit chez les hommes ou chez les femmes.

Question 2 - Exécuter un test permettant de déterminer si la répartition par CSP est significativement différente entre les deux sexes.

```
> SummaryCSPSexe
        groupe
rhesus   Cadre   Employ<U+00E9> Ouvrier
  Hommes    89             275     654
  Femmes    29             401     109
> chisq.test(SummaryCSPSexe, cor=FALSE)

        Pearson's Chi-squared test

data:  SummaryCSPSexe
X-squared = 326.85, df = 2, p-value < 2.2e-16

> Resultats <- chisq.test(SummaryCSPSexe)
> Resultats$expected
        groupe
rhesus      Cadre   Employ<U+00E9>  Ouvrier
  Hommes 77.15093        441.9833 498.8658
  Femmes 40.84907        234.0167 264.1342
```

Nous cherchons à vérifier si la répartition de CSP est significativement différente entre les sexes.
Pour selon nous posons l'hypothèse suivante :
H0 : La répartition par CSP n'est pas significativement différente entre les deux sexes
H1 : La répartition par CSP est significativement différente entre les deux sexes

Pour vérifier cette hypothèse nous avons décidé d'effectuer un Test de CHI 2

Comme tout test de Chi 2 nous avons commencé par créer notre table de contingence avec 3 colonnes et deux lignes. Pour se faire nous avons utilisé la commande summary afin de récupérer la somme de personne dans catégorie professionnel.

```
SummaryCSPSexe <- matrix(c(summary(CSPHommmes$Hommes.CSP), summary(CSPFemmes$Femmes.CSP) ), byrow=TRUE, nrow=2)
dimnames(SummaryCSPSexe)=list(rhesus=c("Hommes","Femmes"), groupe=names(summaryCSPFemmes))
SummaryCSPSexe
```

Cette table de contingence est donc sous la forme suivante :

```
> Resultats <- chisq.test(SummaryCSPSexe)
> Resultats$expected
        groupe
rhesus       Cadre  Employé  Ouvrier
  Hommes 77.15093 441.9833 498.8658
  Femmes 40.84907 234.0167 264.1342
> SummaryCSPSexe
        groupe
rhesus   Cadre Employé Ouvrier
  Hommes    89     275     654
  Femmes    29     401     109
> 
```

Nous affichons la table de valeur théorique effectué par le test de CHi2, par ailleurs nous avons pu remarquer sur la table de contingence qu'aucune des catégories avait des ensembles de moins de 5 individus sinon nous aurions dû effectuer des regroupements

Le test de Chi 2 renvoie alors ce résultat

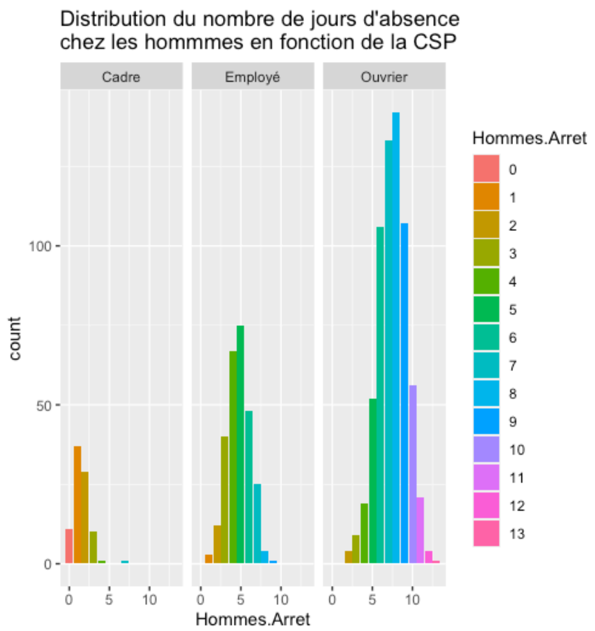
```
        Pearson's Chi-squared test

data:  SummaryCSPSexe
X-squared = 326.85, df = 2, p-value < 2.2e-16
```

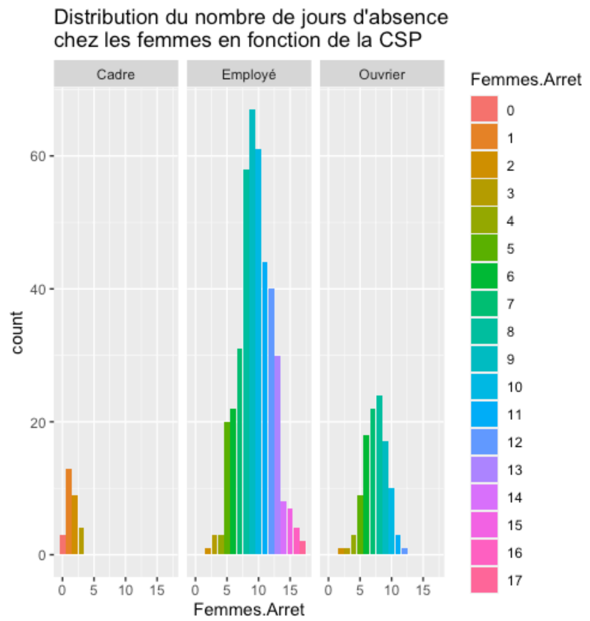
Nous voyons donc une p-value excessivement faible. Nous pouvons donc rejeter H0 avec un taux d'erreur de 5% Ce qui revient à dire que la répartition par CSP est significativement différente en fonction des sexes.

L'autre moyens de pouvoir affirmé cela sera de calculer la zone critique de notre test de chi2. En lisant dans la table de cette loi nous pouvons dire que W=]5.9:+oo] à ddl2,2

Nous pouvons donc dire que la répartition par CSP est significativement différente en fonction du sexe ce qui correspond bien aux données nationales concernant les proportions de cadre, employé et ouvrier.

Question 3 - Dans l'échantillon des hommes (resp. des femmes), représenter graphiquement la distribution du nombre de jours d'absence en fonction de la CSP.





A l'aide de ces 2 graphes montrant la distribution du nombre de jours d'absence en fonction de la CSP, on constate assez aisément que les cadres sont les personnes les moins absentes chez les hommes comme chez les femmes. On constate que chez les hommes, les ouvriers sont ceux qui sont le plus absents. Cela peut potentiellement être dû à la dangerosité de leur métier. Cependant, chez les femmes, on constate que ce sont les employées qui sont les plus absentes. On peut lier cela au fait que les femmes disposent de congés maternité à la naissance de leurs enfants et que les femmes employées sont celles dont la situation serait la plus propice à des naissances. Cela pourrait également expliquer le fait que l'on observe 4 jours d'écart entre le maximum de jours d'arrêt entre les femmes et les hommes.

Question 4 - Exécuter une analyse permettant de déterminer si chez les hommes (resp. chez les femmes), le nombre de jours absence dépend significativement de la CSP.

Pour répondre à cet exercice nous voulions effectuer un test anova afin de voir l'importance de la CSP sur le nombre de jour d'arrêt

Afin de faire notre test anova nous devions vérifier la normalité des échantillons ainsi que l'homogénéité des variances

Nous effectuons donc un test de normalité et un test de bartlett.

Nous avons donc séparé les données en fonction des hommes et des femmes puis nous avions effectué nos tests de Shapiro qui en renvoyer ces résultats.

```
 --------
 Femmes.CSP = Cadre

        Shapiro-Wilk normality test

data:  Femmes.Arret
W = 0.87045, p-value = 0.002064

 --------
 Femmes.CSP = Employé

        Shapiro-Wilk normality test

data:  Femmes.Arret
W = 0.98481, p-value = 0.0003243

 --------
 Femmes.CSP = Ouvrier

        Shapiro-Wilk normality test

data:  Femmes.Arret
W = 0.97245, p-value = 0.02314

 --------

 p-values adjusted by the Holm method:
        unadjusted adjusted
Cadre   0.00206379 0.00412757
Employé 0.00032428 0.00097284
Ouvrier 0.02313745 0.02313745
>
```

Pour les 2 sexe l'échantillons échoue au test de normalité. En effet la p-value très faible signifie que le H0 l'échantillons suis une loi normale peut être rejeté.

Le test de Bartlett échouera aussi ce qui permettra de rejeter l'hypothèse H0 ce qui reviendra à dire que les variances ne sont pas homogènes. Suite à cet échec nous avons essayé des tests non paramétriques afin de vérifier notre hypothèse première.

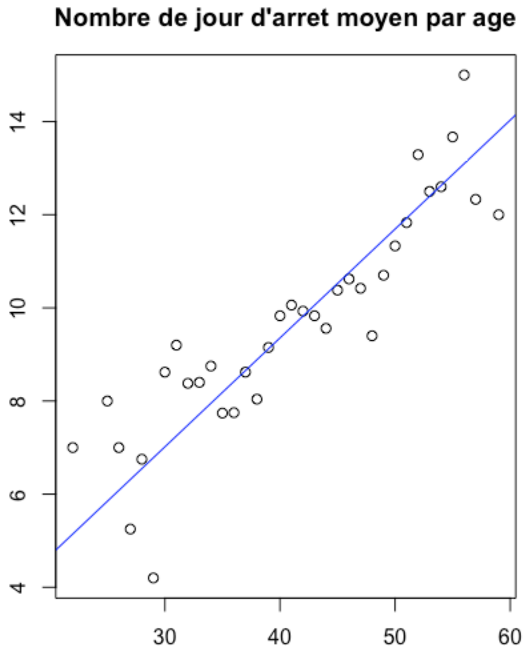
Nous effectuons donc un test de Kruskal Wallis afin de tester les moyennes de différents groupes de notre échantillon.
Ainsi nous posons :
H0: m1 = m2 = m3 homogénéité des moyennes
H1: m1=/= m2 et ou m1=/=m3 et ou m2=/= m3 Au moins une des moyennes n'est pas égal

```
        Kruskal-Wallis rank sum test

data:  Hommes.Arret by Hommes.CSP
Kruskal-Wallis chi-squared = 520.57, df = 2, p-value < 2.2e-16

>
```

Encore une fois nous obtenons une p-valeur inférieur à notre taux d'erreur accepté. Ce qui revient à dire que la condition H0 peur être rejeté et donc que le CSP a bien une incidence sur le nombre de jour d'arrêt.

En effet il est logique d'imaginé que des postes avec plus de responsabilité comme des cadres ont moins de jour d'arrêt en moyenne.

Question 5 - Etudier l'impact de l'âge sur les arrêts-maladie dans le groupe des femmes employées



A l’aide de ce nuage de point représentant le nombre de jour d’arrêt en fonction de l’âge des femmes employées, on constate que la régression linéaire obtenue montre une évolution linéaire du nombre de jour d’arrêt en fonction de l’âge.

L’âge a donc un impact significatif sur les arrêts maladies dans le groupe des femmes employées.

Question 6 - Donner une estimation par intervalle de confiance de niveau 95%, du nombre de jours d'absence d'une femme employée de 45 ans

Pour répondre à cette question nous avons effectué un test de student pour récupérer l’estimation des nombres de jour d’arrêt d’une femme de 45 ans

```
FemmesEmployees45yo <- Sujet_CCR[Sujet_CCR$Sexe == 'Femme' & Sujet_CCR$CSP == 'Employ\u00e9' & Sujet_CCR$Age==45,]

analyse <- t.test(table(FemmesEmployees45yo$Arret), conf.level=0.95 )
print(paste('intervalle de confiance a 95% : [',analyse$conf.int[1],', ' ,analyse$conf.int[2] ,']', sep=' '))
[1] "intervalle de confiance a 95% : [ 1.05864158641997 ,  3.6080250802467 ]"
```

```
(analyse$conf.int[2] - analyse$conf.int[1]) / 2
```

```
> (analyse$conf.int[2] - analyse$conf.int[1]) / 2
[1] 1.274692
```

A un seuil de 95%, l’estimation par intervalle de confiance du nombre de jours d’absence d’une femme employée est de 1.27 jours.