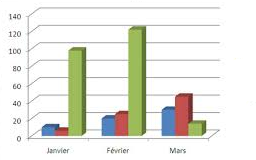
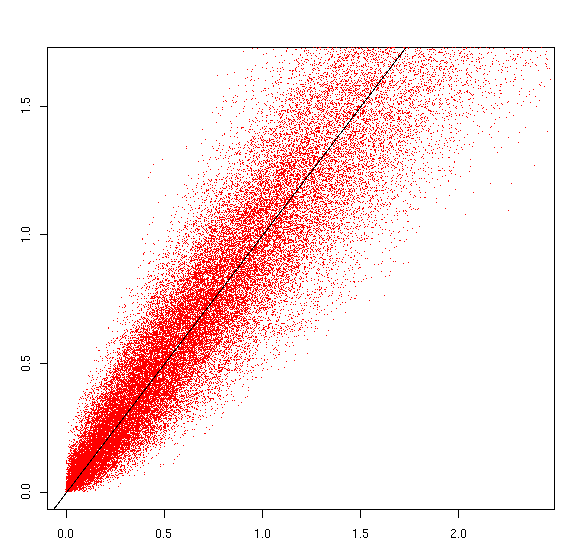
[Date]

**Rapport TP 1**





**Mickaël ALLIGON**

I-Présentation de la base de données

La base de données est constituée de 654 enfants, âgés de 3 à 19 ans. Les données ont été recensées à partir d’une étude américaine « The Childhood Respiratory Disease Study », qui s’est déroulée dans l’East Boston au Massachusetts en 1980. L’objectif principal est d’étudier la différence de capacités pulmonaires entre enfants fumeurs et enfants non-fumeurs.

Les différentes variables sont l’âge, la taille, le sexe, le statut de fumeur, et l’indice fev. L’âge ("age"), est exprimé en années, la taille ("height"), est exprimée en pouces. L’indice fev ("forced expiratory volume"), représente la capacité pulmonaire d’un individu en litres. Le sexe ("sex"), peut prendre deux modalités : "male" pour les hommes et "female" pour les femmes. Le statut de fumeur ("smoker"), peut prendre deux modalités : "Non" pour les non-fumeurs et "current" pour les fumeurs réguliers.

La problématique est la suivante : Le statut de fumeur a-t-il un impact sur la mesure de la capacité pulmonaire d’un individu ? La base de données étant principalement composée de mineurs, et donc de jeunes fumeurs et non-fumeurs, la problématique peut même être plus précise : la capacité pulmonaire est-elle rapidement impactée par le tabagisme ?

Afin de répondre à ces questions, une analyse univariée complète sur chacune des variables de l’étude sera effectuée. Puis dans un second temps, une analyse bivariée suivra avec des tests statistiques afin de comparer toutes les variables deux à deux.

Enfin une conclusion répondra aux problématiques de la base de données.

II-Analyses univariées

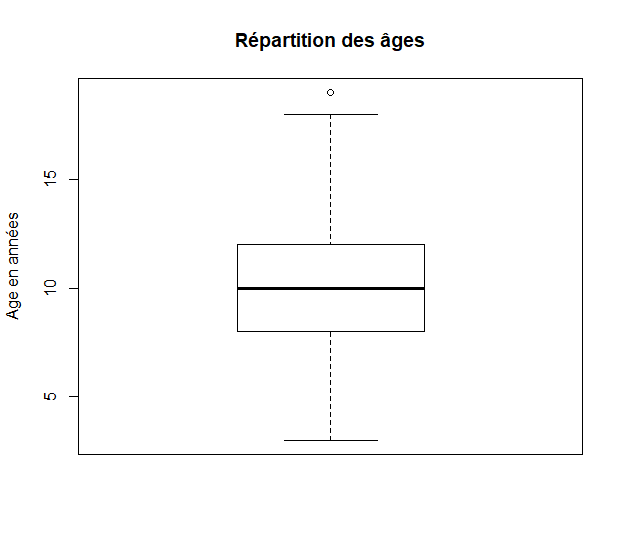
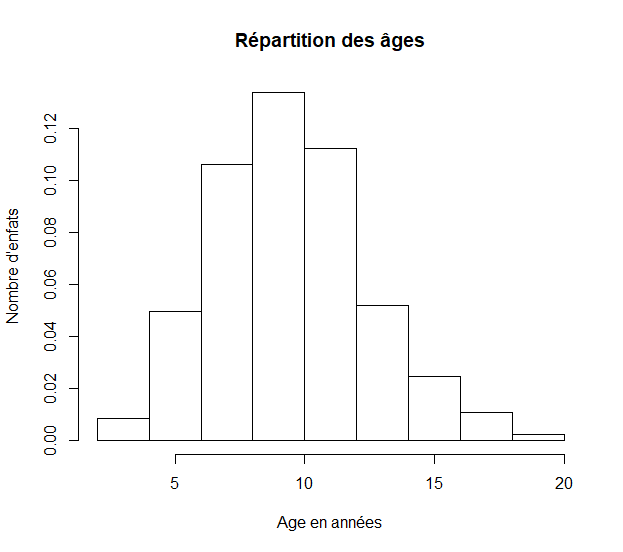
1. Étude de l’âge

L’âge est une variable quantitative discrète, mesurée en années, prenant, d’après la table (1), des valeurs allant de 3 à 19 ans dans la base de données. Sa moyenne est de 9,93 ans et son écart-type de 2,95 ans. La valeur médiane est 10 ans. Un quart des enfants ont moins de 8 ans, et un quart des enfants ont plus de 12 ans.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| 9,93 ans | 2,95 ans | 8 ans | 10 ans | 12 ans | 3 ans | 19 ans |

*Table(1) : distribution de l'âge des enfants*

On remarque grâce à la figure(1) que la distribution est symétrique avec une queue à droite. La variable âge est unimodale, la classe modale étant [8 ;10]. Il y a également une grande concentration entre 8 et 12 ans (la moitié des enfants) avec un étalement aux pôles. Les valeurs extrêmes ne sont pas aberrantes (3 personnes de 19 ans).



*Figure(1) : Représentations graphiques de la distribution de l'âge des enfants*

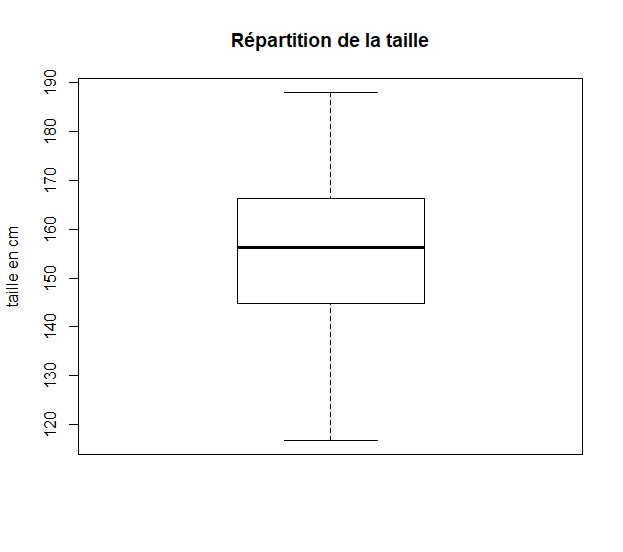
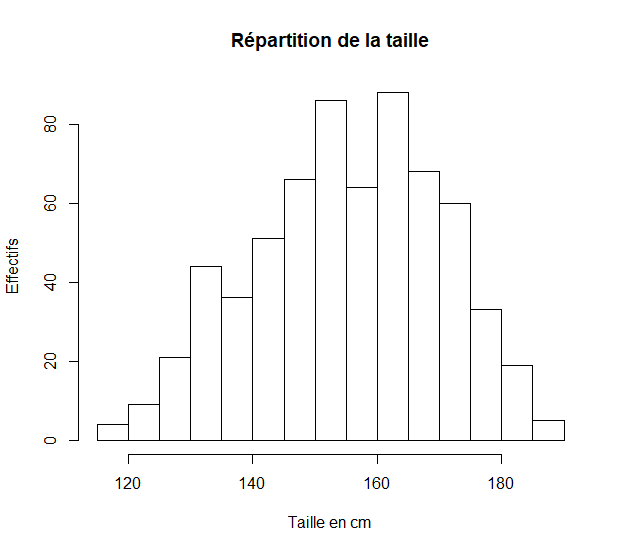
1. Étude de la taille

La taille est une variable quantitative discrète, mesurée en pouces. Comme la mesure standard en France est le centimètre, une conversion a été faite, en multipliant les tailles en pouces par 2,54. D’après la table(2), la taille prend des valeurs allant de 116,84cm à 187,96cm dans la base de données. Sa moyenne est de 115,3cm et son écart-type de 14,49cm. La valeur médiane est 156,21cm. Un quart des enfants mesurent moins de 144,78cm et un quart des enfants mesurent plus de 166,37cm.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| 115,3cm | 14,49cm | 144,78cm | 156,21cm | 166,37cm | 116,84cm | 187,96cm |

*Table(2) : distribution de la taille des enfants*

On remarque grâce à la figure(2) que la distribution est symétrique avec une queue à gauche. Il y a également une légère excroissance sur la classe [130 ; 135]. La variable taille est bimodale, les classes modales étant [150 ; 155] et [160 ; 165], dû à une grande diminution inattendue de l'effectif entre 155cm et 160cm. La concentration est plus grande entre la médiane et le 3ème quartile (respectivement 156cm et 166cm) qu'entre la médiane et le 1er quartile (respectivement 156cm et 144cm). L'écart-type est important, mais est attendu vu la différence d'âge des enfants.



*Figure(2) : Représentations graphiques de la distribution de la taille des enfants*

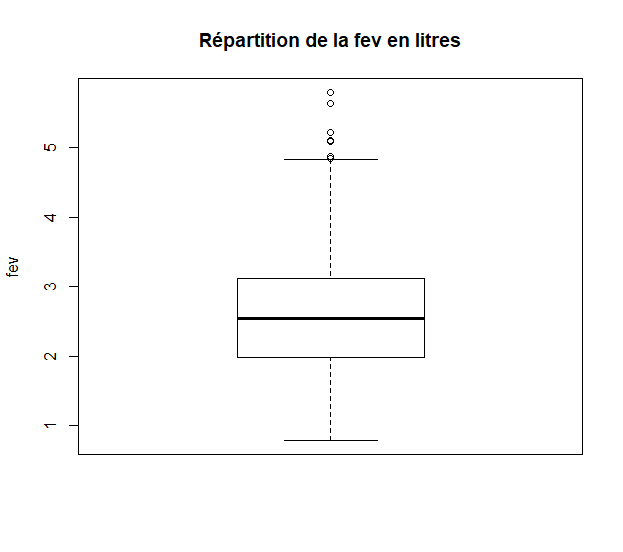
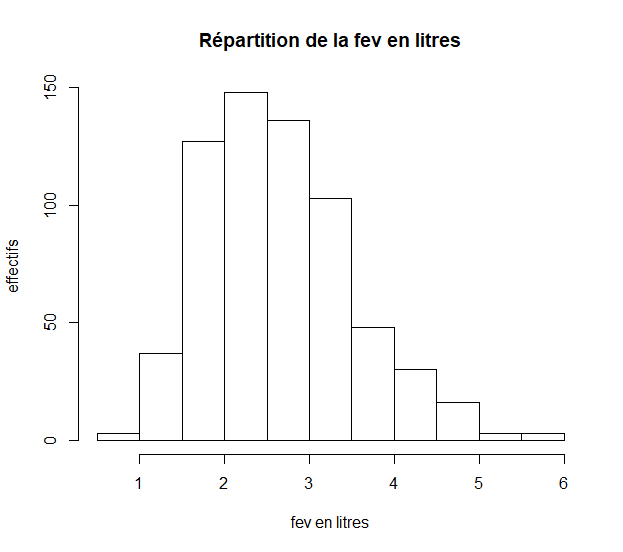
1. Étude de la fev

La fev est une variable quantitative discrète, mesurée en litres, prenant, d’après la table(3), des valeurs allant de 0,79L à 5,793L dans la base de données. Sa moyenne est de 2,64L et son écart-type de 0,87L. La valeur médiane est 2,55L. Un quart des enfants ont une mesure inférieure à 1,98L tandis qu’un quart en ont une supérieure à 3,12L.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| 2,64L | 0,87L | 1,98L | 2,55L | 3,12L | 0,79L | 5,793L |

*Table(3 ) : distribution de la fev*

On remarque, grâce à la figure(3), que la distribution est symétrique avec une queue à droite. La variable fev est unimodale, la classe modale étant [2 ;2,5]. La concentration est très grande entre le 1er et le 3ème quartile (respectivement 1,98L et 3,12L), avec un étalement vers le haut. Les valeurs extrêmes ne sont pas aberrantes (4,842L ; 5,224L ; 4,877L ; 5,0083L ; 5,102L ; 5,793L ; 5,633L ; 5,638L ; 4,872L).



*Figure(3) : Représentations graphiques de la distribution de la fev des enfants*

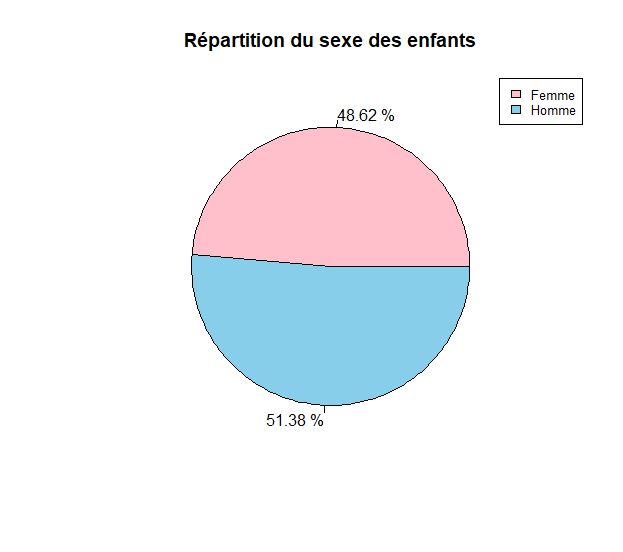
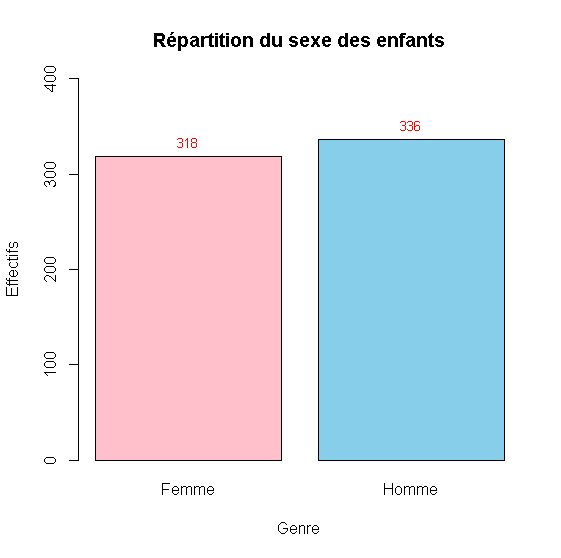
1. Étude du genre

Le sexe est une variable qualitative possédant deux modalités : "male" pour les hommes et "female" pour les femmes. D’après la table(4), il y a 336 hommes pour 318 femmes, soit respectivement 51,4 % et 48,6 % des individus de la base de données.

|  |  |  |
| --- | --- | --- |
|  | Effectifs | Pourcentages |
| Hommes | 336 | 51,4% |
| Femmes | 318 | 48,6% |

*Table(4) : Distribution du genre des enfants*

D’après la figure(4) (et également de la table(4)), on remarque qu'il y a légèrement plus de filles que de garçons, même si pour la taille des effectifs respectifs la différence peut être négligeable.

**

*Figure(4) : Représentations graphiques de la distribution du genre des enfants*

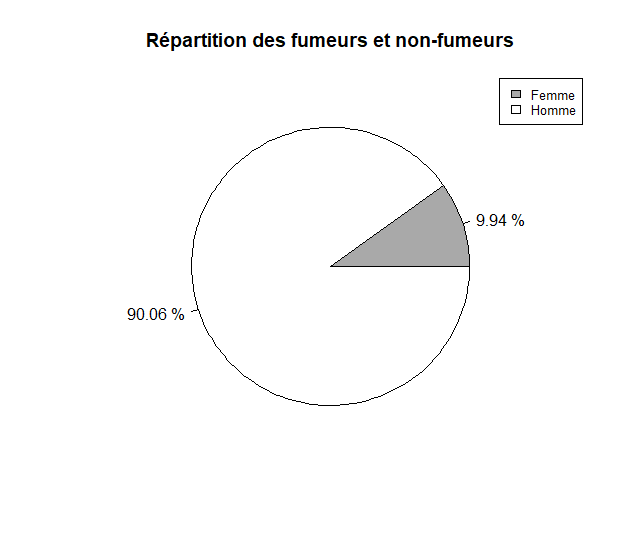
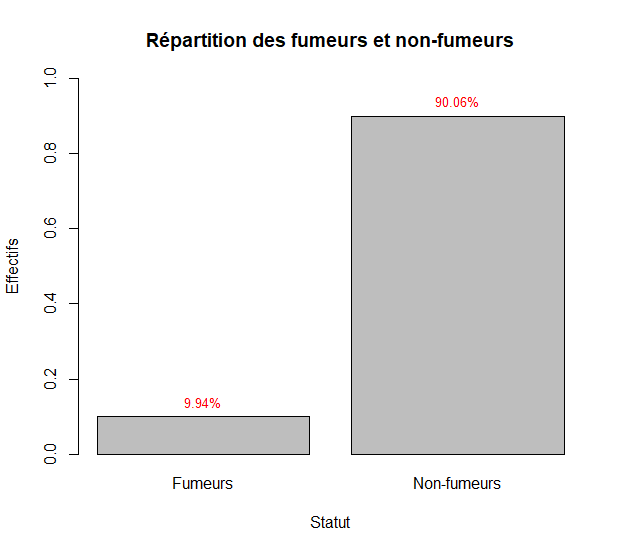
1. Étude du statut de fumeur

Le statut de fumeur est une variable qualitative possédant deux modalités, "non" pour les non-fumeurs et "current" pour les fumeurs. D’après la table(5), il y a 65 fumeurs pour 589 non-fumeurs, soit respectivement 10 % et 90 % des individus de la base de données.

|  |  |  |
| --- | --- | --- |
|  | Effectifs | Pourcentages |
| Fumeurs | 65 | 10% |
| Non-fumeurs | 589 | 90% |

*Table(5) : Distribution du statut de fumeur des enfants*

D’après la figure(5) (et également de la table(5)), on remarque qu'il y a beaucoup plus de non-fumeurs que de fumeurs, ce qui est attendu puisque l'échantillon est composé d'enfants de 3 à 19 ans. C'est pour cela qu'il pourra être nécessaire de découper l'échantillon en classes d'âge, certaines classes d'âge représentant mal la distribution du statut de fumeur (par exemple, les enfants de 3 ans ne fument certainement pas).



*Figure(5) : Représentations graphiques de la distribution du statut de fumeur des enfants*

III-Analyses bivariées et tests

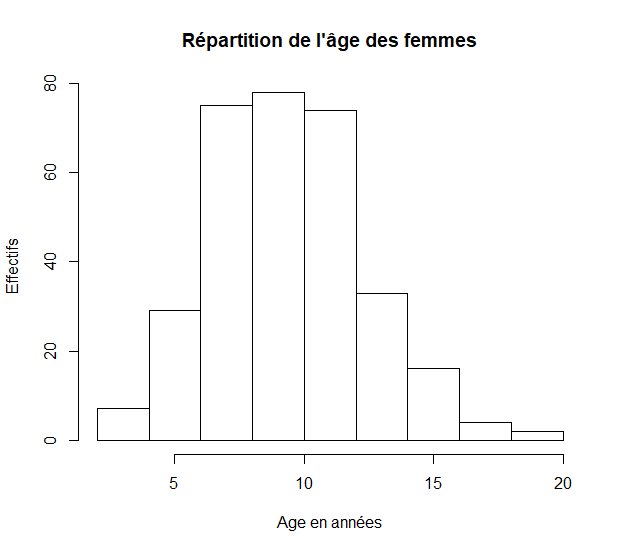
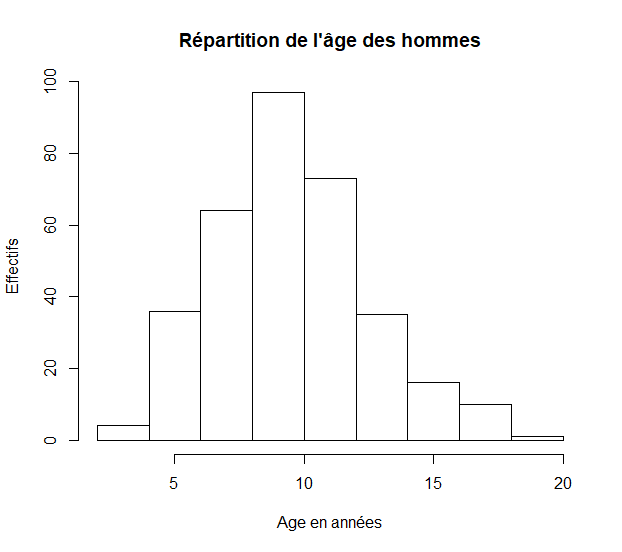
1. Études entre variables quantitatives et variables qualitatives, tests paramétriques
   1. Age et genre

D’après la table(6), on remarque que l’âge semble plutôt bien réparti entre les hommes et les femmes. En effet, tous les indicateurs sont égaux, ou extrêmement proches. La moyenne d’âge des hommes est de 10,01 ans, tandis qu’elle est de 9,84 ans pour les femmes. L’écart-type de l’âge des hommes est de 2,96 ans, contre 2,93 ans pour les femmes. Le 1er quartile (8 ans), la médiane (10 ans), le 3ème quartile (12ans), le minimum (3 ans) et le maximum (19 ans), ont tous la même valeur indépendamment du genre.

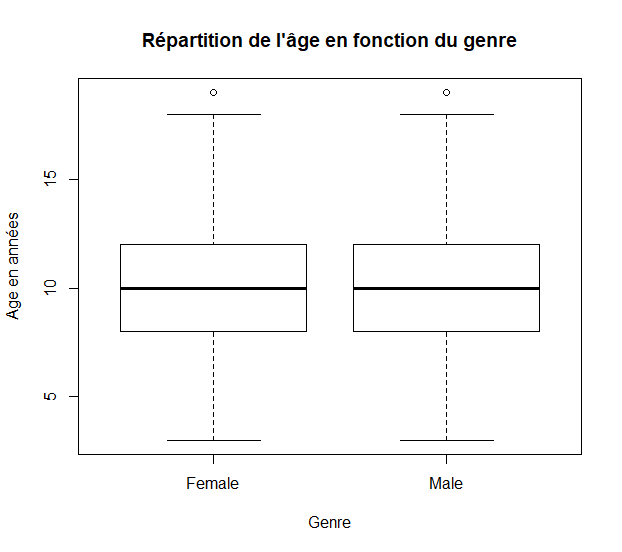
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| Hommes | 10,01 ans | 2,96 ans | 8 ans | 10 ans | 12 ans | 3 ans | 19 ans |
| Femmes | 9,84 ans | 2,93 ans | 8 ans | 10 ans | 12 ans | 3 ans | 19 ans |

*Table(6): Distribution de l’âge en fonction du genre*

D’après la figure(6), on remarque que la distribution de l’âge des femmes a en effet la même allure que celle des hommes : une distribution symétrique avec une queue à droite. Les deux répartitions sont unimodales, de classe modale [8;10].

*Fi*gure(6) : Histogrammes de la distribution de l’âge des enfants par genre

La figure(7) représente d’autant plus l’égalité des indicateurs statistiques : les boîtes à moustache ne diffèrent pas d’un genre à l’autre. Même les valeurs extrêmes, qui ne sont pas aberrantes (19 ans), sont identiques.



*Figure(7) : Boîtes à moustaches de la distribution de l’âge en fonction du genre*

Afin de vérifier si les légères différences ne sont pas significatives, on effectue un test bilatéral de Welch à deux échantillons au niveau de confiance 95 %, afin de comparer la moyenne d’âge de la population masculine, à celle de la population féminine.

* H0 : la moyenne d’âge des femmes est égale à celle des hommes.
* H1 : la moyenne d’âge des femmes est différente de celle des hommes.

On suppose H0 vraie, puis grâce au TLC (théorème limite central), on utilise le test de Welch.

Welch Two Sample t-test

data: Age by Sex

t = -0.74479, df = 650.94, p-value = 0.4567

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.6258831 0.2816558

sample estimates:

mean in group Female mean in group Male

9.842767 10.014881

*Test(1)  : Test de Welch à deux échantillons de l’âge par le sexe*

D’après Test(1), la p-valeur est de 46 %, ce qui est supérieure à 5 % (le risque d’erreur). Donc, au niveau de confiance 95 %, on ne peut montrer de différences significatives entre la moyenne d’âge des hommes et celles des femmes.

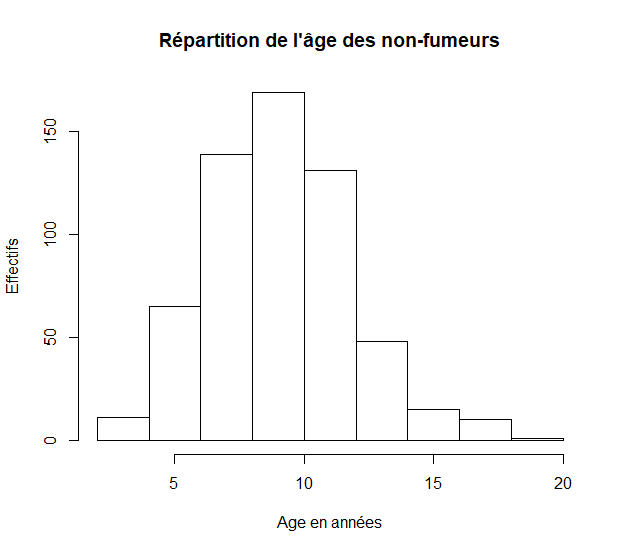
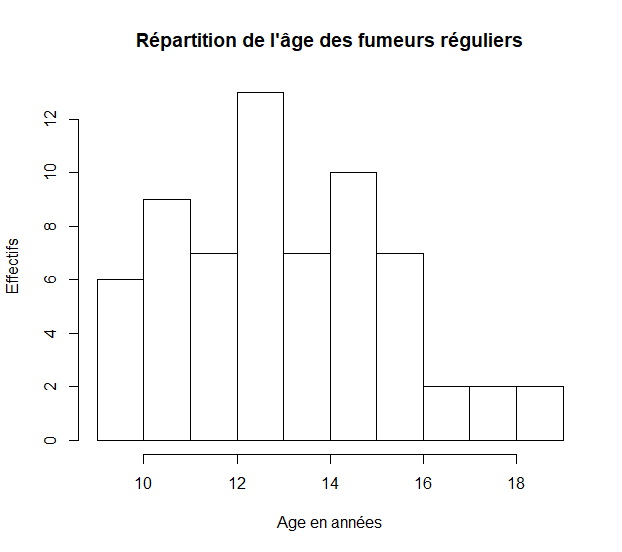
* 1. Age et statut de fumeur

D’après la table(7), on remarque que l’âge semble très mal réparti entre les fumeurs et les non-fumeurs. En effet, la quasi-totalité des indicateurs sont différents. La moyenne d’âge des fumeurs est de 13,52 ans, tandis qu’elle est de 9,53 ans pour les non-fumeurs. L’écart-type des fumeurs (2.34 ans) est légèrement inférieur à celui des non-fumeurs (2,74 ans). Un quart des non-fumeurs ont moins de 8 ans, pour 12 ans chez les fumeurs. De même, la médiane (13 ans) et le 3ème quartile (15 ans) des fumeurs sont plus élevés que ceux des non-fumeurs (respectivement 9 ans et 11 ans). Enfin le plus jeune fumeur a 9 ans, alors que le plus jeune enfant n’a que 3 ans. Finalement, la personne la plus âgée a 19 ans dans les deux catégories. Ces différences sont cependant normales et attendues, du fait que la base de données est composée d’enfants.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| Fumeur | 13,52 ans | 2,34 ans | 12 ans | 13 ans | 15 ans | 9 ans | 19 ans |
| Non-Fumeur | 9,53 ans | 2,74 ans | 8 ans | 9 ans | 11 ans | 3 ans | 19 ans |

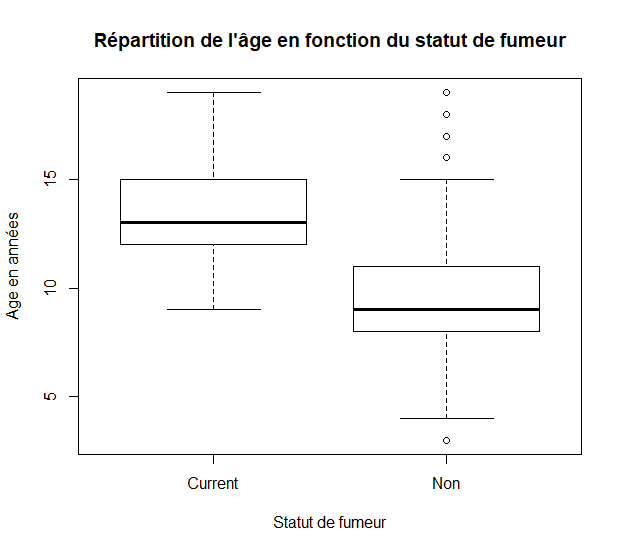
Table(7) : Distribution de l’âge en fonction du statut de fumeur

D’après la figure(8), on remarque que la distribution de l’âge des fumeurs a en effet une allure très différente de celle des non-fumeurs : une distribution symétrique avec une queue à droite pour les non-fumeurs et une distribution asymétrique avec une queue à droite pour les fumeurs. Les deux répartitions sont unimodales, de classe modale [8;10] pour les non-fumeurs, [12 ;13] pour les fumeurs. Il faut néanmoins notifier la faible précision de l’histogramme, dû au nombre de fumeurs peu élevé (moins de 100 personnes).



Figure(8) : Histogrammes de la distribution de l’âge des enfants par statut de fumeur

La figure(9) représente d’autant plus l’inégalité des indicateurs statistiques : si les concentrations autour des médianes semblent réparties de manière égale, il y a un décalage entre les boîtes à moustache très prononcé. En effet, on remarque visuellement que la médiane des non-fumeurs est égale à l’âge du plus jeune fumeur. On voit également que trois quarts des enfants ne fumant pas sont plus jeune que l’enfant représentant le 1er quartile des fumeurs. Les valeurs extrêmes ne sont pas aberrantes.



Figure(9) : Boîtes à moustaches de la distribution de l’âge en fonction du statut de fumeur

Afin de vérifier si les différences sont vr aiment significatives, on effectue un test unilatéral de Welch à deux échantillons au niveau de confiance 95 %, afin de vérifier si la moyenne d’âge des fumeurs est réellement plus élevée que celle des non-fumeurs.

* H0 : la moyenne d’âge des fumeurs est égale à celle des non-fumeurs.
* H1 : la moyenne d’âge des fumeurs est supérieure à celle des non-fumeurs.

On suppose H0 vraie, puis grâce au TLC (théorème limite central), on utilise le test de Welch.

Welch Two Sample t-test

data: Age by Smoker

t = 12.81, df = 84.646, p-value < 2.2e-16

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

3.47048 Inf

sample estimates:

mean in group Current mean in group Non

13.523077 9.534805

Test(2) : Test de Welch à deux échantillons de l’âge par le statut de fumeur

D’après Test(2), la p-valeur est inférieure à 2. 10-16 ce qui est inférieure à 5 % (le risque d’erreur). Donc, au niveau de confiance 95 %, et même ici 99%, l’âge moyen des fumeurs est significativement supérieur à celui des non-fumeurs.

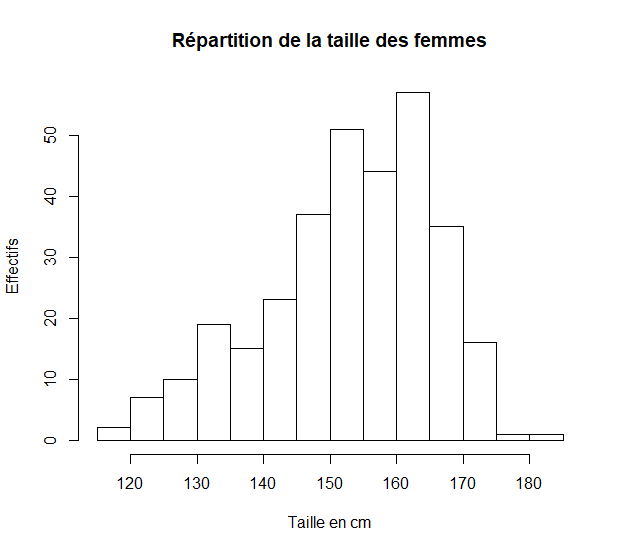
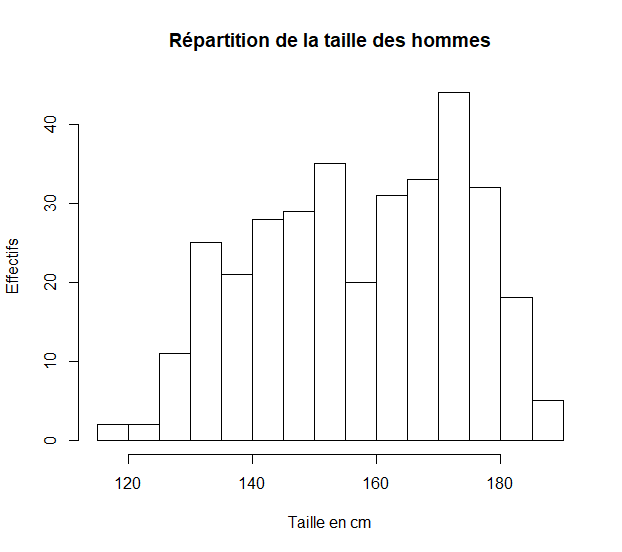
* 1. Taille et genre

D’après la table(8), on remarque que la répartition de la taille entre les hommes et les femmes semble différente. En effet, la totalité des indicateurs sont différents. La taille moyenne des hommes est de 157,5cm, tandis qu’elle est de 152,9cm pour les femmes. L’écart-type des hommes (16,08cm) est supérieur à celui des femmes (12,17cm). Un quart des hommes mesure moins de 144,8cm, pour 146,1cm chez les femmes. Cependant, la médiane (157,5cm) et le 3ème quartile (171,4cm) des hommes sont plus élevés que ceux des femmes (respectivement 154,9cm et 161,3cm). Enfin le plus petit garçon fait 116,8cm, 119,4cm pour la plus petite fille. Finalement, l’homme le plus grand mesure 188cm, tandis que la femme la plus grande atteint 180,3cm.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| Hommes | 157,5cm | 16,08cm | 144,8cm | 157,5cm | 171,4cm | 116,8cm | 188cm |
| Femmes | 152,9cm | 12,17cm | 146,1cm | 154,9cm | 161,3cm | 119,4cm | 180,3cm |

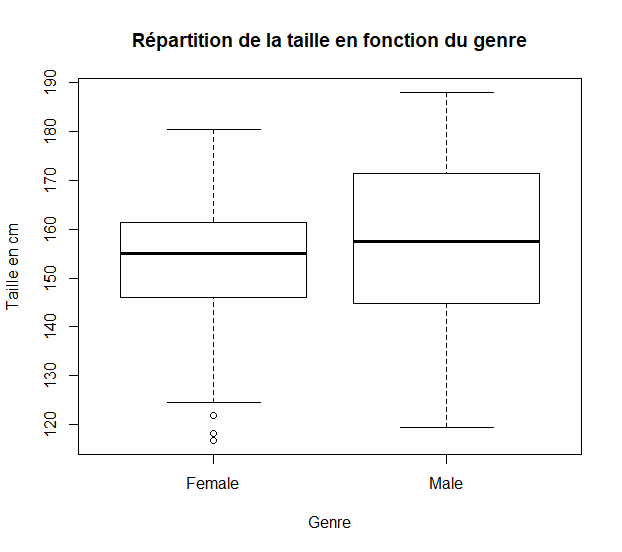
Table(8) : Distribution de la taille des enfants par sexe

D’après la figure(10), on remarque que la distribution de la taille des hommes a en effet une allure très différente de celle des femmes : une distribution symétrique avec une queue à gauche pour les femmes et une distribution asymétrique avec une queue à gauche pour les hommes. Les deux répartitions sont unimodales, de classe modale [170 ;175] pour les hommes, [160 ;165] pour les femmes.



Figure(10) : Histogrammes de la distribution de la taille des enfants par sexe

La figure(11) représente d’autant plus l’inégalité de certains indicateurs statistiques : en effet on remarque que les minimum, médianes et premiers quartiles sont relativement égaux. Cependant on remarque une plus grande différence entre les maximums et troisièmes quartiles. Enfin, la variance de la taille des hommes semble beaucoup plus grande que celle des femmes, car la box est beaucoup plus grande. Les valeurs extrêmes ne sont pas aberrantes.



Figure(11) : Boîte à moustache de la distribution de la taille des enfants en fonction du sexe

Afin de vérifier si les différences sont vraiment significatives, on effectue un test unilatéral de Welch à deux échantillons au niveau de confiance 95 %, afin de vérifier si la taille moyenne des hommes est significativement supérieure à celle des femmes.

* H0 : la taille moyenne des hommes est égale à celle des femmes.
* H1 : la taille moyenne des hommes est supérieure à celle des femmes.

On suppose H0 vraie, puis grâce au TLC (théorème limite central), on utilise le test de Welch.

Welch Two Sample t-test

data: taille by Sex

t = -4.1439, df = 622.34, p-value = 1.943e-05

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -2.774955

sample estimates:

mean in group Female mean in group Male

152.9384 157.5443

Test(3) : Test de Welch à deux échantillons de la taille par rapport au genre

D’après Test(3), la p-valeur est égale à 1,943.10-05 ce qui est inférieure à 5 % (le risque d’erreur). Donc, au niveau de confiance 95 %, et même ici 99%, la taille moyenne des hommes est significativement supérieure à celle des femmes.

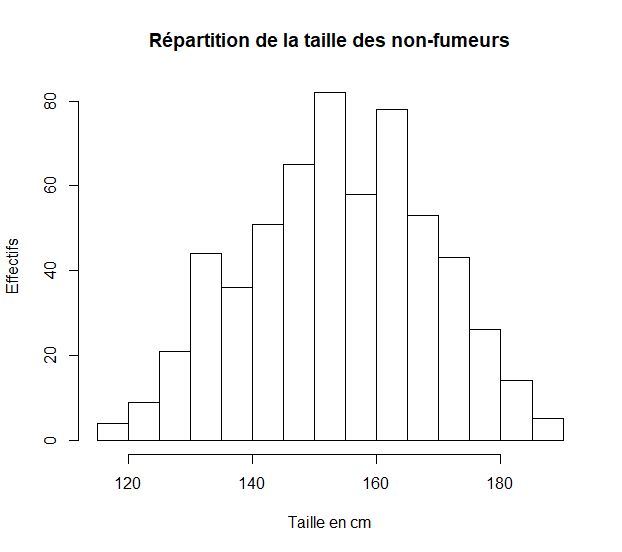
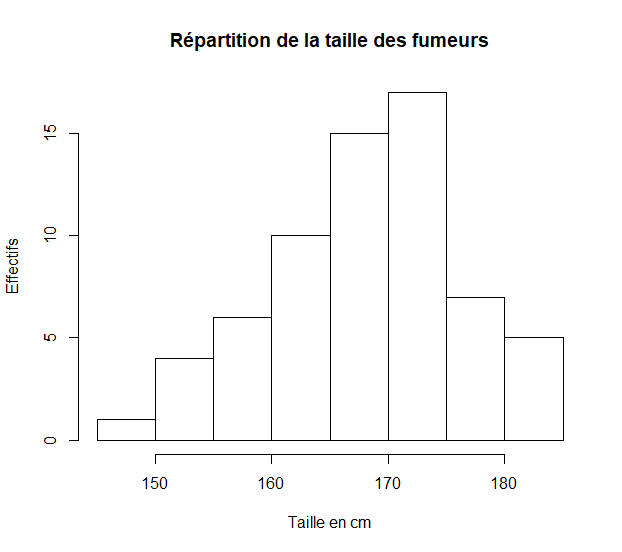
* 1. Taille et statut de fumeur

D’après la table(9), on remarque que la répartition de la taille entre les fumeurs et non-fumeurs semble différente. En effet, la totalité des indicateurs sont différents. La taille moyenne des fumeurs est de 167,5cm, tandis qu’elle est de 154cm pour les non-fumeurs. L’écart-type des fumeurs (8,11cm) est supérieur à celui des non-fumeurs (14,41cm). Un quart des fumeurs mesure moins de 161,3cm, pour 144,8cm chez les non-fumeurs. Cependant, la médiane (167,6cm) et le 3ème quartile (172,7cm) des fumeurs sont également plus élevés que ceux des non-fumeurs (respectivement 154,9cm et 163,8cm). Enfin le plus petit fumeur fait 147,3cm, 116,8cm pour le plus petit non-fumeur. Finalement, le fumeur le plus grand mesure 182,9cm, tandis que le non-fumeur le plus grande atteint 188cm.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moyenne | Écart-type | 1er quartile | Médiane | 3ème quartile | Minimum | Maximum |
| Fumeur | 167,5cm | 8,11cm | 161,3cm | 167,6cm | 172,7cm | 147,3cm | 182,9cm |
| Non-fumeur | 154cm | 14,41cm | 144,8cm | 154,9cm | 163,8cm | 116,8cm | 188cm |

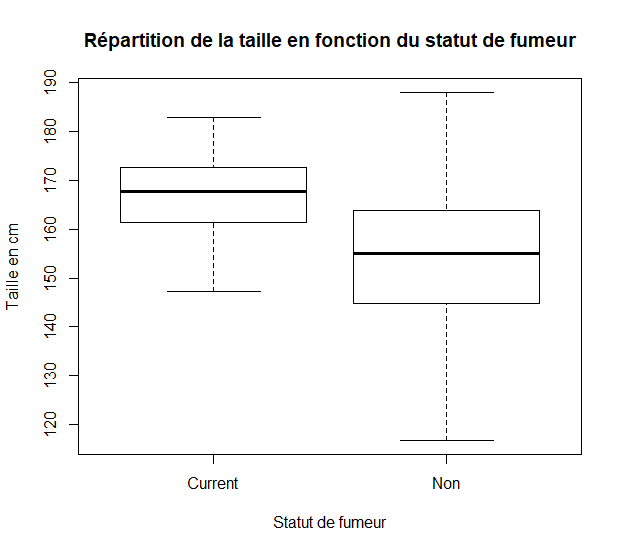
Table(9) : Distribution de la taille des enfants en fonction du statut de fumeur

D’après la figure(12), on remarque que la distribution de la taille des fumeurs a par contre une allure similaire à celle des non-fumeurs : une distribution symétrique avec une queue à gauche. Mais la répartition des fumeurs est unimodale, de classe modale [170 ;175], tandis que celle des non-fumeurs est bimodale, de classes modales [150 ;155] et [160 ;165]. Il faut néanmoins notifier la faible précision de l’histogramme, dû au nombre de fumeurs peu élevé (moins de 100 personnes).



Figure(12) : Histogrammes de la distribution de la taille des enfant en fonction du statut de fumeur

La figure(13) représente d’autant plus l’inégalité des indicateurs statistiques: en effet, tous les indicateurs des fumeurs sont plus grands que ceux des non-fumeurs. Par exemple, on remarque que le plus petit fumeur est plus grand qu’un quart des non-fumeurs. Ou encore que plus de 50% des fumeurs sont plus grand que trois quarts des non-fumeurs. On remarque bien que la variance de la taille des non-fumeurs est plus grande que celle des fumeurs (box plus large).



Figure(13) : Boîte à moustache de la distribution de la taille des enfants en fonction du statut de fumeur

Afin de vérifier si les différences sont vraiment significatives, on effectue un test unilatéral de Welch à deux échantillons au niveau de confiance 95 %, afin de vérifier si la taille moyenne des fumeurs est significativement supérieure à celle des non-fumeurs.

* H0 : la taille moyenne des fumeurs est égale à celle des non-fumeurs.
* H1 : la taille moyenne des fumeurs est supérieure à celle des non-fumeurs.

On suppose H0 vraie, puis grâce au TLC (théorème limite central), on utilise le test de Welch.

Welch Two Sample t-test

data: taille by Smoker

t = 11.615, df = 114.84, p-value < 2.2e-16

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

11.62965 Inf

sample estimates:

mean in group Current mean in group Non

167.5228 153.9563

Test(4) : Test de Welch à deux échantillons de la taille par rapport au statut de fumeur

D’après Test(4), la p-valeur est inférieure à 2,2.10-16 ce qui est inférieure à 5 % (le risque d’erreur). Donc, au niveau de confiance 95 %, et même ici 99%, la taille moyenne fumeurs est significativement supérieure à celle des non-fumeurs, ce qui était attendu vu la différence d’âge entre fumeurs et non-fumeurs montrée précédemment.