TP Bilan : Statistique Inférentielle

Exercice 1 : Maximum de vraisemblance

Nous observons des données qui représentent la durée de vie expérimentale d'un composant électronique observé sur un échantillon de mesure. Nous insérons alors ces 10 observations dans un data frame sur R. On suppose que cette durée de vie notée D suit une loi de Weibull tel que D → $W(\beta, k)$.

On cherche tout d'abord à estimer β avec k connu par la méthode du maximum de vraisemblance. La fonction de répartition de la loi de Weibull est : $f(x) = \frac{k\,x^{k-1}}{\beta^{k}} e^{-(x/\beta)^{k}}$

On calcule d'abord la vraisemblance de l'échantillon x1, …, xn

$$L(\beta) = f(x1) \,.\, f(x2) \,.\, \ldots \,.\, f(xn)$$
$$L(\beta) = k^{n} \,.\, \frac{1}{\beta^{nk}} [\, x1 \,.\, x2 \,.\, \ldots \,.\, xn \,]^{k-1} \,.\, e^{-(x1^{k} + \ldots + xn^{k})}$$

On passe au log tel que :

$$log(L(\beta)) = n\, log\,(k) - nk\, log(\beta) + (k-1)\, log(x1 + \ldots + xn) - (x1^{k} + \ldots + xn^{k}) \,.\, \frac{1}{\beta^{k}}$$

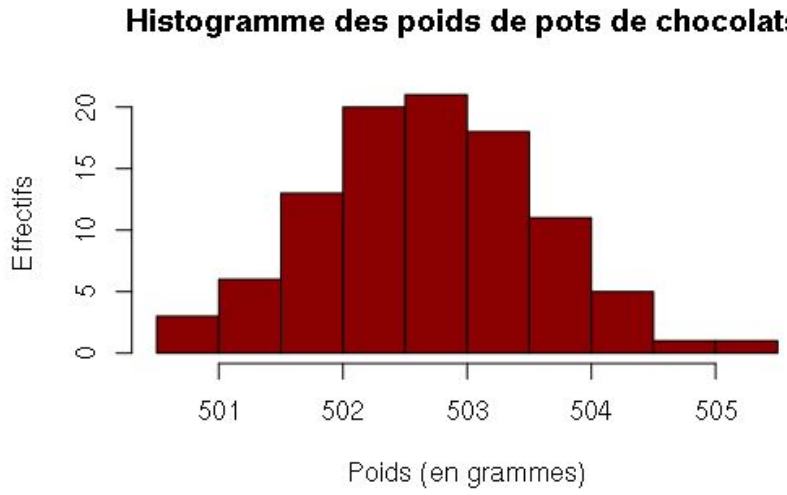
Pour chercher β qui maximise la vraisemblance on résout l'équation $l'(\beta) = 0$.

$$l'(\beta) = 0 - n^{k} / \beta + 0 - (x1^{k} + \ldots + xn^{k}) \,.\, -k/\beta^{k+1}$$
$$\Leftrightarrow \quad \frac{(x_1^{k} + \ldots + x_n^{k})}{\beta^{k+1}} = \frac{-n}{\beta}$$
$$\Leftrightarrow \quad (x_1^{k} + \ldots + x_n^{k}) = n.\beta^{k}$$
$$\Leftrightarrow \quad \beta^{k} = \frac{(x_1^{k} + \ldots + x_n^{k})}{n} \quad \Leftrightarrow \quad \widehat{\beta} = \left(\frac{(x_1^{k} + \ldots + x_n^{k})}{n}\right)^{1/k}$$

On trouve alors l'estimateur $\widehat{\beta}$ = 46.89 grâce à cette estimation. Et $\widehat{k}$ = 2.25.
Ces résultats sont validés car la fonction optim de R renvoie convergence, et ici elle est de 0.

Exercice 2 : Contrôle de fabrication sur mesure

Nous observons maintenant des données représentant les poids en gramme de 99 pots de chocolats sortis d'une machine à conditionner.
Nous observons ainsi ces données et nous produisons un histogramme.



Grâce à cet histogramme, on remarque que les effectifs des observations peuvent ressembler fortement à une loi normale.

De plus, on réalise *un test de Shapiro* pour prouver la normalité avec un risque de $\alpha = 5\%$. On pose ainsi l'hypothèse H0 : "Normalité" et H1 : "Pas de normalité". On trouve une p-valeur de 0.0629, ainsi nous avons donc 6,3% de probabilité de se tromper en rejetant H0, alors on accepte l'hypothèse nulle. Ainsi, on accepte la normalité, alors les poids en gramme de 99 pots de chocolats suivent une loi normale, soit $X \rightarrow N(m, \sigma^2)$.

On cherche maintenant un intervalle de confiance de niveau 95% pour le poids moyen des boîtes, et la variance du poids des boîtes. Tout d'abord, nous estimons la moyenne empirique et la variance empirique de l'échantillon donné.

Nous trouvons alors :

Moyenne empirique	502.9343
Variance empirique non biaisée	0.84

Intervalle de confiance sur le poids moyen des boîtes

Pour construire un intervalle de confiance sur le poids moyen des boîtes dans le cas où σ^2 est inconnu. Il faut, alors, se positionner dans le cas d'un modèle gaussien (ce que nous justifions au dessus). De plus, on remplace σ^2 par un estimateur S^2. Et on considère la statistique de test suivante, qui suit une loi de *Student* à n-1 degrés de liberté :

$$\frac{\overline{Xn} - m}{\sqrt{S^2/n}} \rightarrow T_{n-1}$$

Donc Intervalle de confiance est donc :

$$IC\ (m, 0.95) = [\ \overline{Xn} - t_{n-1\ ;\ 1-\alpha/2} \sqrt{\frac{S^2}{n}}\ ;\ \overline{Xn} + t_{n-1\ ;\ 1-\alpha/2} \sqrt{\frac{S^2}{n}}\ ]$$

On rappelle que cet intervalle de confiance est un IC de niveau asymptotique 1- α (soit de risque α) pour m.

Nous obtenons donc : **IC** (**m, 0.95**) = [**502.75 ; 503.12**]

Avec cet intervalle de confiance, nous montrons que nous avons 95% de probabilité d'avoir des pots de chocolat qui sont entre 502.75 et 503.12 grammes.

Intervalle de confiance sur la variance du poids des boîtes

Dans ce cas d'un modèle gaussien, on cherche l'intervalle de confiance pour la variance σ^2. On remarque que cette statistique de test suit une loi du *Chi-deux* à n-1 degrés de liberté. On obtient la statistique de test suivante :

$$\frac{n V^2}{\sigma^2} = \frac{(n-1) S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

Donc l'intervalle de confiance de niveau 1- α pour σ^2 est le suivant :

$$IC\ (\sigma^2\ , 0.95\ ) = [\ \frac{(n-1) S^2}{u_{n-1\ ;\ 1-\alpha/2}}\ ;\ \frac{(n-1) S^2}{u_{n-1\ ;\ \alpha/2}}\ ]$$

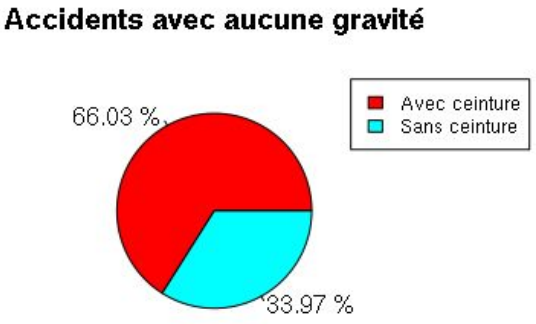
Nous obtenons donc les résultats suivants : **IC**(σ^2 **, 0.95**) = [**0.65 ; 1.14**]

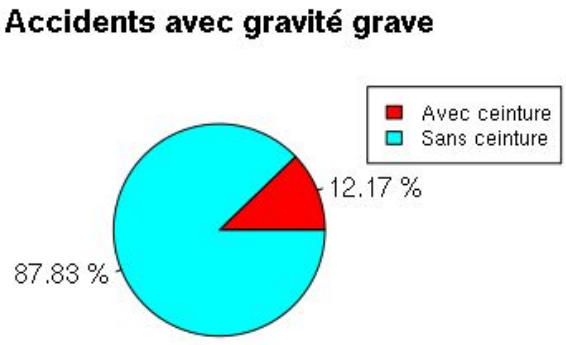
En interprétant cet intervalle de confiance, on peut alors dire que l'on a 95% de probabilité d'avoir une variance entre 0.65 et 1.14.

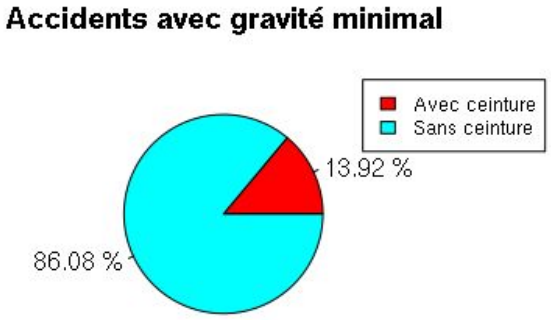
Exercice 3 : Gravité des blessures d'accidents de la route

On observe une étude sur la gravité des accidents de voiture en fonction du port de la ceinture de sécurité (Oui vs Non).

Premièrement, nous observons les différents diagrammes circulaires qui représentent les 4 niveaux de gravité de la proportion :







Nous nous demandons si le port de ceinture de sécurité influence t-il la gravité des blessure. On veut alors tester si il y a indépendance ou non des observations entre elles, afin de voir si porter ou non sa ceinture apporte des accidents. Nous réalisons un *test d'indépendance du chi-deux.*

On pose alors les hypothèses suivantes avec un risque de $\alpha = 5\,\%$:

H0 : " Indépendance des observations " c'est à dire "Ceinture et Gravité sont indépendants"

H1 : " Dépendance des observations " c'est à dire "Ceinture et gravité sont corrélés"

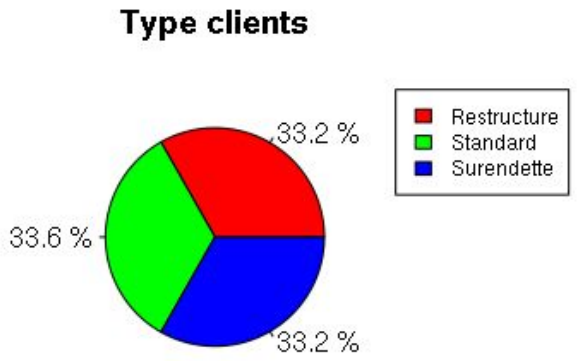
Après réalisation du test on obtient une p-valeur de $2,2 * 10^{-16}$. On n'accepte pas l'hypothèse H0 puisque la p-valeur est inférieur à $2,2 * 10^{-16}$. On conclut donc qu'il y a corrélation entre les observations. Ainsi, le port de la ceinture de sécurité influence la gravité des accidents.

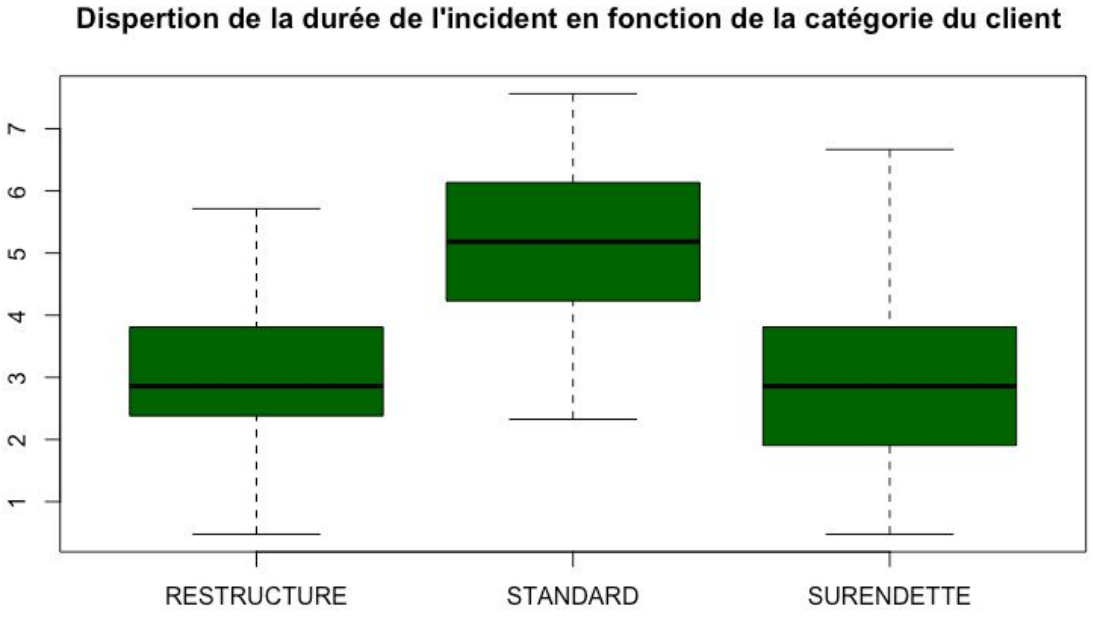
Exercice 4 : Suivi de sinistre des clients d'une banque

On étudie le fichier incident.csv qui contient des informations concernant des clients fragiles d'une institution financière.
On cherche à estimer la durée moyenne en mois entre le moment où un client a été jugé fragile et la survenance d'un premier incident de paiement. On suppose ici que ces durées sont distribuées normalement, ainsi on a alors X → $N(m, \sigma^2)$. On observe les différents indicateurs sur la variable durée tel que :

Min	1er quartile	Median	Mean	3er quartile	Max
0.47	2.38	3.75	3.71	4.76	7.56

De même, on remarque que les différents types de clients, les proportions de ceux ci sont bien répartis, ainsi, il y a 166 clients restructurés, 168 standards, 166, surendettés.





Avec cette boîte à moustache on remarque les 3 types de clients en fonction de la durée de l'incident, en effet on remarque que les clients standards ont une durée d'incident en moyenne plus élevée que les clients restructurés et surendettés.

On cherche à estimer au risque α=5% que la durée moyenne avant la survenance du premier accident de paiement est supérieure à 2 mois, et qui est inférieure à 4 mois. Nous utilisons pour cela le test de Student, car nous savons que la moyenne est inconnue et la variance également. On se rapporte alors à un test de *Student*.
Ainsi nous trouvons :

	Hypothèse H0	Hypothèse H1	p-valeur	Validation ou non
Supérieur à 2 mois	"m = 2"	"m > 2"	$2.2*10^{-16}$	Rejet H0
Inférieur à 4 mois	"m = 4"	"m < 4"	$3.648*10^{-5}$	Rejet H0

Grâce à ces tests on remarque que la moyenne avant le premier incident est supérieure à 2 mois et inférieure à 4 mois avec une confiance de 5%. On cherche un encadrement de la durée moyenne avant le premier incident de paiement plus précis, c'est à dire un intervalle de confiance de risque $\alpha = 5\%$.
L'intervalle est alors :

IC (m,0.95) = [3.57 ; 3.85]

Ainsi nous avons 95% de probabilité de trouver la moyenne entre 3.57 et 3.85 mois.

On veut un intervalle de confiance de niveau 99% de la moyenne et de l'écart type de la durée avant la survenance du premier incident de paiement pour chaque type de client : Pour la moyenne c'est un *test de Student*, et pour l'écart type, *un test du chi-deux.* On suppose être dans un modèle gaussien.

Ainsi pour les moyenne nous avons, à un risque de 1% :

Clients Restructurés	IC (m,0.99) = [2.82 ; 3.26]
Clients Standards	IC (m,0.99) = [4.88 ; 5.37]
Client Surendettés	IC (m,0.99) = [2.68 ; 3.24]

Par exemple, on peut alors dire, qu'avec une probabilité de 99%, les clients standards ont une tendance à avoir une durée en moyenne plus grande que les clients restructurés et surendettés.

De plus, pour l'écart type nous avons, à un risque de 1% :

Clients Restructurés	IC(σ , 0.99) = [0.97 ; 1.29]
Clients Standards	IC(σ , 0.99) = [1.07 ; 1.42]
Client Surendettés	IC(σ , 0.99) = [1.22 ; 1.62]

Test d'égalité des variances :

On étudie maintenant s'il y a une *différence entre la variance des durées* avant la survenance du premier incident de paiement. On effectue alors un test d'égalité des variances. On teste alors l'hypothèse nulle où il y a égalité des variances, contre l'hypothèse 1 où il n'y a pas égalité des variances. Ainsi on veut tester : H0 : " $\sigma_1^2 = \sigma_2^2$ " contre H1 : "" $\sigma_1^2 \neq \sigma_2^2$ " à un risque $\alpha = 5\%$.

La statistique de test est alors :

$$\frac{S^2_{1;n1}}{S^2_{2;n2}} \rightarrow F_{n1\,;\,n2}$$

On compare ainsi les p-valeurs de ces tests.

Echantillon 1	Echantillon 2	P-valeur	Acceptation ou non des hypothèses
Les clients surendettés	Les clients restructurés	0.0034	Rejet H0
Les clients surendettés	Les clients standards	0.099	Acceptation H0
Les clients restructurés	Les clients standards	0.1938	Acceptation H0

Test d'égalité des moyennes :

Et nous cherchons s'il y a une *différence entre la moyenne des durées* avant la survenance du premier incident de paiement. Nous effectuons donc un test sur l'égalité des moyennes. Cependant, ce test indique que les variances doivent être égales pour tester l'égalité des moyennes. On va alors supposer que les variances sont égales pour faire ce test. C'est alors pour cela, que nous avons d'abord testé l'égalité des variances. On teste alors : H0 : " $m_1 = m_2$ " contre H1 : " $m_1 \neq m_2$ ".

La statistique de ce test est :

$$\frac{\overline{X}_{1\,;\,n1} - \overline{X}_{2\,;\,n2}}{\sqrt{S^2\left(\frac{1}{n1}+\frac{1}{n2}\right)}} \rightarrow T_{n1+n2-2}$$

où $S^2 = \frac{n1\, V^{2}_{1,n1} + n2\, V^{2}_{2,n2}}{n1 + n2 - 2}$ qui est un estimateur “global” de la variance commune.

Echantillon 1	Echantillon 2	P-valeur	Acceptation ou non des hypothèses
Les clients surendettés	Les clients restructurés	0.561	Acceptation H0
Les clients surendettés	Les clients standards	$2.2*10^{-16}$	Rejet H0
Les clients restructurés	Les clients standards	$2.2*10^{-16}$	Rejet H0

Exercice 5 : Données accouchement prématurés

Le fichier prematures.csv possède des observations des facteurs prénataux liés à l'accouchement prématuré chez les femmes déjà en travail prématuré. Cette étude porte sur 220 femmes. On observe un bon nombre de variables dans le fichier.

GEST	l'âge gestationnel en semaines à l'entrée dans l'étude
DILATE	la dilatation du col en cm
CONSIS	la consistance du col (1=mou, 2=moyen, 3=ferme)
EFFACE	l'effacement du col (en pourcentage)
CONTR	la présence (=1) ou non (=2) de contraction
MEMBRAN	les membranes ruptures (=1) ou non (=2) ou incertain (=3) AGE : âge de la patiente
AGE	Âge de la patiente
GRAVID	la gestite (nombre de grossesses antérieures y compris celle en cours.
DIAB	la présence (=1) ou non (=2) d'un problème de diabète, ou valeur manquante (=3).
TRANSF	transfert (1) ou non (2) dans un hôpital de soins spécialises
GEMEL	grossesse simple (=1) ou multiple (2)
PREMATURE	accouchement prématuré (positif ou négatif)

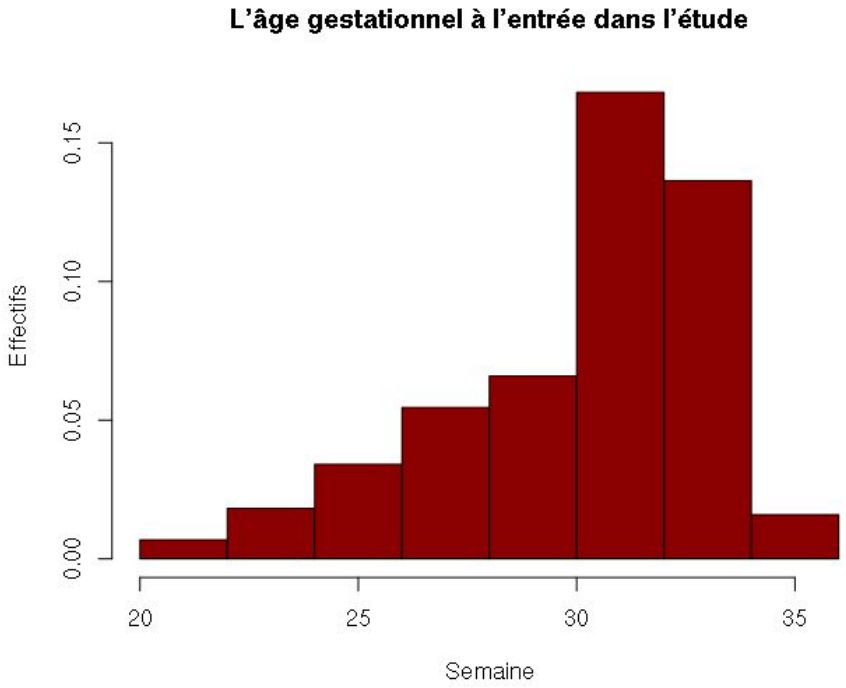
```
       ID              GEST            DILATE
 Min.   :  1.00   Min.   :20.00   Min.   :0.000
 1st Qu.: 55.75   1st Qu.:29.00   1st Qu.:0.000
 Median :110.50   Median :32.00   Median :1.000
 Mean   :110.50   Mean   :30.71   Mean   :1.086
 3rd Qu.:165.25   3rd Qu.:33.00   3rd Qu.:2.000
 Max.   :220.00   Max.   :35.00   Max.   :8.000
     EFFACE           CONSIS          CONTR
 Min.   :  0.00   Min.   :1.000   Min.   :1.000
 1st Qu.:  0.00   1st Qu.:2.000   1st Qu.:1.000
 Median : 50.00   Median :3.000   Median :1.000
 Mean   : 39.45   Mean   :2.382   Mean   :1.091
 3rd Qu.: 52.50   3rd Qu.:3.000   3rd Qu.:1.000
 Max.   :100.00   Max.   :3.000   Max.   :3.000
    MEMBRAN           AGE            GRAVID
 Min.   :1.000   Min.   :17.00   Min.   :0.000
 1st Qu.:2.000   1st Qu.:23.00   1st Qu.:1.000
 Median :2.000   Median :26.00   Median :2.000
 Mean   :1.845   Mean   :26.47   Mean   :2.145
 3rd Qu.:2.000   3rd Qu.:30.00   3rd Qu.:3.000
 Max.   :3.000   Max.   :40.00   Max.   :7.000
     PARIT             DIAB          TRANSF
 Min.   :0.0000   Min.   :1.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000
 Median :1.0000   Median :2.000   Median :2.000
 Mean   :0.7136   Mean   :2.005   Mean   :1.532
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :6.0000   Max.   :9.000   Max.   :2.000
     GEMEL         PREMATURE
 Min.   :1.000   negatif:110
 1st Qu.:1.000   positif:110
 Median :1.000
 Mean   :1.095
 3rd Qu.:1.000
 Max.   :2.000
>
```

On réalise une analyse statistique sur les différentes variables. On peut retrouver les informations des différentes variables ci-contre. Cependant, de nombreuses variables sont qualitatives, et donc ces informations ne sont alors pas très importantes. Ainsi, les variables ID, GEMEL, CONSIS, CONTR, DIAB, TRANSF, PREMATURE, et MEMBRAN sont des variables qualitatives. Ainsi les informations de ces variables ne doivent pas être prises en compte.

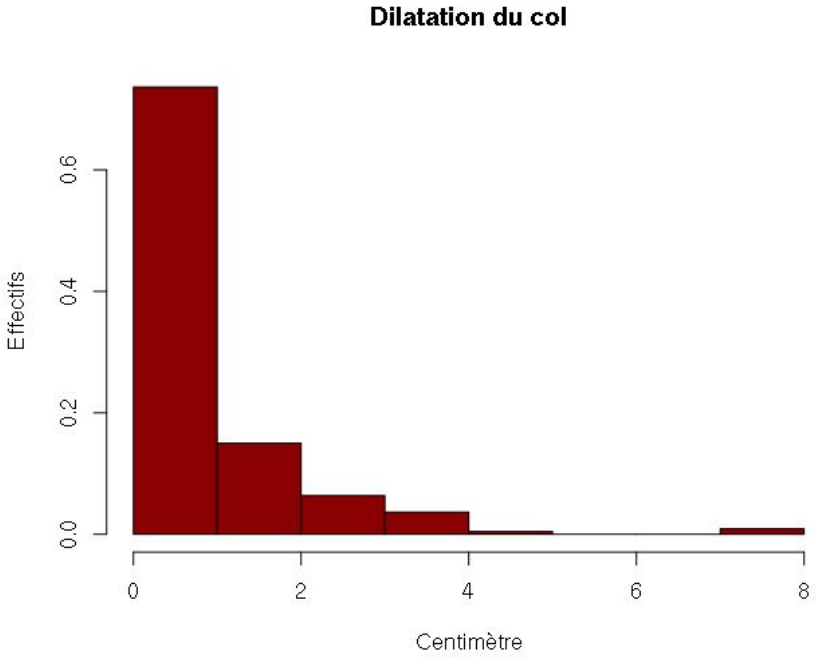
Néanmoins, on remarque l'âge moyen des femmes ayant un enfant prématuré ou non de l'étude qui est de 26.47 ans.

On retrouve ci dessus les analyses descriptives des différentes variables sur l'ensemble de l'échantillon. On

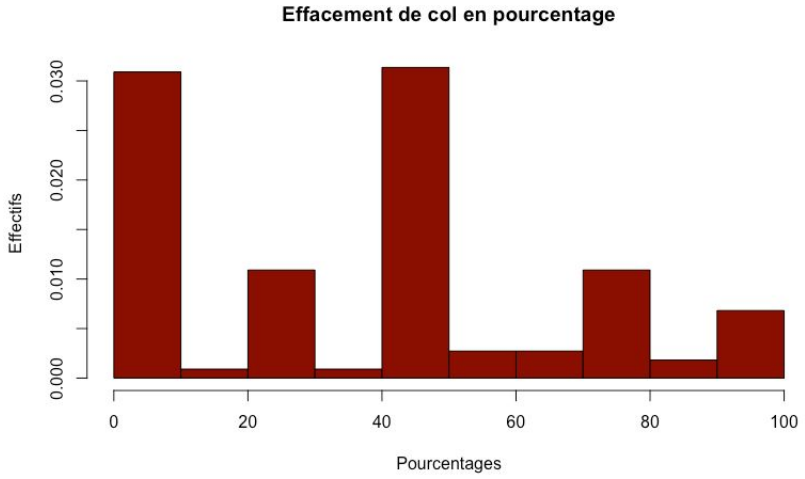
rappelle que les pourcentages ne font peut-être pas 100% car R a fait des arrondis.
GEST : On peut remarquer ici que la moyenne de gestation en semaine est de 32 semaines pour les femmes entrant dans l'étude.



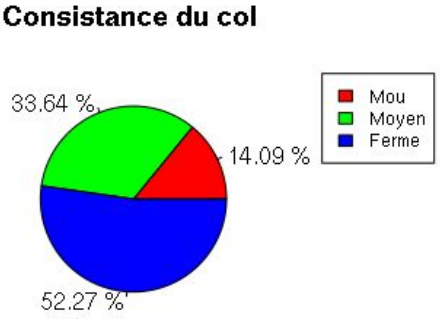
DILATE : La dilatation du col en cm est comprise entre 0 et 2,5 cm. Cependant on retrouve une valeur atypique qui est de 7 cm.



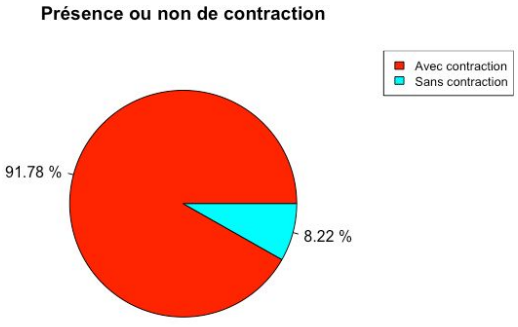
EFFACE :



CONSIS : 52,27% des femmes ont un col fermé, et 14,09% un col mou.

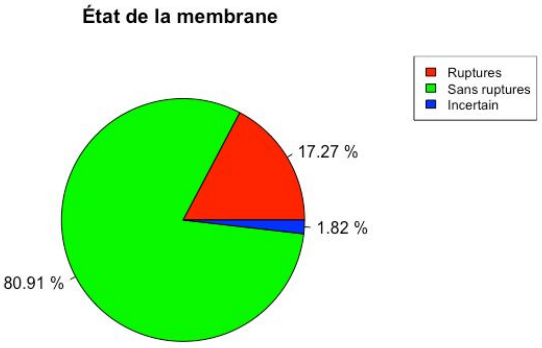


CONTR : 91,78% des femmes ont eu des contractions, contre 8,22% qui n'ont pas eu de contractions.

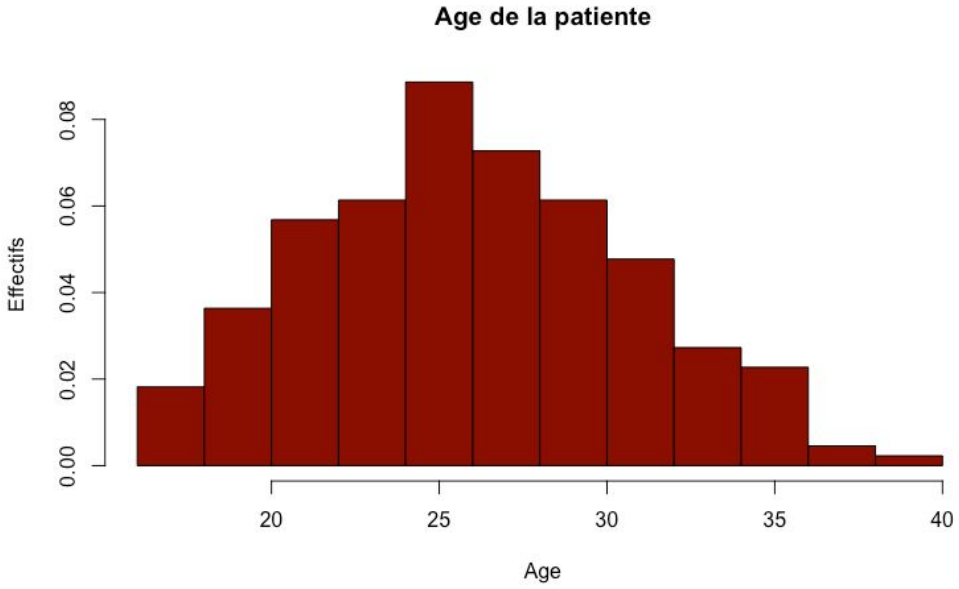


On tient à rappeler que nous avons enlevé les indicateurs 3 de cette variable qualitative, car nous avons trouvé une ligne possédant un 3 en réponse à cette variable qui ne devrait avoir que des 1 et des 2.

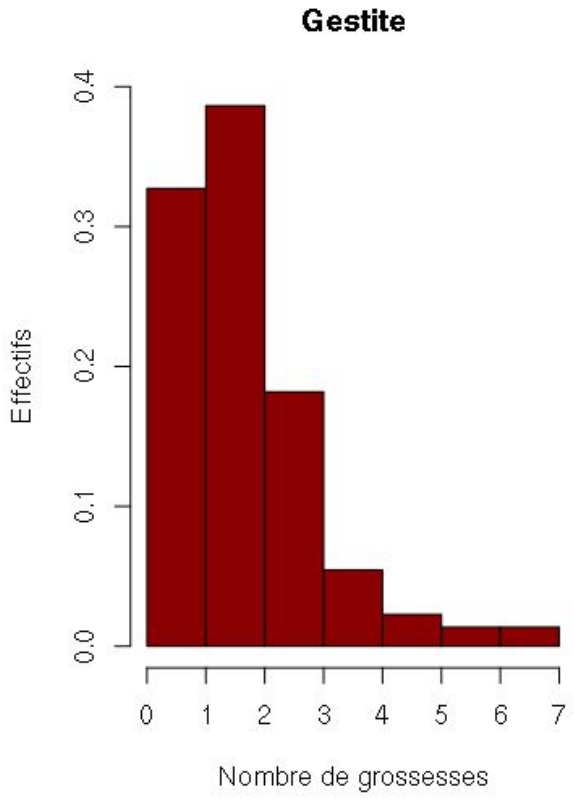
MEMBRAN : 17,27% des femmes n'ont pas eu de rupture de la membrane.



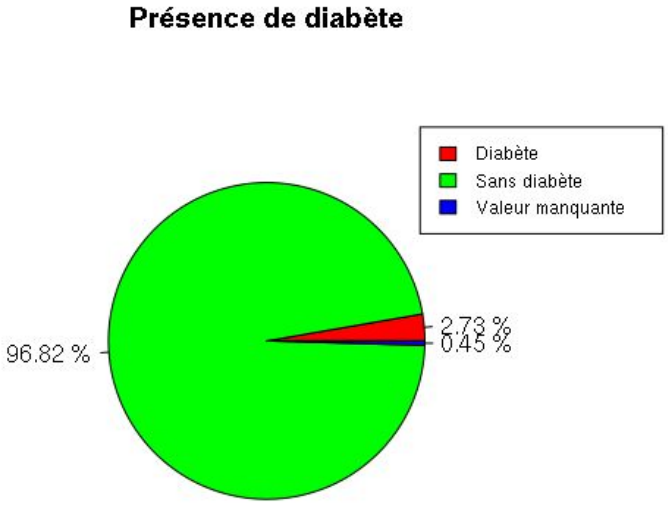
AGE : L'âge moyen de la patiente est 26 ans.



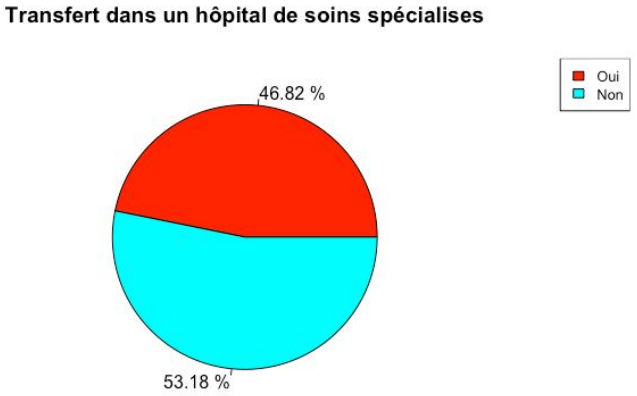
GRAVID : La plupart des femmes ont eu entre 0 et 3 grossesses antérieures.



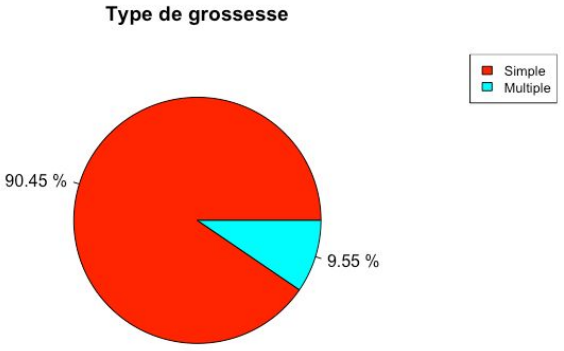
DIAB : Presque 97% des femmes n'ont pas de problème de diabète, et seulement 3% en ont eu.



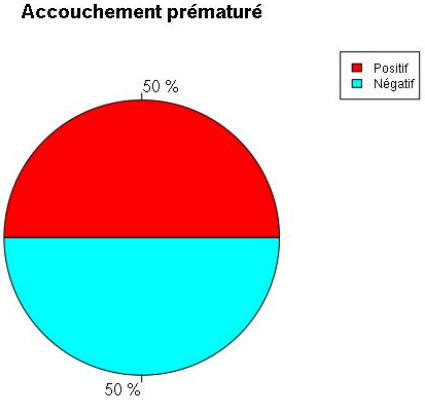
TRANSF : 53% des femmes n'ont pas eu de transfert dans un hôpital de soins spécialisés alors que 47% ont eu un transfert.



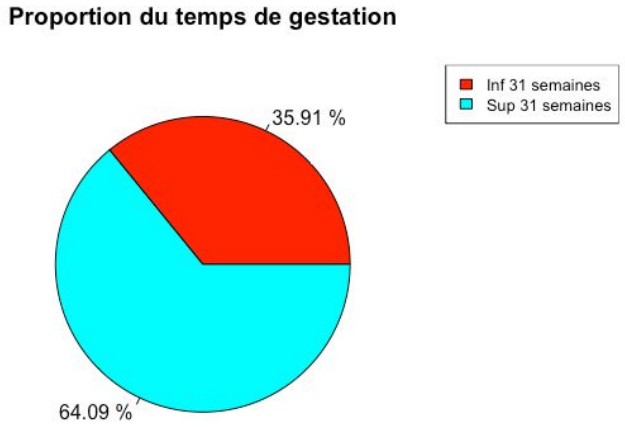
GEMEL : Généralement, les femmes ont eu une grossesse simple, mais 10% ont eu une grossesse multiple.



PREMATURE : Sur l'étude, il y a autant de femmes qui ont eu un accouchement prématuré que celles qui n'en ont pas eu.



On cherche à estimer des données brutes de proportion de femme ayant une période de gestation de moins de 31 semaines. Nous réalisons alors un test sur la densité.
En effectif, on remarque que 79 femmes ont eu une période de gestation de moins de 31 semaines. En proportion cela représente environ 36% des femmes de l'étude.
On peut observer le diagramme circulaire suivant :



Nous nous demandons si l'on peut considérer que la moitié des femmes ayant déjà connues un accouchement prématuré ont une période de gestation inférieur à 31 semaines. Ainsi on peut alors proposer un test sur une proportion. Ou on observe que les deux variables suivent une loi de Bernoulli de paramètre p.
On teste les hypothèses suivantes :

H0 : "p=0,5" contre H1 : "p < 0,5"

Pour ce test la statistique de test est :

$$\frac{\overline{X}_{1,n} - \overline{X}_{2,n}}{\sqrt{p\,(1-p) * (1/n1 + 1/n2)}} \rightarrow N(0,1) \text{ pour n grand}$$

ici n est de 220, donc assez grand.

où $p = \frac{n1 * \overline{X}_{1,n} - n2 * \overline{X}_{2,n}}{n1 + n2}$

La région critique de ce test est : [$-\infty$; 1.959964] .

On trouve une p-valeur de 0.9237 ; ainsi, nous ne rejetons pas H0. On considère alors que la moitié des femmes ayant déjà connues un accouchement prématuré ont une période de gestation inférieure à 31 semaines.

On cherche à savoir si la durée moyenne de gestation est différente entre les femmes ayant déjà connues un accouchement prématuré, et celles qui n'en ont jamais eu. On réalise alors un test sur l'égalité des moyennes sur deux échantillons, où un est l'échantillon des femmes ayant connues un accouchement prématuré et l'autre échantillon où ce sont des

femmes qui n'ont pas connu un accouchement prématuré. Les deux échantillons sont composés de 110 femmes. Les deux échantillons sont indépendants et de variances inconnues. On réalise donc *un test de Student*. On teste les hypothèses suivantes à un risque $\alpha = 5\%$:

H0 : "durée moyenne de gestation des femmes ayant déjà connues un accouchement prématuré = durée moyenne de gestation des femmes n'ayant pas connues un accouchement prématuré"

contre

H1 : "Les moyennes des deux échantillons ne sont pas égales (différentes)".

La p-valeur du test est de 0,022, elle est inférieur à α. Ainsi, on rejette l'hypothèse H0, les durées moyennes de gestation des femmes ayant déjà connues un accouchement prématuré sont différentes des durées moyennes de gestation des femmes n'ayant pas connues un accouchement prématuré.

Lien entre Premature & Gemel

On se demande s'il existe une relation entre l'accouchement prématurée et la type de grossesses (gémellaires oui/non). Nous réalisons alors un test sur l'indépendance des échantillons. Ce test est un test du chi-deux, avec les hypothèses suivantes à un risque $\alpha = 5\%$:

H0 : "Les deux variables accouchement prématuré et type de grossesse sont indépendantes".

Contre

H1 : "Les deux variables accouchement prématuré et type de grossesse sont corrélées (dépendantes)".

La p-valeur du test est de 3.628e-05, elle est inférieur à α. Ainsi, on rejette l'hypothèse H0, Les deux variables accouchement prématuré et type de grossesse sont donc corrélées (dépendantes).

Lien entre Membran & Consis

De plus, on réalise le même test avec les variables membranes rupturées et la consistance du col. Un test d'indépendance des variables, qui est un test du chi-deux. Ainsi on a les hypothèses suivantes à un risque $\alpha = 5\%$.

H0 : "Les deux variables membranes rupturées et consistance du col sont indépendantes"

Contre

H1 : "Les deux variables membranes rupturées et consistance du col sont corrélées (dépendantes)"

La p-valeur du test est de 0.009919, elle est inférieur à α. Ainsi, on rejette l'hypothèse H0. On peut alors dire que les deux variables membranes rupturées et consistance du col sont corrélées (dépendantes). Cependant si l'on avait fixé α à 1%, on aurait plutôt accepté l'hypothèse nulle car la p-valeur est égale à environ 1%.

Cependant on se demande si il est nécessaire d'analyser la modalité "incertaine" de la variable membrane ? Le résultat pourrait être faussé .

Nous allons effectuer un nouveau test du chi-deux en excluant la modalité incertaine, soit en excluant les femmes qui ne savent pas si il y a eu rupture de membrane (soit la réponse incertaine). Ainsi on a les hypothèses suivantes à un risque α=5% :

H0 : "Les deux variables membranes rupturées et consistance du col sont indépendantes"

Contre

H1 : "Les deux variables membranes rupturées et consistance du col sont corrélées (dépendantes)"

La p-valeur du test est de 0.7372, elle est supérieur à α. Ainsi, on ne rejette pas l'hypothèse H0, Les deux variables membranes rupturées et consistance du col sont indépendantes.

On conclut alors que la modalité incertaine avait tendance à faire corréler la variable membran à la variable de consistance du col.