PROJET D'ANALYSE STATISTIQUE DE DONNÉES CLIMATIQUES

Réalisé par :

Yuhe Bai - 9711
Jean-Baptiste de Bellescize - 10161
Hugo Tortosa - 10097

Sous la direction de:

Dr. Omar AL HAMMAL

Table des Matières

1.INTRODUCTION

Dans ce rapport, nous analyserons les données climatiques concernant la distribution de la température maximale quotidienne sur les mois de juillet et d'aôut
sur 2 périodes séparées de 10 ans:

Il s'agit d'une étude comparative des données de température d'une ville américaine entre les mois de Juillet et Août 2007 et ceux de 2017.

Nous effectuerons également une comparaison des résultats avec une valeur médiane 2012.

Tout ceci, dans le but de démontrer scientifiquement l'évolution des variations climatique au cours de ces dix dernières années (2007-2017) et plus précisément sur la première (2007-2012) ou deuxième partie (20012-2017).

Grâce aux données collectées et à leurs analyses, peut-on conclure à une évolution de la température impliquant le réchauffement climatique ?

Nous tenterons de répondre à cette problématique par diverses méthodes expliquées dans la seconde partie de ce rapport.

2. DESCRIPTION DES DONNÉES

Au cours de préparation, nous importons les fichiers contenant la température de 2007, 2017 ainsi que de 2012. Comme il y a quelques erreurs dans les fichiers, il y a plusieurs moyens d'effacer ces erreurs : soit effacer à la main, soit RStudio détecte les températures anormalement élevées et les enlèvent.

2.1 STATISTIQUES DESCRIPTIVES

Statistique (en chiffre):

Pour effectuer une analyse de statique descriptives complètes, nous avons choisi d'utiliser la fonction summary fournie dans RStudio, cette fonction donne le minimum, le maximum, la valeur à ¼(1er quartile), ½(médiane) 3/4(3ème quartile), et la moyenne. Tout cela donne une première idée sur la position et la dispersion pour la variable température, mais pas suffisant. On a donc choisi de rajouter quelques paramètres : la variance, l'écart type et le mode (la température la plus fréquente) et les concaténer sur la liste de summary.

Nous avons étudié pour chaque année et chaque mois. Voici le résultat pour les 3 différentes années :

```
>
> sumAnnee<-data.frame(sum2007,sum2012,sum2017)
> sumAnnee
            sum2007           sum2012           sum2017
1  Min.   :61.90    Min.   :60.00    Min.   :66.50
2  1st Qu.:70.10    1st Qu.:69.45    1st Qu.:72.90
3  Median :73.65    Median :74.60    Median :77.55
4  Mean   :74.30    Mean   :74.13    Mean   :76.96
5  3rd Qu.:79.03    3rd Qu.:78.58    3rd Qu.:80.30
6  Max.   :87.40    Max.   :89.40    Max.   :94.90
7   Variance: 35.1  Variance: 42.18  Variance: 33.57
8 Ecart Type: 5.92 Ecart Type: 6.49 Ecart Type: 5.79
9       Mode: 72.8       Mode: 78.9       Mode: 79.6
>
```

Les résultats de différentes années sont interprétés ensemble donc on visualise bien et on peut bien comparer le résultat de différentes années. Ici on observe que la variance et l'écart type n'ont pas beaucoup de différence mais la position des températures est devenue plus haut : c'est-à-dire le minimum, le maximum ainsi que la moyenne de température sont devenus plus haut, surtout pour l'année 2017 donc ça devient plus chaud.

Résultat pour chaque mois :

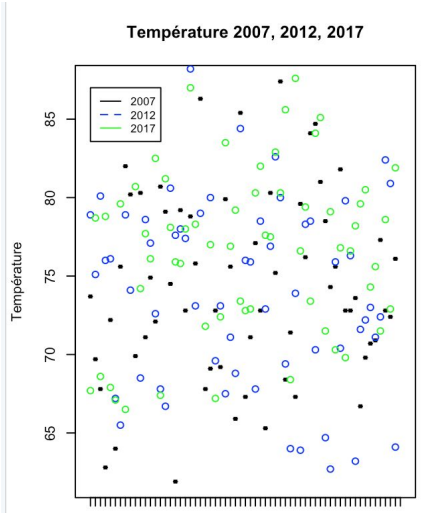
```
>
> sumJuillet<-data.frame(sum2007Juillet, sum2012Juillet, sum2017Juillet)
> sumJuillet
    sum2007Juillet   sum2012Juillet   sum2017Juillet
1  Min.   :66.70    Min.   :62.70    Min.   :67.40
2  1st Qu.:72.30    1st Qu.:71.35    1st Qu.:72.90
3  Median :75.60    Median :76.30    Median :76.60
4  Mean   :76.36    Mean   :75.38    Mean   :76.38
5  3rd Qu.:80.30    3rd Qu.:79.50    3rd Qu.:79.35
6  Max.   :87.40    Max.   :89.40    Max.   :94.90
7  Variance: 31.16  Variance: 44.23  Variance: 37.01
8 Ecart Type: 5.58 Ecart Type: 6.65 Ecart Type: 6.08
9       Mode: 75.6         Mode: 80       Mode: 79.6
>
```


```
>
> sumJuillet<-data.frame(sum2007Juillet, sum2012Juillet, sum2017Juillet)
> sumJuillet
    sum2007Juillet   sum2012Juillet   sum2017Juillet
1  Min.   :66.70    Min.   :62.70    Min.   :67.40
2  1st Qu.:72.30    1st Qu.:71.35    1st Qu.:72.90
3  Median :75.60    Median :76.30    Median :76.60
4  Mean   :76.36    Mean   :75.38    Mean   :76.38
5  3rd Qu.:80.30    3rd Qu.:79.50    3rd Qu.:79.35
6  Max.   :87.40    Max.   :89.40    Max.   :94.90
7  Variance: 31.16  Variance: 44.23  Variance: 37.01
8 Ecart Type: 5.58 Ecart Type: 6.65 Ecart Type: 6.08
9       Mode: 75.6         Mode: 80       Mode: 79.6
>
```

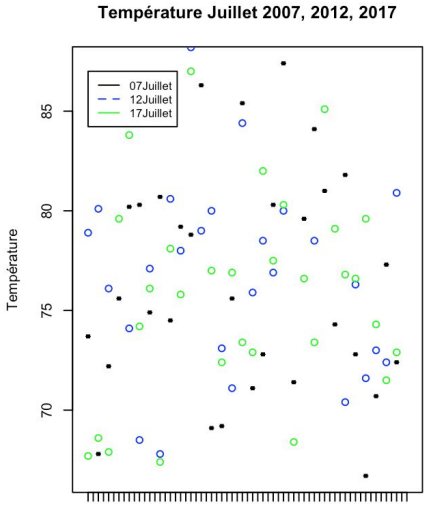

On observe que c'est à peu près la même situation que pour l'année, sauf que Juillet 2017 n'est pas très chaud, c'est août 2017 qui est très chaud par rapport aux autres années.

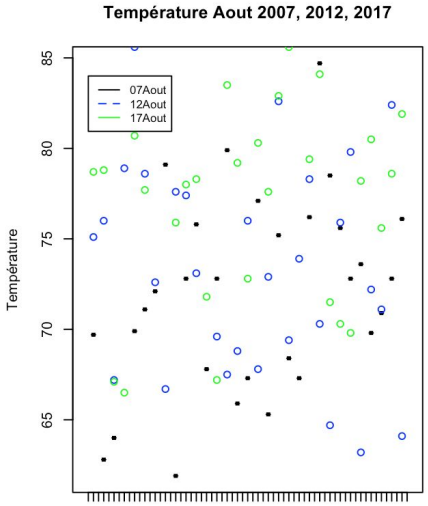
Visualisation : (schéma)

Pour encore une meilleure visualisation, au lieu de mettre des statistiques, nous avons interprétés plusieurs type de diagramme.
Pour visualiser, une première idée est de tracer des points pour toutes les valeurs de température. Voici le graphique pour toutes les températures de 2007, 2012 et 2017 :

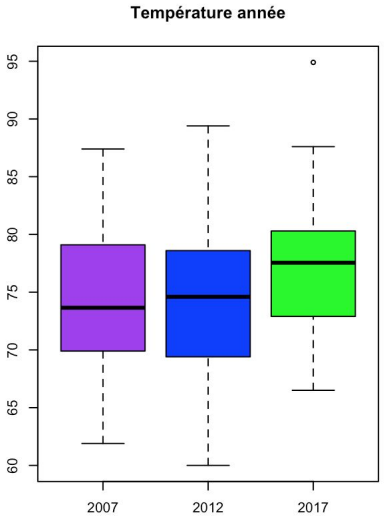


De même, pour les mois de juillet et d'août :



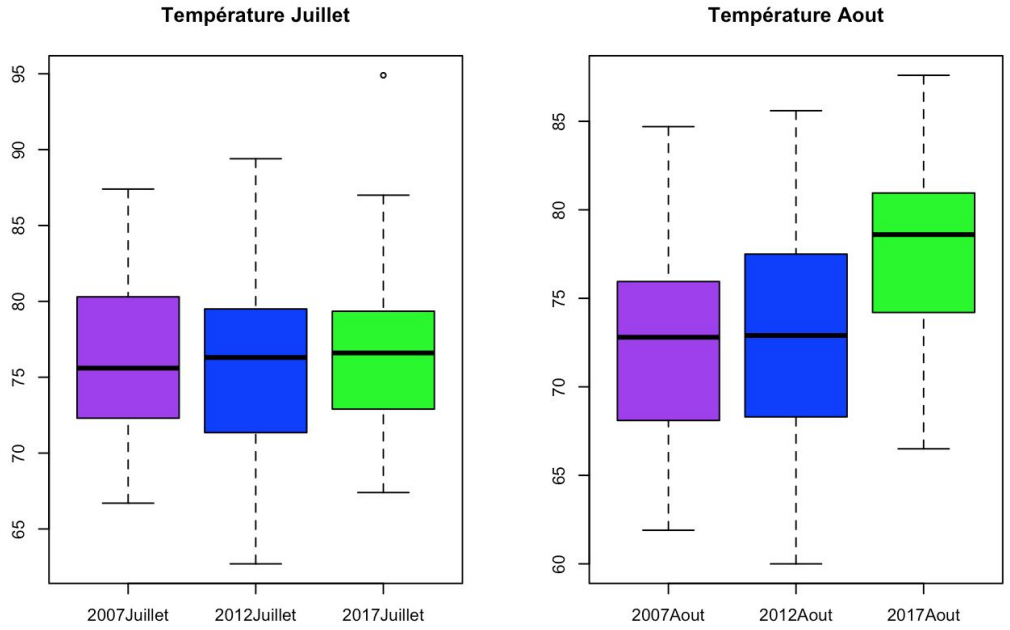


On observe que la température en 2017 est plus haute que les deux autres années, mais ce n'est pas une interprétation très claire.
Pour une meilleure visualisation, nous avons choisi d'utiliser le boxplot. En effet le boxplot ressemble beaucoup à summary sauf que les résultats sont interprétés en mode de box : donc le maximum, le minimum, la valeur à 1 /4 , 1/ 2, 3/ 4 etc. Voici le résultat pour la température de l'année :



On observe que la température en 2007 et 2012 n'ont pas beaucoup de différence, mais la température en 2017 est très haute par rapport aux autres années.
On note qu'il y a un point pour 2017, en effet c'est une valeur extrême(94.90 Farenheit) qui a été retirée pour la précision.

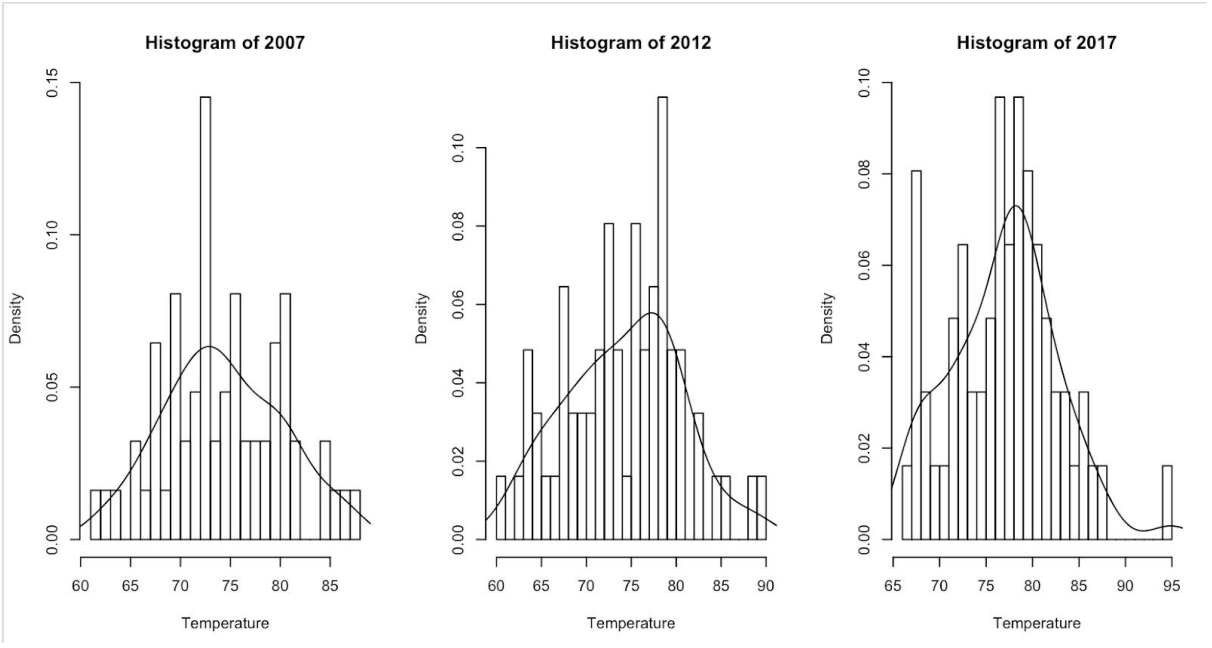
Résultat pour Juillet et Août :



On observe que la température en juillet 2017 est à peu près la même que les deux autres années, mais la température en août 2017 est extrêmement élevée par rapport aux deux autres années, donc c'est plutôt août 2017 qui est très chaude par rapport aux autres années.

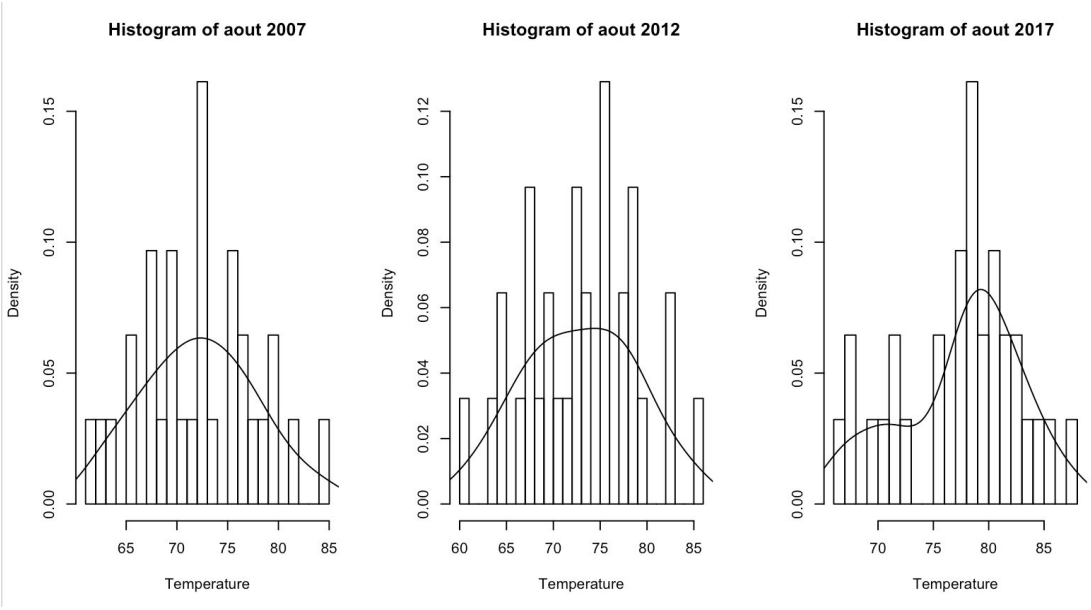
Histogramme :

L'histogramme est un moyen pour étudier la répartition d'une variable. Il permet également d'en déduire la loi de probabilité. Voici l'histogramme pour la température de l'année :



L'axe X est la température, l'axe Y est la densité de probabilité de cette température, c'est-à-dire la chance sur 1 que la température tombe ici. En traçant la ligne de densité, On observe une courbe qui représente très probablement à une loi normale. En effet nous allons beaucoup utiliser cette hypothèses dans les parties suivantes.

L'histogramme pour le mois de juillet et août :



Idem que pour toute l'année, mais on observe que les colonnes ainsi que les piques pour 2017 sont bougé un peu vers la droite par rapport aux autres années, c'est-à-dire qu'il y a plus de chance d'avoir une température plus haute en 2017.

2.2 ESTIMATION PONCTUELLE

Méthodes d'estimation ponctuelle

Nous connaissons deux méthodes d'estimation ponctuelle. Tout d'abord il y a la méthode de maximum de vraisemblance puis la méthode des moments.
Nous allons vous décrire ces méthodes et montrer en quoi elles peuvent servir pour notre projet d'estimation des températures. On notera θ le paramètre inconnu. L'objectif est l'estimation du paramètre θ. Il s'agit de donner, au vu des observations x1,..., xn, une approximation ou une évaluation de θ que l'on espère le plus proche possible de la vraie valeur inconnue.

Maximum de Vraisemblance :

Cette méthode permet de calculer, à partir d'un échantillon observé, la meilleure valeur d'un paramètre d'une loi de probabilité.
Soit X une variable aléatoire réelle, dont on veut estimer le paramètre θ.
On sait que:

$$f(x;\theta) = \begin{cases} f_\theta(x) & \text{si } X \text{ est une v.a. continue} \\ P_\theta(X = x) & \text{si } X \text{ est une v.a. discrète} \end{cases}$$

Hors,

$$f_\theta(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Ceci étant une fonction de θ avec x1, …, xn fixes, c'est une vraisemblance. C'est-à-dire :

$$L(\theta) = f_\theta(x_1, \ldots, x_n; \theta)$$

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n} P(X_i = x_i; \theta) = \prod_{i=1}^{n} P(X = x_i; \theta) & \text{si les } X_i \text{ sont discrètes} \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} f(x_i; \theta) & \text{si les } X_i \text{ sont continues} \end{cases}$$

On cherche à trouver le maximum de cette vraisemblance pour que les probabilités des réalisations observées soient aussi maximum. Ainsi en pratique :

- La condition nécessaire
  $$\frac{\partial L(x_1, \ldots, x_i, \ldots, x_n; \theta)}{\partial \theta} = 0$$
  ou
  $$\frac{\partial \ln L(x_1, \ldots, x_i, \ldots, x_n; \theta)}{\partial \theta} = 0$$
  permet de trouver la valeur $\theta = \hat{\theta}$.
- $\theta = \hat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique $\theta = \hat{\theta}$:
  $$\frac{\partial^2 L(x_1, \ldots, x_i, \ldots, x_n; \theta)}{\partial \theta^2} < 0$$
  ou
  $$\frac{\partial^2 \ln L(x_1, \ldots, x_i, \ldots, x_n; \theta)}{\partial \theta^2} < 0$$

Méthode des moments :

La méthode des moments se base sur une comparaison des moments théoriques E[X k] de la population et les moments expérimentaux.

- Les moments théoriques s'écrivent généralement en fonction de $\theta_1, \ldots, \theta_p$ tel que:

$$\begin{cases} \mathbb{E}[X] &= h_1(\theta_1, \ldots, \theta_p) \\ \mathbb{E}[X^2] &= h_2(\theta_1, \ldots, \theta_p) \\ &\vdots \\ \mathbb{E}[X^p] &= h_p(\theta_1, \ldots, \theta_p) \end{cases}$$

- On peut résoudre ce système d'équations pour obtenir:

$$\begin{cases} \theta_1 &= g_1(\mathbb{E}[X], \ldots, \mathbb{E}[X^p]) \\ \theta_2 &= g_2(\mathbb{E}[X], \ldots, \mathbb{E}[X^p]) \\ &\vdots \\ \theta_p &= g_p(\mathbb{E}[X], \ldots, \mathbb{E}[X^p]) \end{cases}$$

On remplace ensuite les moments théoriques par leurs estimateurs afin d'obtenir :

$$\begin{cases} \hat{\Theta_1} &= g_1(\mathbb{E}[\hat{X}], \ldots, \mathbb{E}[\hat{X}^p]) \\ \hat{\Theta_2} &= g_2(\mathbb{E}[\hat{X}], \ldots, \mathbb{E}[\hat{X}^p]) \\ &\vdots \\ \hat{\Theta_p} &= g_p(\mathbb{E}[\hat{X}], \ldots, \mathbb{E}[\hat{X}^p]) \end{cases}$$

avec :

$$\mathbb{E}[\hat{X}^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \forall k = 1 \ldots p$$

Les propriétés des estimateurs :

L'objectif dun estimateur est detre le plus proche possible de la valeur quil souhaite estimer . cest pour cela que lon utilise ces 3 propriétés qui nous permettend de calculer la qualite dun estimateur

On peut donc souhaiter que l'espérance de $\hat{\theta}$ soit égale à θ, C'est-à-dire que l'estimateur soit le plus proche possible de sa la valeur qu'il estime.

On sait que la valeur de l'estimateur peut fluctuer selon l'échantillon et quelle risque de ne pas coincider avec la valeur qu'il est censée représenter.

ces propriete permettent de calculer la qualite des estimateur.

Le biais

On peut donc souhaiter que l'espérance de $\hat{\theta}$ soit égale à θ, C'est-à-dire que l'estimateur soit le plus proche possible de sa la valeur qu'il estime.

$$\text{Biais}(\hat{\theta}) \equiv \mathbb{E}[\hat{\theta}] - \theta$$

Si le biais est égale à 0, alors l'estimateur nous donne la valeur correcte du paramètre.

La convergence

De plus, on peut dire qu'un estimateur consistant / convergent lorsqu'il converge vers le paramètre qu'il estime, en augmentant la taille de l'échantillon on espère pouvoir diminuer l'erreur. Voir formule ci-dessous :

$$\lim_{n\to\infty} Prob\{|\hat{\Theta} - \theta| < \varepsilon\} = 1$$

Ce qui nous indique que plus la taille de l'échantillon augmente, le plus on s'approche de la valeur à estimer.

La moyenne empirique est un estimateur convergent de l'espérance d'une variable aléatoire.

L'erreur quadratique moyenne

L'erreur quadratique moyenne d'un estimateur est une mesure caractérisant la « précision » de cet estimateur. Voir formule ci-dessous :

$$\textbf{Définition} - \text{MSE}(\hat{\theta}) \stackrel{\text{def}}{=} \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Ce qui nous indique que plus l'erreur quadratique est petite, plus l'estimateur est "précis".

Estimation ponctuelle de l'espérance

Supposons que les variables de températures sont issues d'une population Gaussienne, ainsi nous pouvons considérer que l'estimation de l'espérance est égal à la moyenne. Voici un tableau des estimations de l'espérance de chaque année et de chaque mois :

2007	2012	2017
74.29	74.13	76.96

	2007	2012	2017
Juillet	76.36	75.38387	76.38
Août	72.23	72.88065	77.54

Comme nous pouvons le voir ci-dessus, il y a un changement de température entre 2007 et 2017, ce changement a lieu entre les années 2012 et 2017 car comme nous pouvons

constater avec ces valeurs de l'espérance la température a tendance à baisser entre 2007 et 2012, on a une baisse de 1 degré entre juillet de 2007 et de 2012 suivi d'une augmentation générale de quasiment 2 degrés entre 2012 et 2017.

2.3 ESTIMATION PAR INTERVALLES DE CONFIANCE

Comme étudié précédemment, nous faisons l'hypothèse que la distribution des valeurs de température suit une loi normale. Ainsi nous obtenons la formule de l'intervalle de confiance à distribution normale, noté Ic, pour l'espérance.
On remarque que plus le niveau de confiance est élevé, plus le coefficient critique l'est aussi comme représenté sur le tableau ci-dessous.

$$I_c = \left[\bar{x} - t_\alpha \frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \right]$$

Formulation de l'intervalle de confiance autour d'une moyenne observée $\bar{x}$ avec un écart type observé s sur un échantillon de taille n.

Niveau de confiance C	Niveau de risque α	Coefficient critique t
90%	10%	1,645
95%	5%	1,960
99%	1%	2,576

Un intervalle de confiance à 95% pourcent signifie que si l'on prend 100 échantillons des températures et on calcule leur intervalle de confiance, il y aurait 95 intervalles de confiance qui contiennent la moyenne de température sur toute la population. (À noter que cela ne signifie pas que la moyenne de température de la population a une probabilité de 95% de se trouver dans cet intervalle de confiance.)

La traduction de cette formule dans le logiciel Rstudio se fait de cette manière:

```
Intervalle_de_confiance_2007_95=(mean(d2007[,2])+- 1.96 *(sd(d2007[,2])/sqrt(62)))
Intervalle_de_confiance_2007_99=(mean(d2007[,2])+- 2.58 *(sd(d2007[,2])/sqrt(62)))
```

Ici, on choisit d'étudier l'intervalle de confiance de l'espérance de l'année 2007:

- mean(d2007[,2]) correspond à la moyenne de l'année 2007. [,2] correspond au fait que l'on prend la deuxième colonne du tableau des données de température.
- 1,96 correspond à coefficient critique t=1,96 soit niveau de confiance c= 5 %
- sd(d2007[,2]) correspond à l'écart-type de l'année 2007. [,2] correspond au fait que l'on prend la deuxième colonne du tableau des données de température.
- sqrt(62) correspond à la racine de l'échantillon, ici n= 62 jours

Cependant, ici, nous utiliserons la méthode **t.test(X,Y,conf.level = c)$conf.int** car elle est automatique, plus rapide et surtout plus précise. Ainsi pour l'intervalle de confiance de 2007, le code est:

t.test(d2007[,2], conf.level = 0.95)$conf.int

1. *X= d(2007[,2])*
2. *conf.level = 0,95*
3. *$conf.int*

De **manière générale,**

- X = d(#Année[,2]): Cela correspond à la colonne de valeur des températures
- conf.level = c: On choisit de calculer l'intervalle de confiance de l'année voulue au seuil c = 5 % ou au seuil c = 1%
- $conf.int: correspond à l'affichage dans la console Rstudio que la partie à laquelle on s'intéresse, comme il y a beaucoup d'éléments dans t.test, comme ça il affiche que pour l'intervalle de confiance.

I-Année 2007

(Remarque: Nous avons également effectué l'intervalle de confiance "manuellement")

⟶ Cela affiche à peu près le même résultat (moins précis)

1.Intervalle de confiance à 95%

```
> t.test(d2007[,2], conf.level = 0.95)$conf.int
[1] 72.79226 75.80129
attr(,"conf.level")
[1] 0.95
```

Ici, l'estimation de l'intervalle de confiance de 2007 est Ic = [72.79 ; 75.8] au seuil c = 5%.

2. Intervalle de confiance à 99%

Même chose avec cette fois, c=1%

```
> t.test(d2007[,2], conf.level = 0.99)$conf.int
[1] 72.29626 76.29729
attr(,"conf.level")
[1] 0.99
```

Ici, l'estimation de l'intervalle de confiance de 2007 est Ic = [72.29 ; 76.29] au seuil c = 1%.

a) Mois Juillet 2007

- Intervalle de confiance au seuil c = 5%

```
> t.test(d2007Juillet[,2],conf.level = 0.95)$conf.int
[1] 74.31383 78.40875
attr(,"conf.level")
[1] 0.95
```

Ici, l'estimation de l'intervalle de confiance de 2007 est Ic = [74.31 ; 78.41] au seuil c = 5%.

- Intervalle de confiance Juillet 2007 au seuil c=1%

```
> t.test(d2007Juillet[,2],conf.level = 0.99)$conf.int
[1] 73.60431 79.11828
attr(,"conf.level")
[1] 0.99
```

Ici, l'estimation de l'intervalle de confiance de 2007 est Ic = [73.6 ; 79.11] au seuil c = 1%.

b)Mois Août 2007

- Intervalle de confiance au seuil c = 5%

```
> t.test(d2007Aout[,2],conf.level = 0.95)$conf.int
[1] 70.17684 74.28767
attr(,"conf.level")
[1] 0.95
```

- Intervalle de confiance au seuil 99%

```
> t.test(d2007Aout[,2],conf.level = 0.99)$conf.int
[1] 69.46457 74.99995
attr(,"conf.level")
[1] 0.99
```

II-Année 2017

Ici, l'estimation de l'intervalle de confiance de 2007 est Ic = [75.48 ; 78.43] au seuil c = 5%.

(Remarque: Nous avons également effectué l'intervalle de confiance "manuellement" ---> Cela affiche le même résultat (moins précis car moins de chiffre après la virgule)

- Intervalle de confiance à 95%

```
> t.test(d2017[,2], conf.level = 0.95)$conf.int
[1] 75.48980 78.43278
attr(,"conf.level")
[1] 0.95
```

- Intervalle de confiance à 99%

```
> t.test(d2017[,2], conf.level = 0.99)$conf.int
[1] 75.00468 78.91790
attr(,"conf.level")
[1] 0.99
```

a)Mois Juillet 2017

```
> t.test(d2017Juillet[,2], conf.level = 0.95)$conf.int
[1] 74.14914 78.61215
attr(,"conf.level")
[1] 0.95
```

```
> t.test(d2017Juillet[,2], conf.level = 0.99)$conf.int
[1] 73.37584 79.38545
attr(,"conf.level")
[1] 0.99
```

b)Mois Août 2017

```
> t.test(d2017Aout[,2],conf.level = 0.95)$conf.int
[1] 75.51419 79.56968
attr(,"conf.level")
[1] 0.95
> t.test(d2017Aout[,2], conf.level = 0.99)$conf.int
[1] 74.81150 80.27237
attr(,"conf.level")
[1] 0.99
```

Tableau récapitulatif des intervalles de confiance:

Année	**2007**	**2012**	**2017**
Intervalle de confiance à 95%	[72.79 ; 75.8]	[72.48 ; 75.78]	[75.48 ; 78.43]
Intervalle de confiance à 99%	[72.29 ; 76.29]	[71.94 ; 76.32]	[75.00 ; 78.91]

2.4 TEST D'HYPOTHÈSES

1. Premier test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2017 est compatible avec la température moyenne calculée en 2007.
2. Deuxième test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2007 est compatible avec la température moyenne calculée en 2017.
3. Troisième test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2007 est compatible avec la température moyenne calculée en 2012.
4. Quatrième test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2012 est compatible avec la température moyenne calculée en 2007.
5. Cinquième test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2012 est compatible avec la température moyenne calculée en 2007.
6. Sixième test d'hypothèse : **H0** : l'échantillon de mesures de températures relevées en 2012 est compatible avec la température moyenne calculée en 2007.

Comme étudié précédemment, nous faisons l'hypothèse que la distribution des valeurs de température suit une loi normale. Ainsi il y a une fonction t.test fournie dans R qui permet de faire le test d'hypothèse.

Explication t.test

```
# tests d'hypothese

testhypothese07_17 <- t.test(d2007[,2], mu=Esperance2017)
testhypothese07_17
testhypothese17_07 <- t.test(d2017[,2], mu=Esperance2007)
testhypothese17_07

testhypothese07_12 <- t.test(d2007[,2], mu=Esperance2012)
testhypothese07_12
testhypothese12_07 <- t.test(d2012[,2], mu=Esperance2007)
testhypothese12_07

testhypothese17_12 <- t.test(d2017[,2], mu=Esperance2012)
testhypothese17_12
testhypothese12_17 <- t.test(d2012[,2], mu=Esperance2017)
testhypothese12_17
```

On prend un intervalle de confiance avec la valeur par défaut qui est de 0.05. De plus on a choisi de faire un test qui n'est pas paired car il n'y a pas de lien direct entre les données des deux séries.

Afin de réaliser ce test d'hypothèse il faut un constructeur dans ce cas nous allons utiliser l'espérance mu. La règle de décision d'acceptation porte sur la comparaison de l'estimation de la moyenne sur un intervalle. Pour cela nous allons utiliser deux intervalles d'acceptation des hypothèse: 95% et 99%.

La p-value est la probabilité, sous H0, d'obtenir une statistique aussi extrême que la valeur observée sur l'échantillon. Aussi, pour un seuil de significativité alpha donné, on compare p_value et alpha(α), afin d'accepter, ou de rejeter H0.

- si $p \leq \alpha$, on va rejeter l'hypothèse H_0 (en faveur de H_1)
- si $p > \alpha$, on va rejeter H_1 (en faveur de H_0).

On peut alors interpréter la p-value comme le plus petit seuil de significativité pour lequel l'hypothèse nulle est acceptée.

Un autre moyen de faire le test d'hypothèse est de vérifier si la moyenne calculée d'un échantillon est dans l'intervalle de confiance de l'autre échantillon, ici dans le t.test l'intervalle de confiance est à 95% par défaut, donc on peut faire le test d'hypothèse que pour 95% et c'est plus compliqué de le faire alors qu'avec p-value, on peut facilement déterminer en la comparant à 0.05 ou 0.01. Ainsi nous avons choisi de faire le test d'hypothèse avec les p-values.

Interprétation :

1] Nous avons donc utilisé la méthode t.test() afin de comparer les données de 2007 avec la moyenne calculée de 2017

```
> testhypothese07_17 <- t.test(d2007[,2], mu=Esperance2017)
> testhypothese07_17

        One Sample t-test

data:  d2007[, 2]
t = -3.5414, df = 61, p-value = 0.0007694
alternative hypothesis: true mean is not equal to 76.96129
95 percent confidence interval:
 72.79226 75.80129
sample estimates:
mean of x
 74.29677
```

Étant donnés que la valeur de p-value est plus petite que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse nulle H0 pour l'hypothèse alternative H1. Nous pouvons même la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2007 est NON compatible avec la température moyenne calculée en 2017.

2] Les données de 2017 avec la moyenne calculée de 2007

```
> testhypothese17_07 <- t.test(d2017[,2], mu=Esperance2007)
> testhypothese17_07

        One Sample t-test

data:  d2017[, 2]
t = 3.6208, df = 61, p-value = 0.0005986
alternative hypothesis: true mean is not equal to 74.29677
95 percent confidence interval:
 75.48980 78.43278
sample estimates:
mean of x
 76.96129
```

Étant donnés que la valeur de p-value est plus petite que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse nulle H0 pour l'hypothèse alternative H1. Nous pouvons même la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2017 est NON compatible avec la température moyenne calculée en 2007.

3] Les données de 2007 avec la moyenne calculée de 2012

```
> testhypothese07_12 <- t.test(d2007[,2], mu=Esperance2012)
> testhypothese07_12

        One Sample t-test

data:  d2007[, 2]
t = 0.21866, df = 61, p-value = 0.8276
alternative hypothesis: true mean is not equal to 74.13226
95 percent confidence interval:
 72.79226 75.80129
sample estimates:
mean of x
 74.29677
```

Étant donnés que la valeur de p-value est plus grande que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse alternative H1 et accepter l'hypothèse nulle H0. Nous pouvons aussi la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2007 est compatible avec la température moyenne calculée en 2012.

4] Les données de 2012 avec la moyenne calculée de 2007

```
> testhypothese12_07 <- t.test(d2012[,2], mu=Esperance2007)
> testhypothese12_07

        One Sample t-test

data:  d2012[, 2]
t = -0.19945, df = 61, p-value = 0.8426
alternative hypothesis: true mean is not equal to 74.29677
95 percent confidence interval:
 72.48286 75.78166
sample estimates:
mean of x
 74.13226
```

Étant donnés que la valeur de p-value est plus grande que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse alternative H1 et accepter l'hypothèse nulle H0. Nous pouvons aussi la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2012 est compatible avec la température moyenne calculée en 2007.

5] Les données de 2017 avec la moyenne calculée de 2012

```
> testhypothese17_12 <- t.test(d2017[,2], mu=Esperance2012)
> testhypothese17_12

        One Sample t-test

data:  d2017[, 2]
t = 3.8444, df = 61, p-value = 0.0002909
alternative hypothesis: true mean is not equal to 74.13226
95 percent confidence interval:
 75.48980 78.43278
sample estimates:
mean of x
 76.96129
```

Étant donnés que la valeur de p-value est plus petite que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse nulle H0 pour l'hypothèse alternative H1. Nous pouvons même la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2017 est NON compatible avec la température moyenne calculée en 2012.

6] Les données de 2012 avec la moyenne calculée de 2017

```
> testhypothese12_17 <- t.test(d2012[,2], mu=Esperance2017)
> testhypothese12_17

        One Sample t-test

data:  d2012[, 2]
t = -3.4297, df = 61, p-value = 0.001089
alternative hypothesis: true mean is not equal to 76.96129
95 percent confidence interval:
 72.48286 75.78166
sample estimates:
mean of x
 74.13226
```

Étant donnés que la valeur de p-value est plus petite que l'intervalle de confiance de 0.05 nous pouvons donc rejeter l'hypothèse nulle H0 pour l'hypothèse alternative H1. Nous pouvons même la rejeter pour un intervalle de confiance de 0.01. Ainsi l'échantillon de mesures de températures relevées en 2012 est NON compatible avec la température moyenne calculée en 2017.

2.5 TEST DE COMPARAISON DE MOYENNES

Au début nous prenons l'hypothèse que la distribution des valeurs de température suit une loi normale : c'est la prémisse pour faire le t-test. Sinon il faut utiliser d'autres méthodes qui seront développées dans la partie suivante.

Le test-t de Student est un test statistique permettant de comparer les moyennes de deux groupes d'échantillons. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes au point de vue statistique.

Ici, nous utilisons la formule d'un **test de Student pour séries non-appariées car** la valeur des deux séries ont un lien : elles partagent la même unité en température pour une même période de mesure et ont la même taille d'échantillon. Cependant, on part du principe que les deux moyennes des deux groupes d'échantillons sont différents et n'ont aucun lien. Le test est donc **indépendant.**
Ci-dessous, la formule donnant la loi de Student pour séries non-appariées:

Formule

- Soit A et B deux groupes différents à comparer.
- Soit m_A et m_B la moyenne du groupe A et celui du groupe B, respectivement.
- Soit n_A et n_B la taille du groupe A et celle du groupe B, respectivement.

La **valeur t de Student** est donnée par la formule suivante:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

S^2 est la **variance** commune aux deux groupes. Elle est calculée par la formule suivante :

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au **risque alpha** = 5% pour un degré de liberté :

$$d.d.l = n_A + n_B - 2$$

I-Test entre 2007 et 2017

```
> t.test(d2007[,2],d2017[,2])

        Welch Two Sample t-test

data:  d2007[, 2] and d2017[, 2]
t = -2.5318, df = 121.94, p-value = 0.01262
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.7479340 -0.5810983
sample estimates:
mean of x mean of y
 74.29677  76.96129
```

Analyse: entre 2007 et 2017

Ici, on compare les moyennes de température des années 2007 et 2017.:

L'intervalle de confiance de la différence des moyennes à 95% est également montrée (intervalle de confiance= [-4.7479340,-0.5810983]) intervalle non-centré ; et enfin, on a la valeur moyenne des deux groupes (température moyenne de x:2007 = 74.29677, température moyenne y:2017 =76.96129)

Dans le résultat ci-dessus : t est la statistique de student (t = -2.5318), df est le degré de liberté (df= 121.94), p-value est le degré de significativité du test (p-value = 0.01262).

La valeur t mesure la taille de la différence par rapport à la variation des données des échantillons. Une statistique t supérieure à 1.96 (ou inférieure à -1.96) indique que le coefficient est significatif avec un degré de confiance supérieur à 95%. Le signe ne fait aucune différence car les deux signes sont interprétés de la même manière - en tant que preuve contre l'hypothèse nulle (l'hypothèse selon laquelle il n'y a pas de différence significative entre des populations spécifiées).

Ici, t = -2.5318 soit t < -1.96, on en déduit que le coefficient est significatif avec un degré supérieur à 95%.

Ici aussi, le degré de liberté (d.d.l), donné par la formule page 1, est correspond à la somme des tailles du groupe A et du groupe B étudiés (soit 2007 et 2012 par exemple) moins 2. ce qui est égal à (62 + 62 - 2 = 122)

La valeur p-value est la probabilité, sous H0, d'obtenir une statistique aussi extrême (pour ne pas dire aussi grande) que la valeur observée sur l'échantillon. Aussi, pour un seuil de significativité α(alpha) donné, on compare p et α(alpha), afin d'accepter, ou de rejeter H0,

- si $p \leq \alpha$, on va rejeter l'hypothèse H_0 (en faveur de H_1)
- si $p > \alpha$, on va rejeter H_1 (en faveur de H_0).

On peut alors interpréter la p-value comme le plus petit seuil de significativité pour lequel l'hypothèse nulle est acceptée.

	H_0 vraie	H_1 vraie
accepter H_0	OK	*erreur type 2*
rejeter H_0	*erreur type 1*	OK

Ici, p-value = 0.01262 soit **p-value <α(alpha)** ⇔0.01262 < 0.05.
Ho = pas de différence significative, H1 = différence significative
On déduit qu'on rejette l'hypothèse Ho (en faveur de H1).
la valeur **p-value** indique un preuve **contre** l'hypothèse nulle **Ho**.

Entre 2007 et 2017, on conclut que la moyenne des température a évolué de manière significative.

II-Test entre 2007 et 2012

```
> t.test(d2007[,2],d2012[,2])

        Welch Two Sample t-test

data:  d2007[, 2] and d2012[, 2]
t = 0.14735, df = 120.98, p-value = 0.8831
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.045818  2.374850
sample estimates:
mean of x mean of y
 74.29677  74.13226
```

Analyse: entre 2007 et 2012

On compare les moyennes de température des années 2007 et 2012:
Ici, p-value = 0.8831 soit **α(alpha) < p-value**
Ho = pas de différence significative, H1 = différence significative
On déduit qu'on accepte l'hypothèse Ho (rejet de H1).
la valeur **p-value** indique un preuve **en faveur** de l'hypothèse nulle **Ho**.
Entre 2007 et 2012, on conclut que la moyenne des température n'a pas significativement évolué de manière drastique.

III-Test entre 2012 et 2017

```
> t.test(d2012[,2],d2017[,2])

        Welch Two Sample t-test

data:  d2012[, 2] and d2017[, 2]
t = -2.5593, df = 120.44, p-value = 0.01173
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.0175656 -0.6404989
sample estimates:
mean of x mean of y
 74.13226  76.96129
```

Analyse: entre 2012 et 2017

On compare les moyennes de température des années 2012 et 2017 :
Ici, p-value = 0.01173 soit **p-value <α(alpha)** ⇔ 0.01262 < 0.05.
Ho = pas de différence significative, H1 = différence significative
On déduit qu'on rejette l'hypothèse Ho (en faveur de H1).
la valeur p-value indique un preuve contre l'hypothèse nulle Ho.
Entre 2012 et 2017, on conclut que la moyenne des température a évolué de manière significative.

3. CONCLUSION

Tout au long de notre analyse statistique,nous avons pu observer une augmentation de température entre les années 2007 et 2017, grâce aux valeur obtenu en 2012 nous pouvons savoir plus précisément que la température a subi son augmentation entre 2012 et 2017. Lors du calcul de l'espérance (voir tableau partie 2.2) nous avons pu constater que l'augmentation est plus importante au mois d'août qu'au mois de juillet. Cela ce traduit par des valeurs min et max plus élevé pour les séries 2017 que celle de 2007 et 2012. On a donc pu vérifier ces résultat en utilisant un test d'hypothèse dans lequel on a pu accepter ou rejeter l'hypothèse null qui indique que la l'échantillon de mesures de températures relevées une certaine année est compatible avec la température moyenne calculée une autre année. Les test d'hypothèse ont effectivement confirmé nos résultat préalables. Nous avons fait de même avec le test de Student permettant de comparer les moyennes de deux groupes d'échantillons et de savoir si les moyennes des deux groupes sont significativement différentes. Les valeurs de p_value ont corroboré le travail déjà effectué.

Nous avons donc trouvé des valeurs logique tout au long de notre projet qui correspondent à une augmentation de la température. Nous pourrions donc utiliser ce projet pour prouver qu'il y a belle et bien un réchauffement climatique mais il faut bien prendre en compte que l'augmentation a lieu essentiellement entre 2012 et 2017 contrairement aux études qui indique une augmentation constante tout au long des deux dernières décennies, et que ces valeurs correspondent à un seul endroit donc pour pouvoir affirmer cela il faudrait effectuer ces tests à plusieurs endroits distincts. Cela peut aussi s'expliquer par le fait que 2012 fut une année froide. C'est pour cela qu'il faudrait collecter les données annuelles pour faire une analyse plus profonde et en tirer des conclusions certaines

Comment changerait l'analyse si les variables provenaient d'une distribution non-gaussienne ?

Nous avons supposé que les observations étaient indépendantes et qu'elles étaient issues d'une population Gaussienne. En effet beaucoup d'analyses que l'on fait sont basés sur la distribution gaussienne, il faut vérifier que ce soit une loi normale d'abord pour pouvoir utiliser ces analyses.
Nous avons eu une première visualisation à partir de l'histogramme, pour un résultat plus précis, nous avons utilisé le Shapiro test pour vérifier si la température suit bien une loi normale.

Hypothèse nulle : l'échantillon suit une loi normale.

```
> shapiro.test(d2007[,2])

        Shapiro-Wilk normality test

data:  d2007[, 2]
W = 0.9882, p-value = 0.8165

> shapiro.test(d2012[,2])

        Shapiro-Wilk normality test

data:  d2012[, 2]
W = 0.98641, p-value = 0.7247

> shapiro.test(d2017[,2])

        Shapiro-Wilk normality test

data:  d2017[, 2]
W = 0.97545, p-value = 0.2483
```

En regardant la valeur W et la p-value, si la valeur W est proche de 1 et la p-value est supérieur à 0.05, on accepte l'hypothèse nulle et on peut dire que c'est une distribution gaussienne. Ainsi la valeur de températures des trois années suivent tous la loi normale, on obtient le même résultat pour chaque mois. Donc on peut bien effectuer un T test.

Si les variables proviennent d'une distribution non-gaussienne, un test non-paramétrique tel que le test de Wilcoxon est recommandé comme une alternative au test de Student.
Le test de Wilcoxon (ou de Mann-Whitney) est un test non-paramétrique de comparaison de moyennes de deux échantillons. Ce test est dit non-paramétrique car il ne fait aucune hypothèse sur la distribution des échantillons.

H0 : la valeur moyenne des deux échantillons ne sont pas significativement différents.

- si $p \leq \alpha$, on va rejeter l'hypothèse H_0 (en faveur de H_1)
- si $p > \alpha$, on va rejeter H_1 (en faveur de H_0).

2007-2017

```
> wilcox.test(d2007[,2], d2017[,2])

        Wilcoxon rank sum test with continuity correction

data:  d2007[, 2] and d2017[, 2]
W = 1433.5, p-value = 0.01472
alternative hypothesis: true location shift is not equal to 0
```

En regardant la p-value, ici p-value<0.05 donc on rejette H0.
Entre 2007 et 2017, on conclut que la moyenne des température a significativement évolué.

2007-2012

```
> wilcox.test(d2007[,2], d2012[,2])

        Wilcoxon rank sum test with continuity correction

data:  d2007[, 2] and d2012[, 2]
W = 1924, p-value = 0.994
alternative hypothesis: true location shift is not equal to 0
```

Ici p-value > 0.05 donc on accepte H0,
Entre 2007 et 2012, on conclut que la moyenne des température n'a pas significativement évolué de manière drastique.

2012-2017

```
> wilcox.test(d2012[,2], d2017[,2])

        Wilcoxon rank sum test with continuity correction

data:  d2012[, 2] and d2017[, 2]
W = 1452, p-value = 0.01896
alternative hypothesis: true location shift is not equal to 0
```

Ici p-value < 0.05 donc on rejette H0.
Entre 2012 et 2017, on conclut que la moyenne des température a significativement évolué.
On obtient le même résultat qu'avec T test comme prévu.

En effet, même si ce n'est pas une loi normale, on peut quand même utiliser les méthodes d'analyses pour la distribution gaussienne pour un échantillon de taille >30, et c'est grâce au théorème central limite. Il établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. C'est-à-dire pour la taille d'échantillon supérieur à 30, la somme (la moyenne) de tous ces échantillons suit une loi normale. Ainsi la moyenne de température de chaque année suit une loi normale même si leur distribution de température ne suit pas une loi normale, on peut donc appliquer le t.test.
C'est un théorème très important et il nous permet de pouvoir appliquer beaucoup de méthodes d'analyses basés sur la loi normale aux distributions non-gaussienne.

4. BIBLIOGRAPHIE

- Statistical Tools For High-Throughput Data Analysis

http://www.sthda.com/french/wiki/test-de-student-formules

- Test de Student non-appariées avec R

http://www.sthda.com/french/wiki/test-de-student-non-apparie-avec-r-comparaison-de-moyennes-de-deux-groupes-d-echantillons-independants

- R pour les statophobes

https://perso.univ-rennes1.fr/denis.poinsot/Statistiques_%20pour_statophobes/R%20pour%20les%20statophobes.pdf

- Estimations et intervalles de confiance

https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-estim.pdf

- Intervalle de confiance et de fluctuation

https://euler.ac-versailles.fr/IMG/pdf/fluctuconf2.pdf

5. ANNEXE

```
rm(list = setdiff(ls(), lsf.str())) # remove all variables except for functions
par(mfrow=c(1,3))

# 2
# Import the file of 2007 and 2017 with read.csv
d2007 <- read.csv(file= "/Users/Cecile/Documents/ISEP/A11/ProbStat/dJul07_62.csv",
header = TRUE)
d2012 <- read.csv(file
="/Users/Cecile/Documents/ISEP/A11/ProbStat/dJul12_62_groupe3.csv", header = TRUE)
d2017 <- read.csv(file= "/Users/Cecile/Documents/ISEP/A11/ProbStat/dJul17_62.csv",
header = TRUE)

# Delete error data by hand (or in excel directly)
# d2007 <- d2007[-c(63),]
# d2017 <- d2017[-c(53),]

# Delete error by detection
d2007=d2007[which(d2007$Température<=150),]
d2012=d2012[which(d2012$Température<=150),] # normally there is no error, just in case
d2017=d2017[which(d2017$Température<=150),]

# Obtain data by month

d2007Juillet <- d2007[c(1:31),]
d2007Aout <- d2007[-c(1:31),]

d2017Juillet <- d2017[c(1:31),]
d2017Aout <- d2017[c(32:62),]

d2012Juillet <- d2012[c(1:31),]
d2012Aout <- d2012[c(32:62),]

# 2.1
# Summary of year

sum2007 = summary(d2007)[c(1:6),2]
sum2007 = c(sum2007,paste("Variance:",round(var(d2007[,2]),2)))
sum2007 = c(sum2007,paste("Ecart Type:",round(sd(d2007[,2]),2)))
sum2007 = c(sum2007,paste("Mode:",getmode(d2007[,2])))
sum2007

sum2012 = summary(d2012)[c(1:6),2]
```

```
sum2012 = c(sum2012,paste("Variance:",round(var(d2012[,2]),2)))
sum2012 = c(sum2012,paste("Ecart Type:",round(sd(d2012[,2]),2)))
sum2012 = c(sum2012,paste("Mode:",getmode(d2012[,2])))
sum2012

sum2017 = summary(d2017)[c(1:6),2]
sum2017 = c(sum2017,paste("Variance:",round(var(d2017[,2]),2)))
sum2017 = c(sum2017,paste("Ecart Type:",round(sd(d2017[,2]),2)))
sum2017 = c(sum2017,paste("Mode:",getmode(d2017[,2])))
sum2017

sumAnnee<-data.frame(sum2007,sum2012,sum2017)
sumAnnee

# Summary Juillet

sum2007Juillet = summary(d2007Juillet)[c(1:6),2]
sum2007Juillet = c(sum2007Juillet,paste("Variance:",round(var(d2007Juillet[,2]),2)))
sum2007Juillet = c(sum2007Juillet,paste("Ecart Type:",round(sd(d2007Juillet[,2]),2)))
sum2007Juillet = c(sum2007Juillet,paste("Mode:",getmode(d2007Juillet[,2])))
sum2007Juillet

sum2012Juillet = summary(d2012Juillet)[c(1:6),2]
sum2012Juillet = c(sum2012Juillet,paste("Variance:",round(var(d2012Juillet[,2]),2)))
sum2012Juillet = c(sum2012Juillet,paste("Ecart Type:",round(sd(d2012Juillet[,2]),2)))
sum2012Juillet = c(sum2012Juillet,paste("Mode:",getmode(d2012Juillet[,2])))
sum2012Juillet

sum2017Juillet = summary(d2017Juillet)[c(1:6),2]
sum2017Juillet = c(sum2017Juillet,paste("Variance:",round(var(d2017Juillet[,2]),2)))
sum2017Juillet = c(sum2017Juillet,paste("Ecart Type:",round(sd(d2017Juillet[,2]),2)))
sum2017Juillet = c(sum2017Juillet,paste("Mode:",getmode(d2017Juillet[,2])))
sum2017Juillet

sumJuillet<-data.frame(sum2007Juillet, sum2012Juillet, sum2017Juillet)
sumJuillet

# Summary Aout
sum2007Aout = summary(d2007Aout)[c(1:6),2]
sum2007Aout = c(sum2007Aout,paste("Variance:",round(var(d2007Aout[,2]),2)))
sum2007Aout = c(sum2007Aout,paste("Ecart Type:",round(sd(d2007Aout[,2]),2)))
sum2007Aout = c(sum2007Aout,paste("Mode:",getmode(d2007Aout[,2])))

sum2012Aout = summary(d2012Aout)[c(1:6),2]
sum2012Aout = c(sum2012Aout,paste("Variance:",round(var(d2012Aout[,2]),2)))
```

```
sum2012Aout = c(sum2012Aout,paste("Ecart Type:",round(sd(d2012Aout[,2]),2)))
sum2012Aout = c(sum2012Aout,paste("Mode:",getmode(d2012Aout[,2])))

sum2017Aout = summary(d2017Aout)[c(1:6),2]
sum2017Aout = c(sum2017Aout,paste("Variance:",round(var(d2017Aout[,2]),2)))
sum2017Aout = c(sum2017Aout,paste("Ecart Type:",round(sd(d2017Aout[,2]),2)))
sum2017Aout = c(sum2017Aout,paste("Mode:",getmode(d2017Aout[,2])))

sumAout<-data.frame(sum2007Aout, sum2012Aout, sum2017Aout)
sumAout

# plot the tempratures
# A little bug for the dates, but no influence on data.
# Year
plot(d2007, type="p", main="Température 2007, 2012, 2017")
points(d2012, col="blue")
points(d2017, col="green")
legend(1, 87, legend=c("2007", "2012", "2017"),col=c("black", "blue", "green"), lty=1:2,
cex=0.8)

# Juillet
plot(d2007Juillet, type="p", main="Température Juillet 2007, 2012, 2017")
points(d2012Juillet, col="blue")
points(d2017Juillet, col="green")
legend(1, 87, legend=c("07Juillet", "12Juillet", "17Juillet"),col=c("black", "blue", "green"),
lty=1:2, cex=0.8)

# Aout
plot(d2007Aout, type="p", main="Température Aout 2007, 2012, 2017")
points(d2012Aout, col="blue")
points(d2017Aout, col="green")
legend(1, 84, legend=c("07Aout", "12Aout", "17Aout"),col=c("black", "blue", "green"), lty=1:2,
cex=0.8)

# Boxplot
# boxplot year
boxplot(d2007$Température, d2012$Température, d2017$Température,
main="Température année",
      names=c("2007","2012","2017"),col=c("purple","blue","green"))

# boxplot Juillet
boxplot(d2007Juillet$Température, d2012Juillet$Température, d2017Juillet$Température,
main="Température Juillet",
      names=c("2007Juillet","2012Juillet","2017Juillet"),col=c("purple","blue","green"))
```

```
# boxplot Aout
boxplot(d2007Aout$Température, d2012Aout$Température, d2017Aout$Température,
main="Température Aout",
        names=c("2007Aout","2012Aout","2017Aout"),col=c("purple","blue","green"))

# Histogram year

hist(d2007[,2], breaks=30, freq=FALSE, main= "Histogram of 2007",xlab="Temperature")
lines(density(d2007[,2]),col = "black")
hist(d2012[,2], breaks=30, freq=FALSE, main= "Histogram of 2012",xlab="Temperature")
lines(density(d2012[,2]),col = "black")
hist(d2017[,2], breaks=30, freq=FALSE, main= "Histogram of 2017",xlab="Temperature")
lines(density(d2017[,2]),col = "black")

# Histogram juillet

hist(d2007Juillet[,2], breaks=30, freq=FALSE, main= "Histogram of juillet
2007",xlab="Temperature")
lines(density(d2007Juillet[,2]),col = "black")
hist(d2012Juillet[,2], breaks=30, freq=FALSE, main= "Histogram of juillet
2012",xlab="Temperature")
lines(density(d2012Juillet[,2]),col = "black")
hist(d2017Juillet[,2], breaks=30, freq=FALSE, main= "Histogram of juillet
2017",xlab="Temperature")
lines(density(d2017Juillet[,2]),col = "black")

# Histogram aout

hist(d2007Aout[,2], breaks=30, freq=FALSE, main= "Histogram of aout
2007",xlab="Temperature")
lines(density(d2007Aout[,2]),col = "black")
hist(d2012Aout[,2], breaks=30, freq=FALSE, main= "Histogram of aout
2012",xlab="Temperature")
lines(density(d2012Aout[,2]),col = "black")
hist(d2017Aout[,2], breaks=30, freq=FALSE, main= "Histogram of aout
2017",xlab="Temperature")
lines(density(d2017Aout[,2]),col = "black")
```

 # 2.2

```
# Estimation ponctel de l'esperance

Y2007= d2007[,2]
Y2017= d2017[,2]
Y2012= d2012[,2]
size2007=length(Y2007)
```

```
size2012=length(Y2012)
size2017=length(Y2017)

J2007= d2007[c(1:31),2]
J2012= d2012[c(1:31),2]
J2017= d2017[c(1:31),2]
sizeJ2007=length(J2007)
sizeJ2012=length(J2012)
sizeJ2017=length(J2017)

A2007= d2007[c(32:62),2]
A2012= d2012[c(32:62),2]
A2017= d2017[c(32:62),2]
sizeA2007=length(A2007)
sizeA2012=length(A2012)
sizeA2017=length(A2017)

sum=sum(Y2007)
Esperance2007=sum/size2007
Esperance2007

sum7=sum(Y2012)
Esperance2012= sum7/size2012
Esperance2012

sum2=sum(Y2017)
Esperance2017=sum2/size2017
Esperance2017

sum3=sum(J2007)
EsperanceJ2007=sum3/sizeJ2007
EsperanceJ2007


sum4=sum(J2017)
EsperanceJ2017=sum4/sizeJ2017
EsperanceJ2017


sum5=sum(A2007)
EsperanceA2007=sum5/sizeA2007
EsperanceA2007

sum6=sum(A2017)
EsperanceA2017=sum6/sizeA2017
```

```
EsperanceA2017

sum8=sum(J2012)
EsperanceJ2012= sum8/sizeJ2012
EsperanceJ2012

sum9=sum(A2012)
EsperanceA2012= sum9/sizeA2012
EsperanceA2012
```

 # 2.3

```
# Intervalle de confiance year
# Intervalle de confiance 2007
t.test(d2007[,2], conf.level = 0.95)$conf.int
t.test(d2007[,2], conf.level = 0.99)$conf.int

# La facon manuelle
# alpha = 5%
Intervalle_de_confiance_2007_95 = (mean(d2007[,2])- 1.96 *(sd(d2007[,2])/sqrt(62))) # min
Intervalle_de_confiance_2007_95 = paste(Intervalle_de_confiance_2007,(mean(d2007[,2])+
1.96 *(sd(d2007[,2])/sqrt(62)))) # max
Intervalle_de_confiance_2007_95
# alpha = 1%
Intervalle_de_confiance_2007_99 = (mean(d2007[,2])- 2.576 *(sd(d2007[,2])/sqrt(62))) # min
Intervalle_de_confiance_2007_99 = paste(Intervalle_de_confiance_2007,(mean(d2007[,2])+
2.576 *(sd(d2007[,2])/sqrt(62)))) # max
Intervalle_de_confiance_2007_99

# Intervalle de confiance 2012
t.test(d2012[,2], conf.level = 0.95)$conf.int
t.test(d2012[,2], conf.level = 0.99)$conf.int

# Intervalle de confiance 2017
t.test(d2017[,2], conf.level = 0.95)$conf.int
t.test(d2017[,2], conf.level = 0.99)$conf.int

#Interavlle de confiance pour chaque mois
#Juillet
#intervalle pour juillet 2007
t.test(d2007Juillet[,2], conf.level = 0.95)$conf.int
t.test(d2007Juillet[,2], conf.level = 0.99)$conf.int

#intervalle pour juillet 2012
t.test(d2012Juillet[,2], conf.level = 0.95)$conf.int
t.test(d2012Juillet[,2], conf.level = 0.99)$conf.int
```

```
#intervalle pour juillet 2017
t.test(d2017Juillet[,2], conf.level = 0.95)$conf.int
t.test(d2017Juillet[,2], conf.level = 0.99)$conf.int

#Aout
#intervalle pour Aout 2007
t.test(d2007Aout[,2], conf.level = 0.95)$conf.int
t.test(d2007Aout[,2], conf.level = 0.99)$conf.int

#intervalle pour Aout 2012
t.test(d2012Aout[,2], conf.level = 0.95)$conf.int
t.test(d2012Aout[,2], conf.level = 0.99)$conf.int

#intervalle pour Aout 2017
t.test(d2017Aout[,2], conf.level = 0.95)$conf.int
t.test(d2017Aout[,2], conf.level = 0.99)$conf.int
```

 # 2.4

```
# tests d'hypothese
# 2007 et 2017
testhypothese07_17 <- t.test(d2007[,2], mu=Esperance2017)
testhypothese07_17
testhypothese17_07 <- t.test(d2017[,2], mu=Esperance2007)
testhypothese17_07

# 2007 et 2012
testhypothese07_12 <- t.test(d2007[,2], mu=Esperance2012)
testhypothese07_12
testhypothese12_07 <- t.test(d2012[,2], mu=Esperance2007)
testhypothese12_07

# 2012 et 2017
testhypothese17_12 <- t.test(d2017[,2], mu=Esperance2012)
testhypothese17_12
testhypothese12_17 <- t.test(d2012[,2], mu=Esperance2017)
testhypothese12_17
```

#2.5

```
# test de comparaison
t.test(d2007[,2],d2012[,2])
t.test(d2007[,2],d2017[,2])
t.test(d2012[,2],d2017[,2])
```

 # Conclusion

```
# Test de normality
shapiro.test(d2007[,2])
shapiro.test(d2012[,2])
shapiro.test(d2017[,2])

shapiro.test(d2007Juillet[,2])
shapiro.test(d2012Juillet[,2])
shapiro.test(d2017Juillet[,2])

shapiro.test(d2007Aout[,2])
shapiro.test(d2012Aout[,2])
shapiro.test(d2017Aout[,2])

# Test non parametric
wilcox.test(d2007[,2], d2017[,2])
wilcox.test(d2007[,2], d2012[,2])
wilcox.test(d2012[,2], d2017[,2])


#functions
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```