

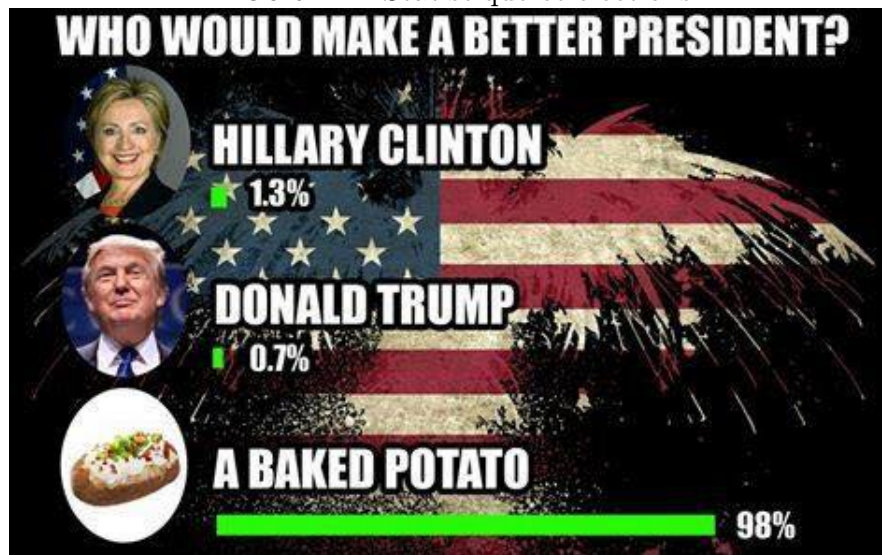
T. D. n° 2 4056

Résumé

Ce document est le T. D. n° 2 du module 4056. Il reprend rapidement des éléments du cours et propose une mise en pratique interactive des notions de tests statistiques, de variance et d'intervalle de confiance.

1 Élection américaine

FIGURE 1 – Statistique et élections



source :www.liberalvaluesblog.com

1.1 Chargement des données

Vous avez parié beaucoup d'argent sur la victoire d'Hillary Clinton et vous avez malheureusement perdu. A défaut d'être riche, vous allez essayer de savoir pourquoi les sondages ont échoué.

Cet ensemble de données est une collection de sondages nationaux et locaux menés de novembre 2015 à novembre 2016 sur l'élection présidentielle américaine de 2016. Les données donnent les résultats des sondages bruts et pondérés par état, date, genre, etc. Il a été généré par le site de fivethirtyeight.com.

Il y a 27 variables.

Les champs incluent :

- *cycle* : année de l'élection
- *branch* : type d'élection

- *type* : type du sondage
- *matchup* : candidats proposés
- *forecastdate* : date d'extraction des données (11/01/16)
- *state* : état(s) sondé(s)
- *startdate* : début du sondage
- *enddate* : fin du sondage
- *pollster* : organisme sondeur
- *grade* : facteur qualitatif du sondage calculé par fivethirtyeight
- *samplesize* : échantillons de personnes interrogées
- *populaion* : population interrogée (A = adulte, RV = inscrit sur les listes, LV = enclin à voter, V = votants)
- *poll_wt* : le poids affecté au sondage par fivethirtyeight
- *rawpoll_clinton* : scores du sondage associé à Clinton
- *rawpoll_trump* : scores du sondage associé à Trump
- *rawpoll_johnson* : scores du sondage associé à Johnson
- *rawpoll_mcmullin* : scores du sondage associé à Mc Mullin (valeur souvent absente)
- *adypoll_clinton* : scores du sondage associé à Clinton après ajustement de fivethirtyeight
- *adypoll_trump* : scores du sondage associé à Trump après ajustement de fivethirtyeight
- *adypoll_johnson* : scores du sondage associé à Johnson après ajustement de fivethirtyeight
- *adypoll_mcmullin* : scores du sondage associé à Mc Mullin après ajustement de fivethirtyeight
- *multiversions* : variable non exploitée ici (probablement associée à variante du sondage)
- *url* : url du sondage
- *poll_id* : identifiant du sondage
- *question_id* : identifiant du type de question ici
- *createddate* : variable non exploitée ici
- *timestamp* : variable non exploitée ici

Après avoir téléchargé le fichier *presidential_polls.csv* chargez-le sous **R** (vous stockerez ces données sous la variable *Data*). Il est également possible de télécharger le fichier *.csv* depuis : <https://www.kaggle.com/fivethirtyeight/2016-election-polls>.

```
>Data <- read.csv("C:/Users/claeys/Documents/Cours/cour ESIEA/4a/
  presidential_polls.csv", sep=',')
```

Stockez les données dans un *dataframe* appelé *dataf*. Appliquez la fonction *summary()* à votre *dataframe*.

```
> summary(dataf)
      cycle      branch      type
      Min.    :2016    President:10236    now-cast    :3412    Clinton vs.
      Trump vs. Johnson:10236
```

1st Qu.:2016 polls-only:3412
 Median :2016 polls-plus:3412
 Mean :2016
 3rd Qu.:2016
 Max. :2016

forecastdate	state	startdate	enddate	
11/1/16:10236	U.S.	:3105	10/20/2016:	450
10/31/2016:	324			
	Florida	: 360	10/25/2016:	393 9/1/2016
	:	300		
	North Carolina:	303	10/10/2016:	270
			10/26/2016:	273
	Pennsylvania	: 285	9/27/2016 :	261
			10/30/2016:	258
	Ohio	: 273	10/17/2016:	231
			10/24/2016:	240
	New Hampshire :	246	10/24/2016:	204
			10/25/2016:	228
	(Other)	:5664	(Other)	:8427 (Other)
		:8613		
	pollster	grade	samplesize	
	population			
Ipsos	:2373	A-	:2796	Min. : 35 a :
57				
Google Consumer Surveys	:2070	B	:2622	1st Qu.: 440 lv
:8880				
SurveyMonkey	:1059	C-	:1146	Median : 772 rv
:1173				
USC Dornsife/LA Times	: 342		:1134	Mean : 1116 v :
126				
CVOTER International	: 327	C+	: 846	3rd Qu.: 1225
The Times-Picayune/Lucid:	288	B+	: 552	Max. :40816
(Other)	:3777	(Other):	1140	NA's :3
poll_wt	rawpoll_clinton	rawpoll_trump	rawpoll_johnson	
rawpoll_mcmullin				
Min. :0.000000	Min. :11.04	Min. : 4.00	Min. : 0.000	
Min. : 9.00				
1st Qu.:0.000444	1st Qu.:38.00	1st Qu.:34.88	1st Qu.: 6.000	
1st Qu.:20.00				
Median :0.008206	Median :43.00	Median :39.00	Median : 7.250	
Median :26.00				
Mean :0.221769	Mean :41.72	Mean :39.23	Mean : 7.713	
Mean :23.23				
3rd Qu.:0.101939	3rd Qu.:46.00	3rd Qu.:44.00	3rd Qu.: 9.000	
3rd Qu.:29.00				
Max. :7.591351	Max. :88.00	Max. :68.00	Max. :25.000	
Max. :31.00				
			NA's :3756	
			NA's	
			:10185	
adjpoll_clinton	adjpoll_trump	adjpoll_johnson	adjpoll_mcmullin	
multiversions				

Min. :16.42	Min. : 3.234	Min. : -3.372	Min. :13.97
:10203			
1st Qu.:40.15	1st Qu.:37.973	1st Qu.: 3.322	1st Qu.:21.41
*: 33			
Median :43.73	Median :42.035	Median : 4.502	Median :26.09
Mean :43.02	Mean :42.049	Mean : 4.804	Mean :24.88
3rd Qu.:46.51	3rd Qu.:45.374	3rd Qu.: 5.855	3rd Qu.:29.67
Max. :86.70	Max. :71.865	Max. :19.278	Max. :31.98
		NA's :3756	NA's :10185

url
<https://datast...2M/page/GsS:1251>
[http://www.reuters.co\[...\]:792](http://www.reuters.co[...]:792)
[https://drive.google.\[...\]:765](https://drive.google.[...]:765)
[http://polling.reuter\[...\]:519](http://polling.reuter[...]:519)
[https://www.surveymon\[...\]:459](https://www.surveymon[...]:459)
[http://polling.reuter\[...\]:426](http://polling.reuter[...]:426)
 (Other) : 6024

poll_id	question_id	createddate	timestamp
Min. :35362	Min. :47244	9/9/16 : 609	14:57:58 1 Nov
2016:3412			
1st Qu.:44851	1st Qu.:68747	5/5/16 : 528	15:03:47 1 Nov
2016:3412			
Median :45926	Median :71774	10/27/16: 414	15:09:38 1 Nov
2016:3412			
Mean :45318	Mean :70588	11/1/16 : 378	
3rd Qu.:47008	3rd Qu.:73498	9/28/16 : 354	
Max. :48000	Max. :75067	10/10/16: 330	
		(Other) :7623	

Dans ce jeu donnée, il existe trois types de sondage.

- poll-plus : Combine les sondages avec un indice économique. Il suppose que la course va resserrer un peu.
- poll-only : Une version plus simple? Il suppose que les sondages les plus récents reflètent la meilleure prévision pour novembre, bien qu'avec beaucoup d'incertitude.
- now-cast : Une projection de ce qui se passerait dans une élection hypothétique tenue le jour du sondage. Beaucoup plus agressif que les autres modèles.

Les sondages de type *poll-plus* commence en supposant que les sondages d'électeurs probables sont meilleurs pour les républicains; *poll-only* ne fait pas une telle hypothèse. Les sondages de type *poll-plus* soustraient les points des candidats tiers au début de la course, tandis que les sondages *poll-only* ne le font pas. Les sondages de type *poll-plus* et *poll-only* utilisent une régression basée sur la démographie et l'historique des votes passés. Les sondages de type *now-cast* sont des sondages en ligne. Il pondère les sondages récents plus fortement et est plus agressif dans le calcul d'une ligne de tendance.

À vous !

- Selon vous, quel type de sondage présente le plus d'incertitude ?
- Au cours du temps, les sondages auront-ils tendance à se rapprocher ou au contraire, à s'éloigner ? Pourquoi ?
- Pensez-vous que les sondages *now-cast* soient fiables ? Pourquoi ?
- Pour une analyse globale, regardez-vous plutôt la moyenne ou la médiane ? Pourquoi ?
- Commentez les résultats de la fonction **summary()**. Quels «sondage type» avons nous ?

1.2 Test de K-S

La variable *poll_wt* traduit la valeur du sondage en fonction de la qualité de son fournisseur (ils sont tous notés de D à A +) et la taille de l'échantillon. Mais il s'avère être également fondé en grande partie sur combien de temps avant l'élection les données ont été recueillies.

Chaque sondage a deux séries de pourcentages pour chacun des quatre candidats présidentiels : Clinton, Trump, McMullin et Johnson, mais Mc Mullin manque souvent. Les deux ensembles sont étiquetés «*raw*» et «*adj*». L'approche de *FiveThirtyEight* est que la valeur «*adj*» reflète l'étalonnage pour le biais statistique historique des sondages individuels (c.-à-d. si les sondages ont constamment surestimé leur prédiction pour un parti politique dans le passé, leur proportion rapportée pour le parti est ajustée à la baisse). Les données ajustées et non ajustées sont disponibles. Ni les séries brutes ni les séries rajustées s'élèvent à 100%, vraisemblablement en raison de répondants indécis ou non-informatifs. Nous affichons l'historique des scores de Clinton et de Trump pour les valeurs brutes

```
> hist(dataf$rawpoll_clinton)
> hist(dataf$rawpoll_trump)
```

FIGURE 2 – Histogramme des scores de Clinton (bruts)

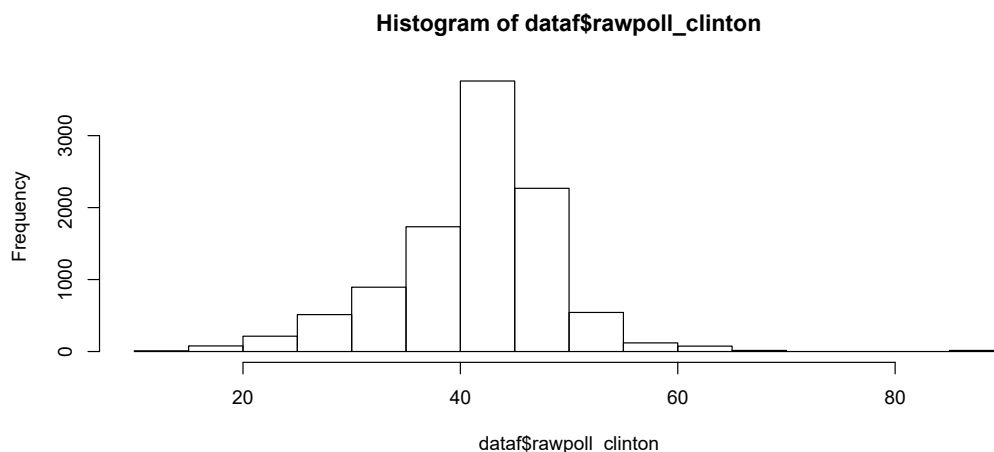
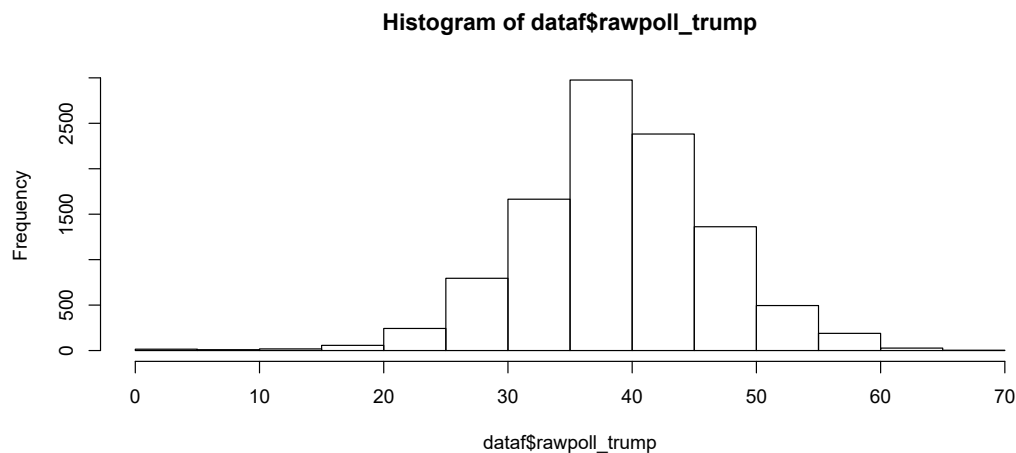


FIGURE 3 – Histogramme des scores de Trump (bruts)



Vous allez appliquer le test de Kolmogorov-Smirnov aux séries de points x_1 et x_2 définies avec les commandes ci-dessous :

```
> x1 <- rnorm(nrow(dataf), mean = mean(dataf$rawpoll_clinton))
> plot(density(x1))
> ks.test(dataf$rawpoll_clinton, x1)
```

Two-sample Kolmogorov-Smirnov test

```
data: dataf$rawpoll_clinton and x1
D = 0.40875, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Warning message:

```
In ks.test(dataf$rawpoll_clinton, x1) :
  les valeurs p seront approximées en présence d'ex-aequo
```

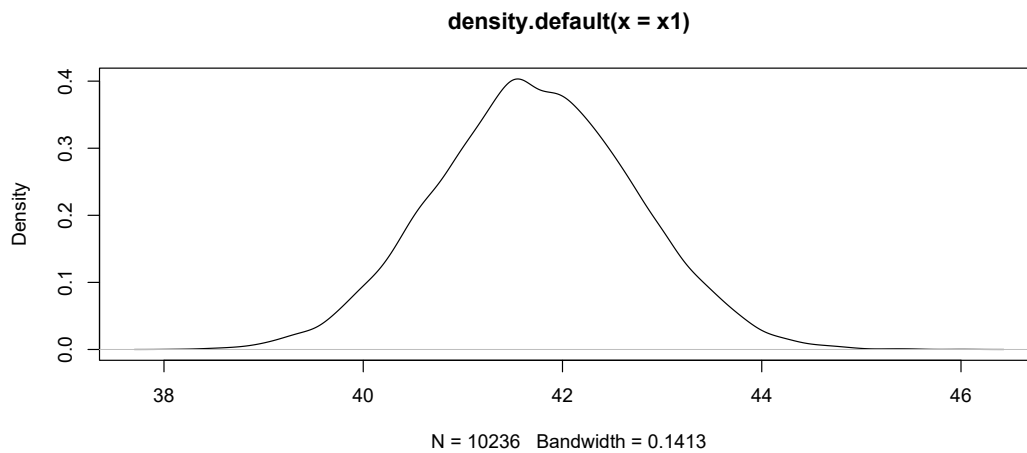
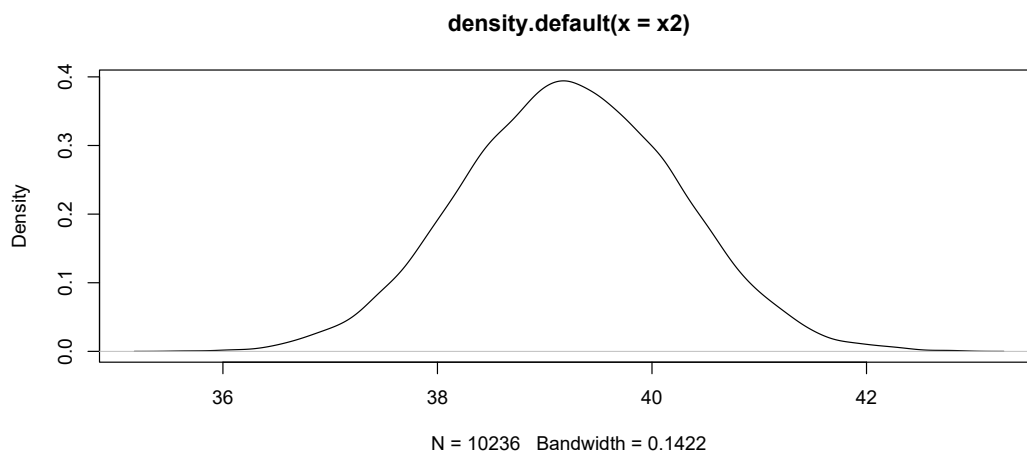
```
> x2 <- rnorm(nrow(dataf), mean = mean(dataf$rawpoll_trump))
> plot(density(x2))
> ks.test(dataf$rawpoll_trump, x2)
```

Two-sample Kolmogorov-Smirnov test

```
data: dataf$rawpoll_trump and x2
D = 0.38189, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Warning message:

```
In ks.test(dataf$rawpoll_trump, x2) :
  les valeurs p seront approximées en présence d'ex-aequo
```

FIGURE 4 – Densité de probabilité de x_1 FIGURE 5 – Densité de probabilité de x_2 

À vous !

- Affichez les histogrammes des scores bruts de Clinton et de Trump mais cette fois ajustés par FiveThirtyEight.
- Comparez les différences avec les histogrammes des scores de Clinton et Trump bruts.
- A quoi correspond x_1 ? Expliquez la Figure 4.
- A quoi correspond x_2 ? Expliquez la Figure 5.
- Quelle est la particularité du test de Kolmogorov-Smirnov ? Pourquoi l'avons nous utilisé avec le couple $(x_1, \text{rawpoll_clinton})$ et $(x_2, \text{rawpoll_trump})$?
- Quel est le résultat du test de Kolmogorov-Smirnov sur le couple couple $(x_1, \text{rawpoll_clinton})$?

- g) Quel est le résultat du test de Kolmogorov-Smirnov sur le couple couple ($x2$, *rawpoll_trump*) ?
- h) Quel autre test auriez vous plus appliquer ?
- i) Pensez-vous que de regarder le résultat global de tous les sondages bruts suffise à vous donner une bonne estimation du résultats des élections ? Pourquoi ?

1.3 Le principe de l'Analyse de la variance

L'analyse de la variance permet d'étudier le comportement d'une variable quantitative à expliquer en fonction d'une ou de plusieurs variables qualitatives. Lorsque l'on souhaite étudier le comportement de plusieurs variables à expliquer en même temps, on utilisera une analyse de la variance multiple (MANOVA). Si un modèle contient des variables explicatives catégorielles et continues on utilisera plutôt une analyse de la covariance (ANCOVA).

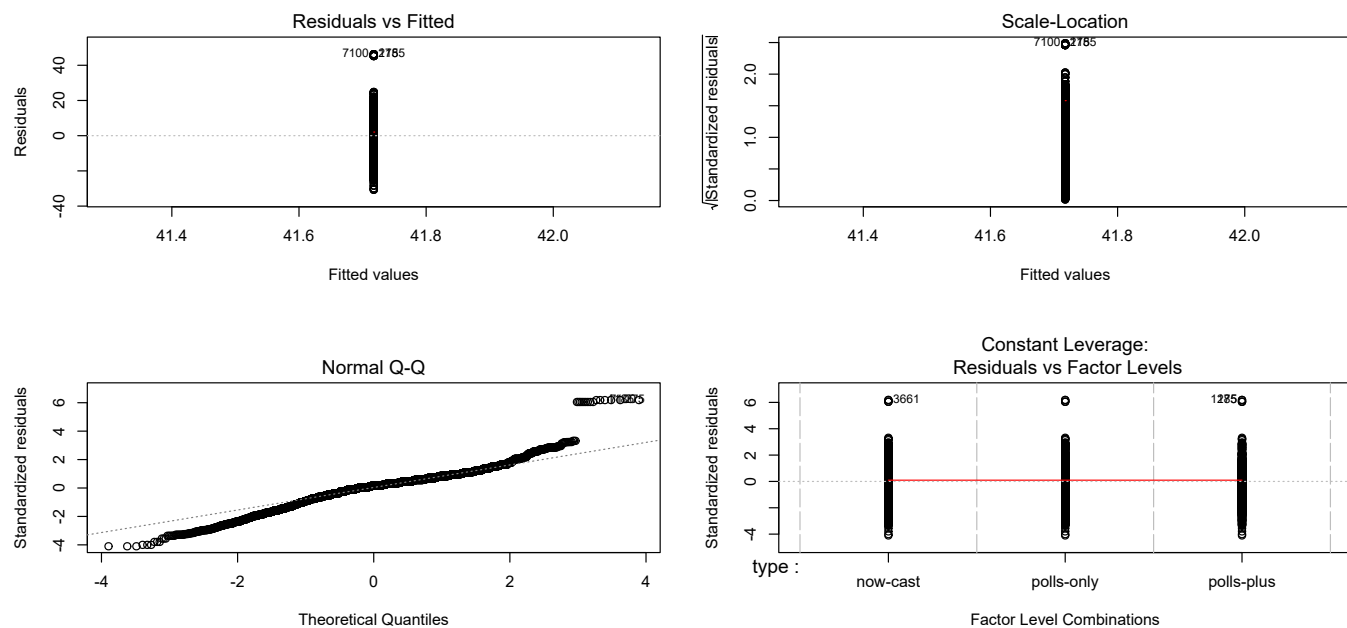
Intuitivement, l'analyse de la variance à facteurs fixes vise à comparer des moyennes sur plusieurs populations. Lors du premier T. D., vous aviez comparé la variance entre deux genres de jeux. Certain tests¹ vous permettent de savoir si la variances de deux populations sont égales. Cependant, pour un nombre d'échantillons supérieurs à deux, nous utilisons une ANOVA (ANalysis Of VAriance).

L'hypothèse à vérifier (H_0) est que tous les échantillons ont la même moyenne. L'hypothèse alternative est qu'au moins l'un d'eux possède une moyenne sensiblement différente des autres. Le but est en principe d'identifier un facteur de variabilité, notamment dans le cadre de plans d'expériences. On souhaite savoir si Clinton a eu la même moyenne pour tout les type de sondage

```
>fit <- aov(rawpoll_clinton ~ type, data=dataf)
>layout(matrix(c(1,2,3,4),2,2)) # optional layout
>plot(fit) # diagnostic plots
```

1. notamment le test de Fisher

FIGURE 6 – Anova de rawpoll_clinton selon différents type de tests



Le principe est celui de la décomposition de la variance (modèle et erreur). L'ANOVA ne compare pas deux variances de populations mais bien les deux composantes d'une même variance. Le test ne « fonctionnera » que si les moyennes sont les mêmes dans tous les échantillons. En statistique, l'on parle d'hétéroscédasticité lorsque les variances des variables examinées sont différentes.

La courbe résiduel vs fitted value montre s'il y a un modèle dans les résidus, et idéalement devrait montrer une diffusion similaire pour chaque condition. Ici aucune modélisation n'a pu être correctement réalisée. La courbe Normal Q-Q recherche la normalité des résidus. Si elles ne sont pas normales, les hypothèses permettant de réaliser l'ANOVA sont potentiellement non respectées. La courbe Scale-Location, comme la courbe Normal Q-Q, test spécifiquement si les résidus augmentent avec les valeurs ajustées. Comme nous nous pouvons le voir le modèle est particulièrement mauvais. La courbe constant Leverage donne une idée des niveaux du facteur qui sont les mieux adaptés.

À vous !

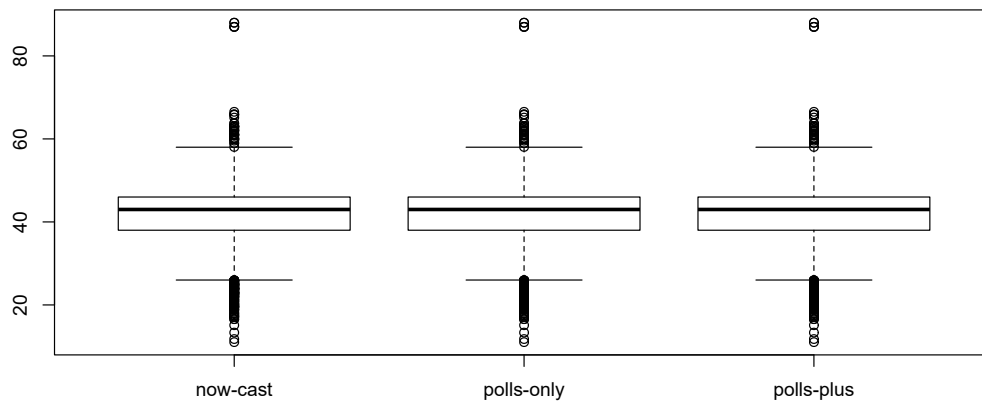
- Interprétez la courbe Normal Q-Q.
- Suite aux résultats du test de l'ANOVA, le test est-il adapté ?
- Par conséquent, quelles sont les deux pré-requis indispensables avant d'utiliser ANOVA ?

1.4 Une alternative à l'ANOVA : le test de Kruskal-Wallis

Le test de Kruskal-Wallis est considéré comme l'alternative non paramétrique de l'ANOVA dès que la distribution sous-jacente des données n'est plus gaussienne. Il est extrêmement populaire en data Science. Commencez par charger la librairie *graphics* et afficher la boîte à moustaches des valeurs obtenu par Clinton pour les trois types de tests

```
> require(graphics)
> boxplot(rawpoll_clinton ~ type, data = dataf)
```

FIGURE 7 – Boxplot des score de Clinton type de tests



On applique le test de Kruskal-Wallis sur la variance avec les trois type de tests. L'hypothèse H_0 Il permet de tester si k échantillons ($k > 2$) proviennent de la même population, ou de populations ayant des caractéristiques identiques, au sens d'un paramètre appelé position (le paramètre de position est conceptuellement proche de la médiane, mais le test de Kruskal-Wallis prend en compte plus d'information que la position au seul sens de la médiane)

```
> kruskal.test(rawpoll_clinton ~ type, data=dataf)
```

Kruskal-Wallis rank sum test

data: rawpoll_clinton by type

Kruskal-Wallis chi-squared = 0, df = 2, p-value = 1

À vous !

- Interprétez le résultat de Kruskal-Wallis .
- Réalisez le le test de Kruskal-Wallis sur *rawpoll_trump* et *type*.
- Concluez sur les différents types de sondage.

- d) Réalisez le test de Kruskal-Wallis sur *rawpoll_clinton* et *genre*.
- e) Réalisez le test de Kruskal-Wallis sur *rawpoll_trump* et *genre*.
- f) Concluez sur les qualités des sondages.

1.5 Intervalle de confiance

Faire une prédiction est une bonne chose, mais les statisticiens n'aiment pas toujours trop s'engager avec des variables aléatoires. Nous utilisons un estimateur θ qui permet de calculer une valeur prédite (par exemple les résultats de Clinton) à partir d'un échantillon qui devrait être proche de la réalité sans pour autant savoir si cette valeur est totalement fiable. L'intervalle de confiance fournit un intervalle dans lequel se trouve θ avec une probabilité de $1 - \alpha$, α étant le risque que l'on se fixe, généralement petit. Par exemple je vous prédis avec un risque de 0% que votre moyenne à ce module se trouvera dans l'intervalle $[0 - 20]$ (théoriquement...).

On suppose que les observations x_1, \dots, x_n sont issues de n variables aléatoires indépendantes X_1, \dots, X_n de même loi $\mathcal{N}(\mu, \sigma^2)$.

Si la loi n'est pas gaussienne, on suppose alors que la taille de l'échantillon est grande ($n \geq 30$ en pratique). Le théorème central limite (T.C.L.) permet de faire des approximations par des lois normales, ce qui donnera des intervalles de confiance approximatifs mais parfois suffisant en pratique. Vous allez afficher les valeurs de *rawpoll_clinton* pour le genre. Vous stockerez ces valeurs dans la variable *rowClinA*

```
> rowClinA <- dataf[ which(dataf$grade=='A+'),]$rawpoll_clinton
```

Nous allons chercher à calculer son intervalle de confiance à $1 - \alpha$ en fixant α à 5% à partir de la moyenne μ noté $IC_{1-\alpha}(\mu)$ et pour variance σ^2 noté $IC_{1-\alpha}(\sigma^2)$. Il est possible de calculer l'intervalle de confiance lorsque σ est inconnu mais nous ne verrons pas ce cas dans ce T. D. Soit \bar{X} , l'estimateur ponctuel de μ calculé sur l'échantillon ω de taille n . L'I.C. selon μ , ω , et n se est donné par :

$$IC_{1-\alpha}(\mu) = \left[\bar{X}(\omega) - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}(\omega) + t_\alpha \frac{\sigma}{\sqrt{n}} \right] \quad (1)$$

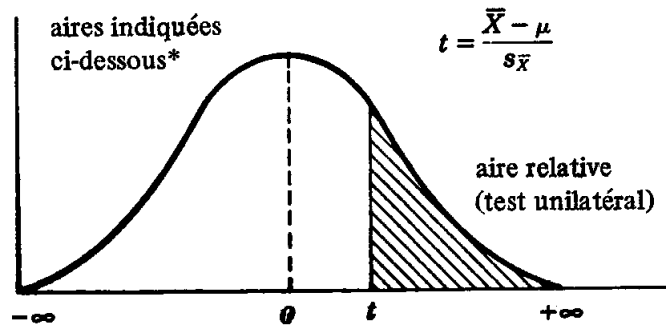
la valeur de t_α est donnée dans la table de Student, selon le risque α associée. Ainsi :

$$P(|U| \leq t_\alpha) = 1 - \alpha \Leftrightarrow P(U \leq t_\alpha) = 1 - \frac{\alpha}{2} \quad (2)$$

U étant une valeur aléatoire future.

DISTRIBUTION t DE STUDENT

Table des valeurs de t en fonction du nombre ν de degrés de liberté et de l'aire relative comprise entre t et $+\infty$



	0,50	0,40	0,30	0,20	0,10	0,05	0,025	0,010	0,005	0,001	0,0005	$\alpha = 1 - P$
$\nu \backslash P$	0,50	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995	
1	0,000	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6	
2	0,000	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60	
3	0,000	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94	
4	0,000	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610	
5	0,000	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859	
6	0,000	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959	
7	0,000	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405	
8	0,000	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041	
9	0,000	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781	
10	0,000	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587	
11	0,000	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437	
12	0,000	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318	
13	0,000	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221	
14	0,000	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140	
15	0,000	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073	
16	0,000	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015	
17	0,000	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965	
18	0,000	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922	
19	0,000	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883	
20	0,000	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850	
21	0,000	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819	
22	0,000	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792	
23	0,000	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767	
24	0,000	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745	
25	0,000	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725	
26	0,000	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707	
27	0,000	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690	
28	0,000	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674	
29	0,000	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659	
30	0,000	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646	
40	0,000	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551	
60	0,000	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	2,232	3,460	
80	0,000	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415	
100	0,000	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389	
200	0,000	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339	
∞	0,000	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291	

FIGURE 8 – Table de Student

```
> plotmeans(dataf$adjpoll_clinton ~ dataf$grade, xlab="grade",
+           ylab="Clinton result", main="Mean Plot\nwith 95% CI")
There were 24 warnings (use warnings() to see them)
```

À vous !

- Calculer et stocker la variance σ des résultats *rawpoll_clinton*, pour les sondages de type $A+^2$, dans une variable *v*.
- Chercher l'utilité de la fonction **sd()** dans **R**.
- Afficher **sd(rowClinA)**, associée aux résultats *rawpoll_clinton*, pour les sondages de type $A+$.
- Comparer cette valeur avec celle de la variance associée aux résultats *rawpoll_clinton* (fonction **var()**).
- dans quels cas la fonction **sd()** peut être utile ?
- Calculer l'intervalle de confiance (risque α de 5%) à partir des résultats *rawpoll_clinton* pour les sondages de type $A+$ (vous utiliserez la fonction **sqrt()** et la fonction **var()**).
- Comparer ce résultat avec celui de la figure 9 qui affiche les résultats *adjpoll_clinton*.
- Formuler une hypothèse sur mauvaises prédictions des sondages lors de la campagne.
- Croyez vous aux sondages ? Pourquoi ?

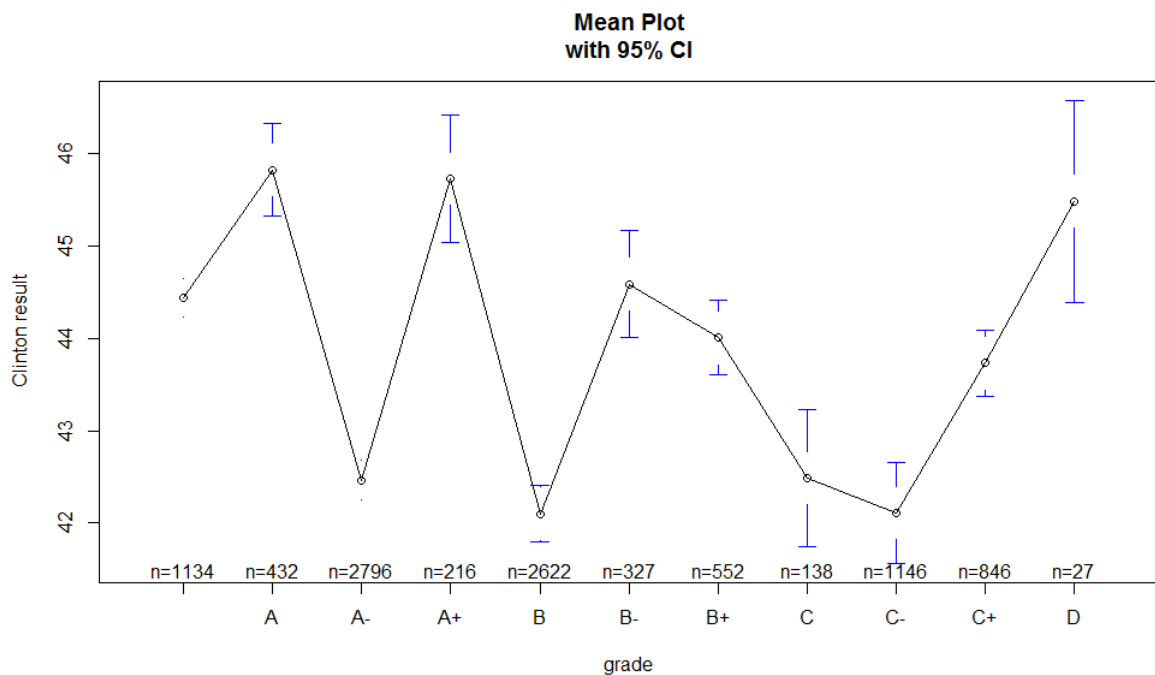


FIGURE 9 – Intervalle de confiance des scores *adjpoll_clinton* pour différents grades de tests

2. vous pouvez stocker ces résultats dans une variable appelée *rowClinA*