BE de Statistiques

Étude d'une base de données IENAC

Groupe 20

Dzieciol Nicolas
Pastre Guillaume
Bonneval Fabien
Buisan Guilhem

Sommaire

Table des matières

Introduction

Dans ce sujet, on se propose d'étudier les caractéristiques des étudiants ingénieur à l'ENAC afin de déterminer s'il existe des relations entre ces variables.

Nous nous intéresserons à un échantillon de 55 étudiants, et considérerons les variables présentées à la page suivante.

Pour effectuer une analyse la plus complète possible, nous regarderons en premier lieu le jeu de données, ensuite nous effectuerons une analyse descriptive des variables les plus parlantes, et enfin nous procéderons à une étude inférentielle pour valider ou invalider nos tests d'hypothèses.

I. Lecture du jeu de données

Question 2.2 : Les natures des différentes variables sont regroupées dans le tableau suivant :

Sexe	Qualitative Nominale
Bac	Qualitative Nominale
Mention	Qualitative Ordinale
Filière	Qualitative Nominale
Note Écrit	Quantitative Continue
Note Orale	Quantitative Continue
Moyenne	Quantitative Continue
Rang	Quantitative Discrète
Voeux	Quantitative Discrète
Concours	Qualitative Nominale
Note Analyse	Quantitative Continue
Note Proba	Quantitative Continue
Succès	Qualitative Nominale

Nature des variables

La variable succès représente l'obtention ou non d'une note supérieure à 10/20 en probabilités.

II. Étude descriptive

II.1. Étude unidimensionnelle

II.1.1 Variables Qualitatives

Question 3.1.1.a : Tables des fréquences

F	M
8	47

Variable "sexe"

A	M	PC	SI	SVT
0	30	3	4	13

Variable "bac"

P	AB	B	TB
0	5	24	21

Variable "mention"

CI	CPP	MP	MP*	PC	PC*	PSI	PSI*
1	1	21	2	7	5	6	9

Variable "filière"

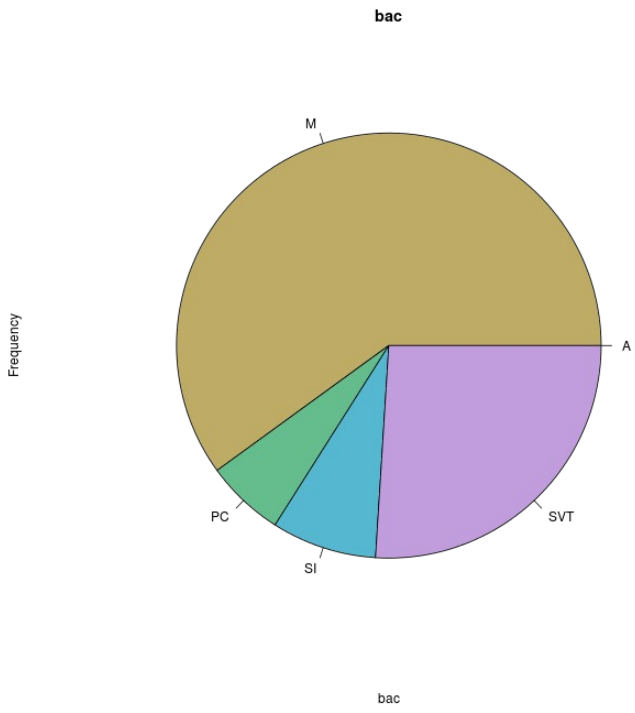
Civil	Fonctionnaire
48	7

Variable "concours"

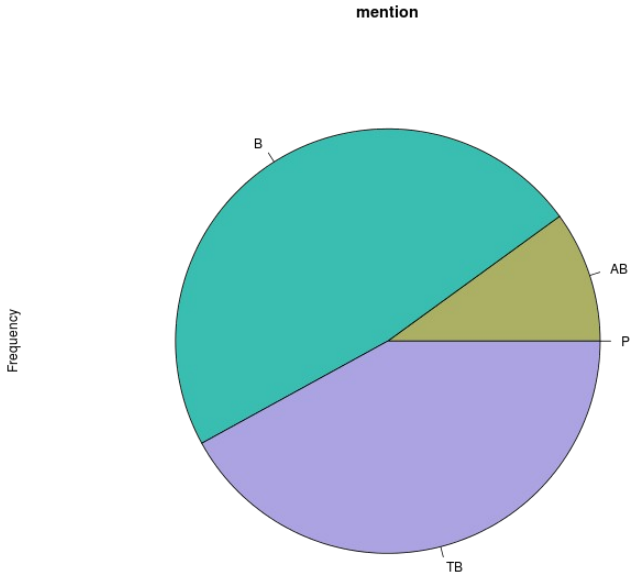
Echec	Succès
16	39

Variable "succès"

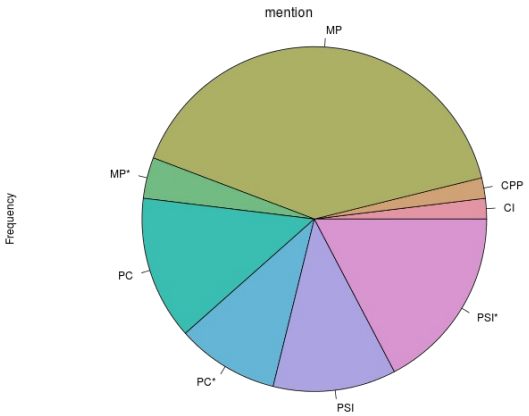
Question 3.1.1.b : Diagrammes



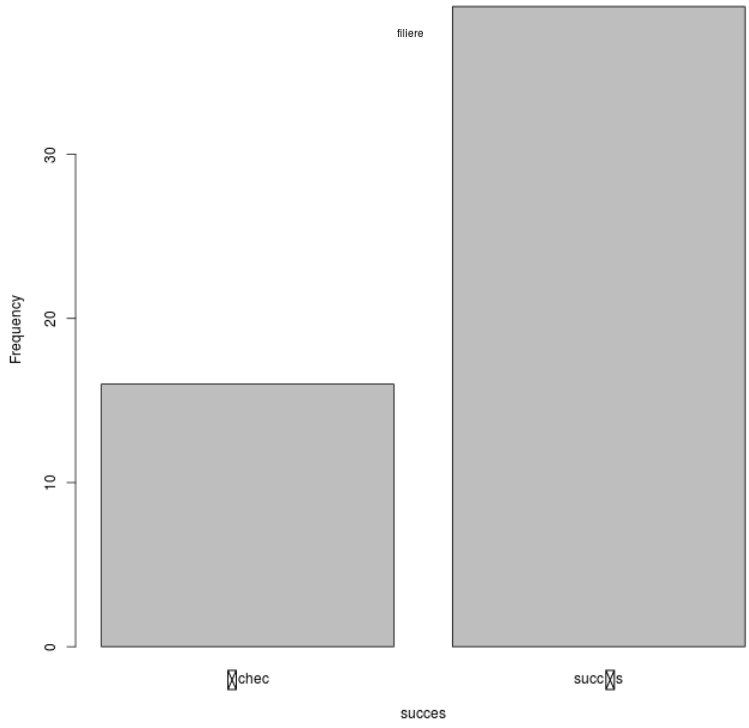
Spécialité au bac des étudiants : une majorité a fait spé maths



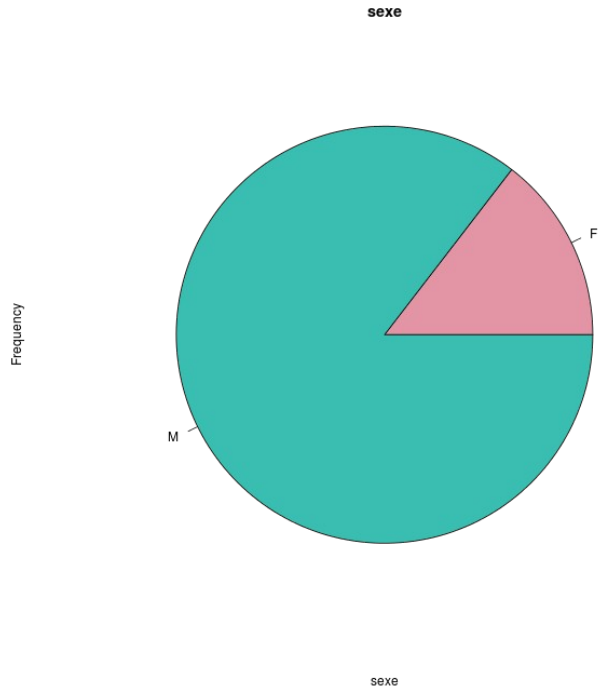
Mention au bac des étudiants : Peu ont eu AB comparé au nombre de B et TB



Provenance des étudiants de prépa : Une majorité vient de MP



Proportion des étudiants qui ont validé leur examen en probabilités ou non : environ 16 % ne l'ont pas validé



Proportion du sexe des étudiants : Bien plus de garçons sont présents dans l'échantillon que de filles

Question 3.1.1.c : Les graphes permettent de visualiser rapidement les proportions pour les variables qualitatives, mais, dans le cas du circulaire, ne présente pas les chiffres et est donc imprécis. On voit par exemple qu'il y a bien plus de garçons que de filles, que la majorité viennent d'un bac M avec pour la quasi totalité une mention B ou TB.

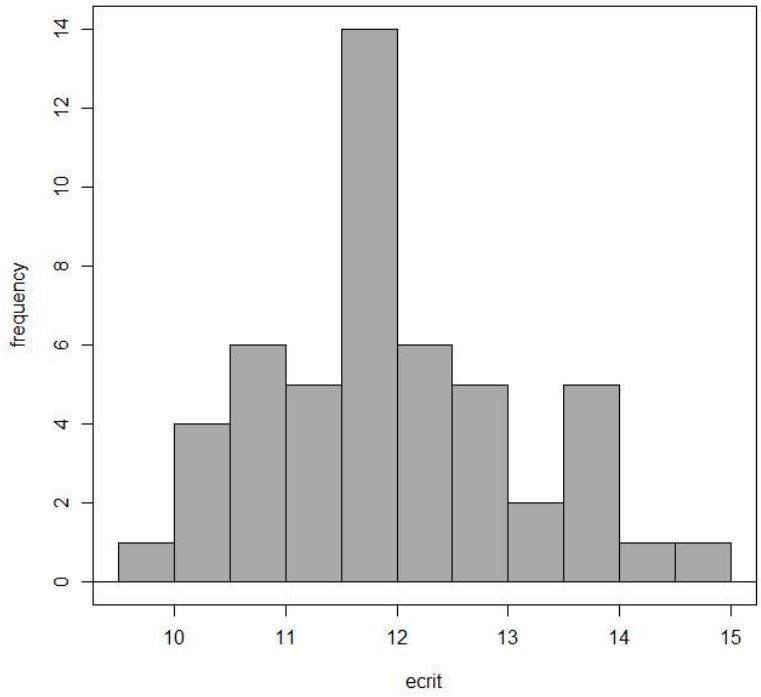
II.1.2 Variables Quantitatives

Question 3.1.2.a Résumés numériques

Variable	Moyenne	Ecart-Type	Ecart Interquartile (Q3-Q1)	Minimum	Q1	Mediane	Q3	Maximum	n	NA
Analyse	14.78182	2.6260159	3.250	7.00	13.5	15.0	16.750	19.00	55	0
Ecrit	11.9638	1.1181105	1.35	9.95	11.3325	11.865	12.6825	14.69	50	5
Moyenne	12.8834	0.8080099	0.700	11.91	12.38	12.68	13.08	15.08		5

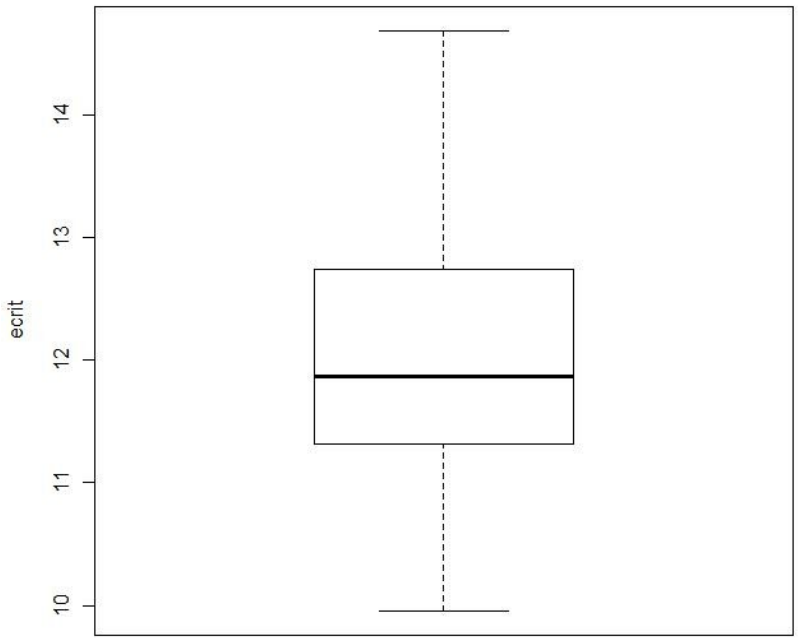
Oral	14.3428	1.5271917	2.225	10.99	13.4275	14.125	15.6525	17.35		5
Proba	11.27636	2.9688796	4.250	4.10	9.15	11.55	13.4	17.00		0
Rang	788.48	350.65002	482.0	48	595.25	788.5	1077.25	1320		5
Voeux	3.28	2.2227029	2.0	1	2	3	4	13	50	5

Question 3.1.2.b Diagrammes



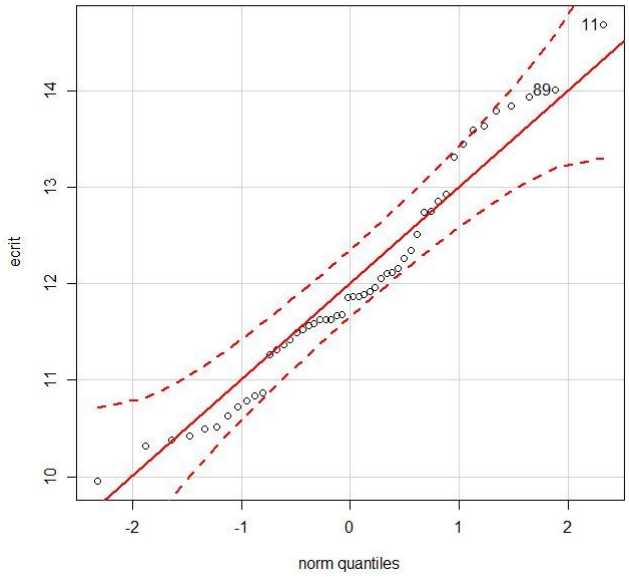
Ecrit : Histogramme

On observe les proportions des étudiants suivant la fourchette de leur note à l'écrit.



Ecrit : Dispersion

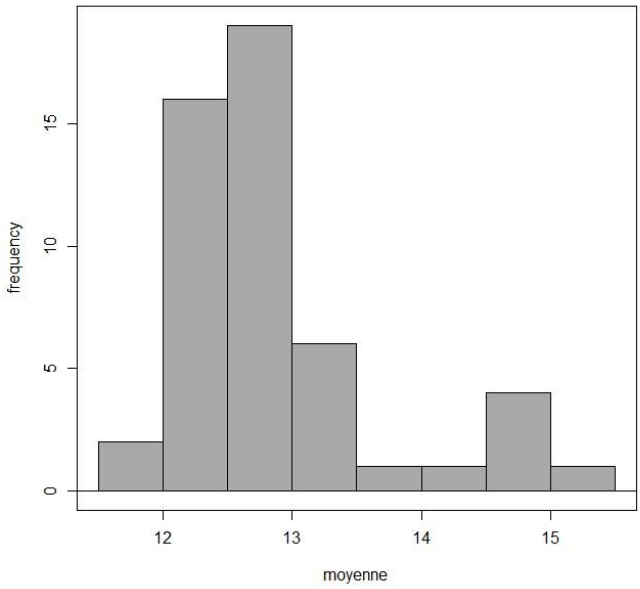
Le diagramme de dispersion de l'écrit : on observe un minimun d'environ 10, une moyenne d'un peu moins de 12, d'un maximum vers 14,5, et observons de même les quartiles.



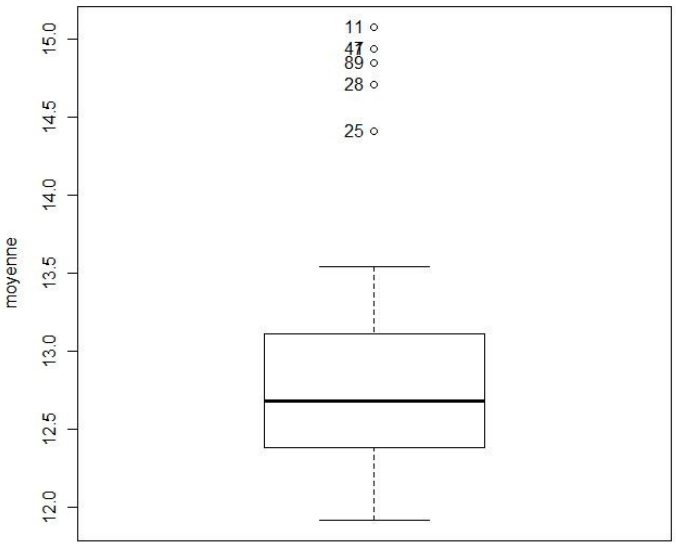
Le diagramme quantile-quantile de l'écrit, pour voir les écarts par rapport aux résultats que l'on observerait si la distribution suivait une loi normale.

Ecrit : quantile-quantile

L'histogramme de répartition de la variable moyenne : une grande majorité des étudiants voient leur moyenne entre 12 et 13

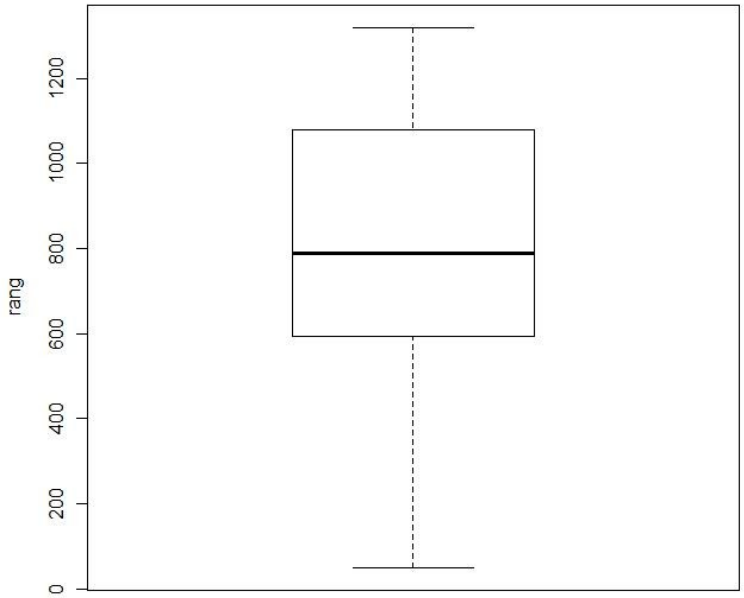


Moyenne : histogramme



Les paramètres statistiques de la variable moyenne

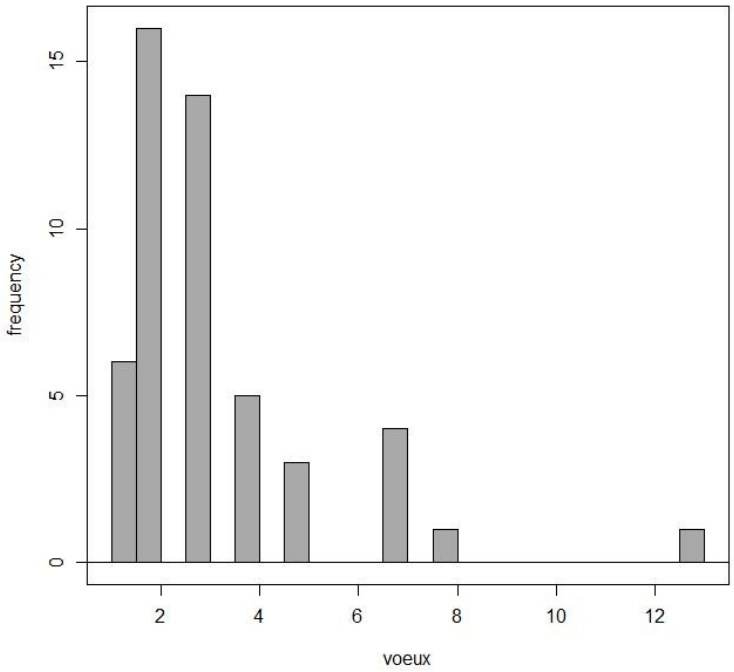
Moyenne : dispersion



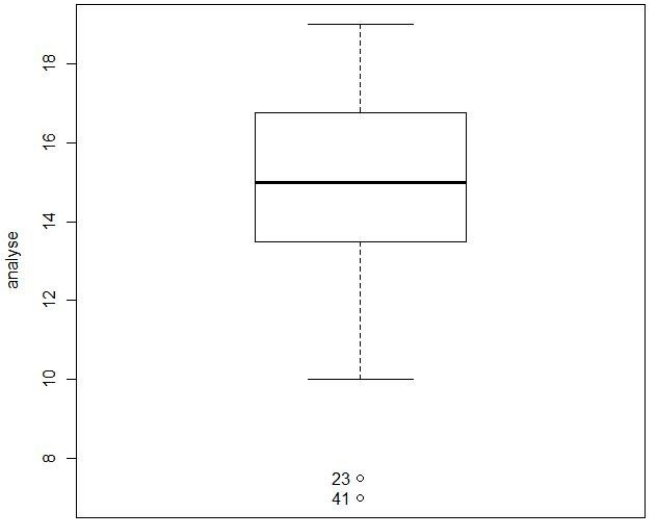
L'étalement des rangs aux concours des étudiants

Rang : dispersion

Les répartitions de l'obtention de leur premier/deuxième vœux sur le service scei



Voeux : en bâtons



La dispersion des notes en analyse

Analyse : Dispersion

Question 3.1.2.c. Ces diagrammes pour les variables quantitatives : histogrammes pour variables continues, bâtons pour variables discrètes, permettent d'avoir un aperçu de l'ensemble des informations.

Mais parfois un manque de sens se fait sentir : une observation des rangs alors que les étudiants ne viennent pas de la même filière. On voit tout de même deux groupes, probablement les fonctionnaires et les civils.

Les diagrammes de dispersions sont utiles pour voir la répartition autour de la médiane, l'homogénéité, ce qui est particulièrement utile pour les notes.

L'étude unidimensionnelle des variables permet d'en apprendre plus sur elles, voir les répartitions par exemple permet de se faire une première idée sur leurs contenus, mais en revanche ne permet pas de repérer et mettre en évidence les éventuels lien entre les variables.

Il faut donc faire appel à l'étude bidimensionnelle pour étudier par exemple s'il peut y avoir corrélation entre notes d'analyse et de proba ou encore mettre en exergue certaines tendances.

II.2 Étude bidimensionnelle

II.2.1 Croisement entre une variable quantitative et une variable qualitative

Question 3.2.1.1.a

	Analyse					Proba				
bac	A	M	PC	SI	SVT	A	M	PC	SI	SVT
	NA	15.00000	14.66667	14.50000	14.76923	NA	11.48333	11.18333	12.88750	10.67692
concours	Civil		Fonctionnaire			Civil		Fonctionnaire		
	15.78571		14.63542			11.01771		13.05000		
filière	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*	CI	CPP	MP	MP*	PC	PC *	PSI	PSI*
	18.0	7.50	14.904	18.250	14.642	15.200	13.750	14.833	15.3	8.3	11.507	14.175	10.242	11.880	11.908	10.650
mention	P	AB	B	TB					P	AB	B	TB				
	NA	15.20000	14.27083	15.50000					NA	10.62000	11.00625	11.95952				
sexe	F		M						F		M					
	14.8125		14.7766						13.01250		10.98085					
succès	Échec		Succès						Échec		Succès					
	13.00000		15.51282						7.565625		12.798718					

Résumé numérique entre variables qualitative et quantitatives (moyenne)

	Analyse					Proba				
bac	A	M	PC	SI	SVT	A	M	PC	SI	SVT
	NA	15.00	14.50	15.25	16.00	NA	11.875	10.350	12.150	10.800
concours	Civil				Fonctionnaire	Civil				Fonctionnaire
	14.5				17.0	11.5				13.85
filière	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*	CI	CPP	MP	MP*	PC	PC *	PSI	PSI*
	18.00	7.50	14.50	18.25	14.50	15.00	13.25	16.00	15.300	8.300	11.450	14.175	10.350	12.350	13.875	10.800
mention	P	AB	B	TB		P	AB	B	TB	
	NA	16.5	14.5	16.0		NA	11.00	11.25	12.00	
sexe	F			M		F			M	
	15.5			15.0		13.00			11.35	
succès	Échec			Succès		Échec			Succès	
	13.75			16.00		8.125			12.550	

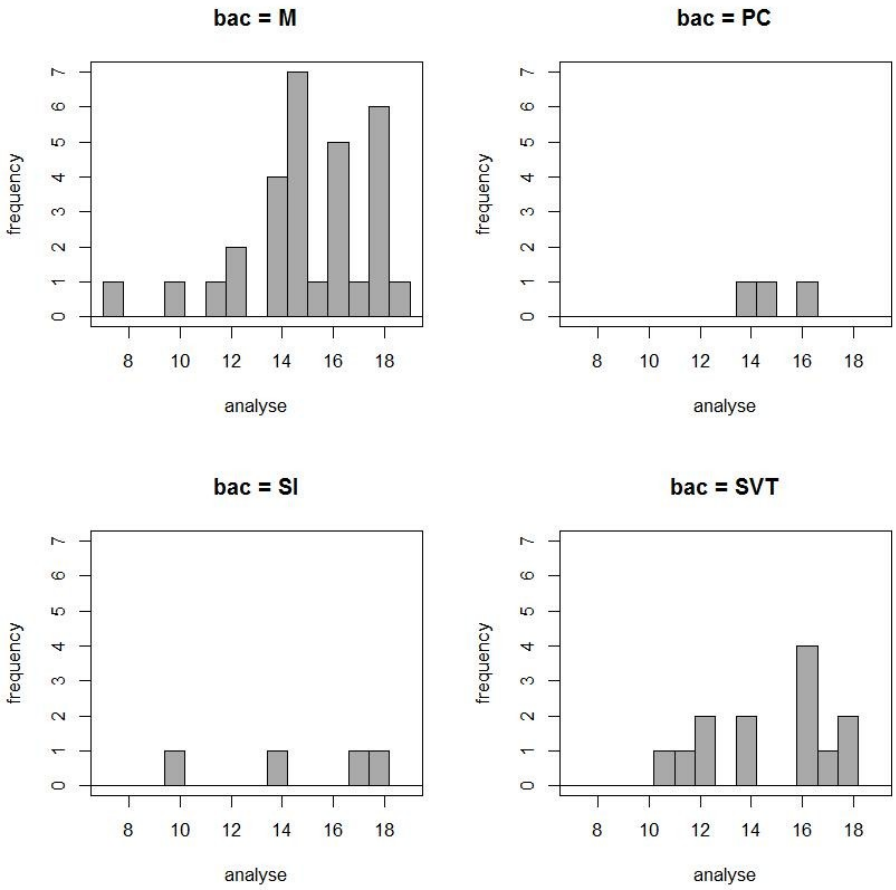
Résumé numérique entre variables qualitative et quantitatives (médiane)

	Analyse					Proba				
bac	A	M	PC	SI	SVT	A	M	PC	SI	SVT
	NA	2.593094	1.258306	3.488075	2.429124	NA	2.913652	1.985783	3.200879	2.961321
concours	Civil				Fonctionnaire	Civil				Fonctionnaire
	2.371685				4.081025	2.960098				2.546730
filière	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*	CI	CPP	MP	MP*	PC	PC *	PSI	PSI*
	NA	NA	2.457577	1.060606	1.858419	2.165965	3.024124	2.969261	NA	NA	2.42844	1.66170	2.2687	2.6061	4.8411	3.1089
mention	P	AB	B	TB		P	AB	B	TB	
	NA	2.863564	2.698547	2.109502		NA	2.817490	2.803504	3.001525	
sexe	F			M		F			M	
	2.344256			2.694214		1.700158			3.048584	
succès	Échec			Succès		Échec			Succès	
	3.011091			2.082152		1.721019			1.783273	

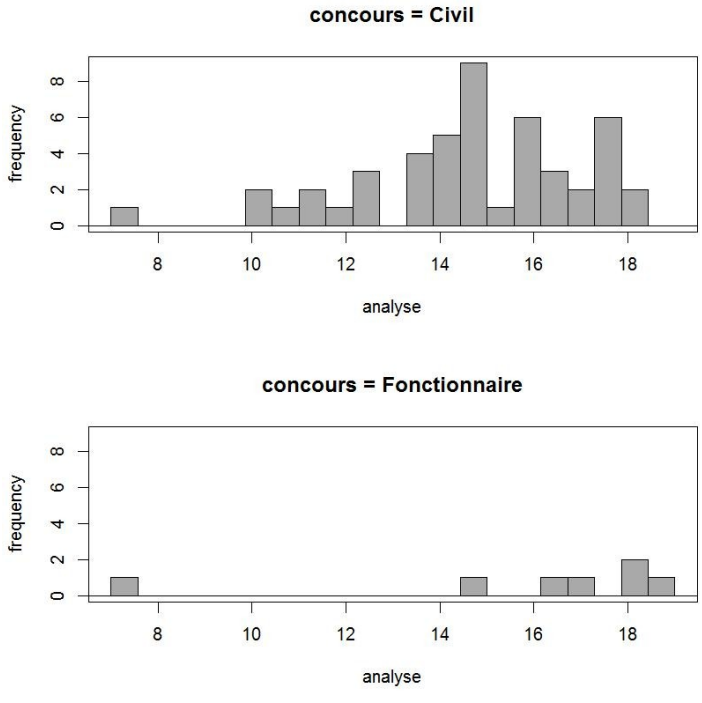
Résumé numérique entre variables qualitative et quantitatives (écart type)

Question 3.2.1.1.b

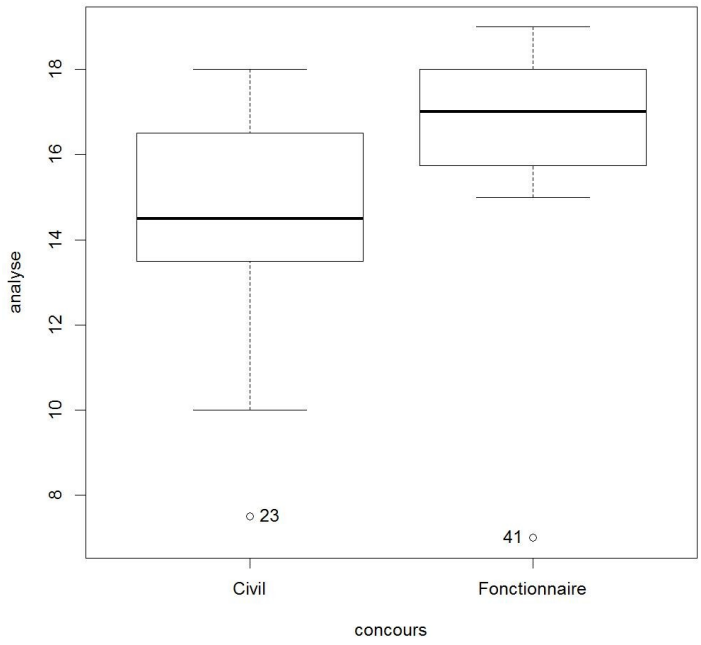
Variable *analyse :*



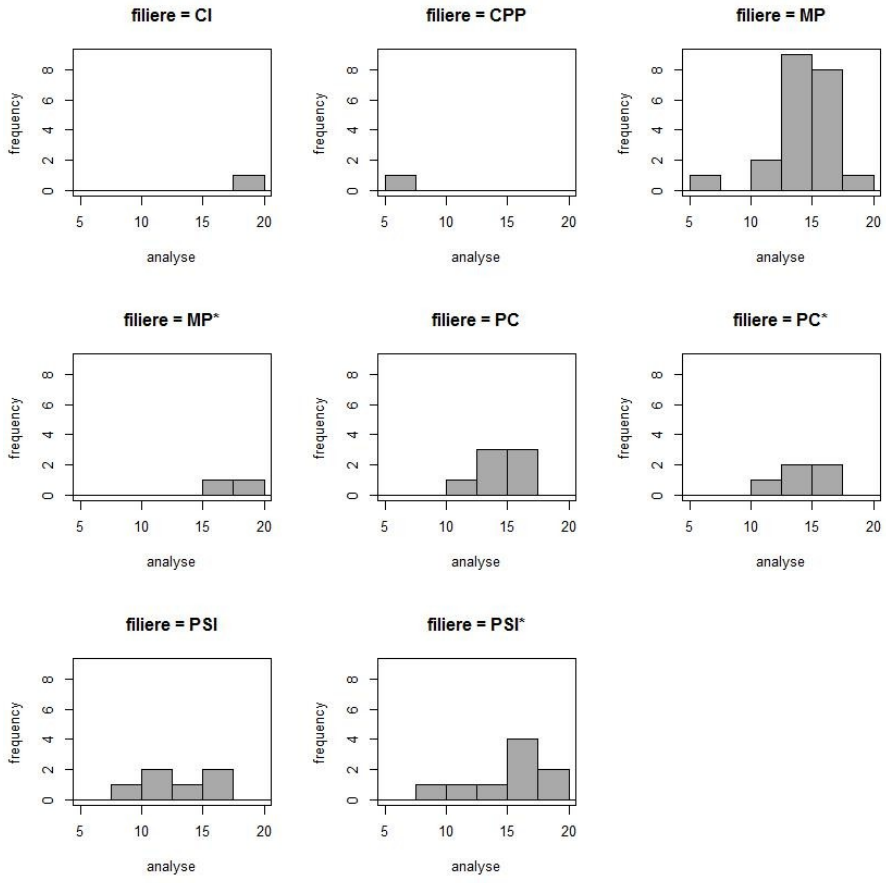
Analyse-bac : histogrammes



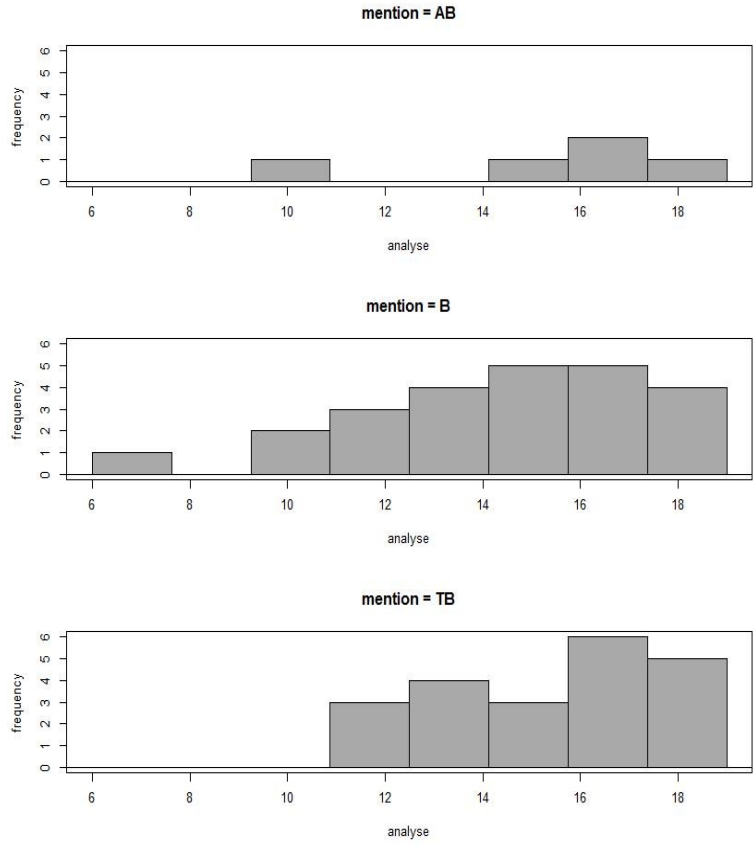
Analyse-concours : histogrammes



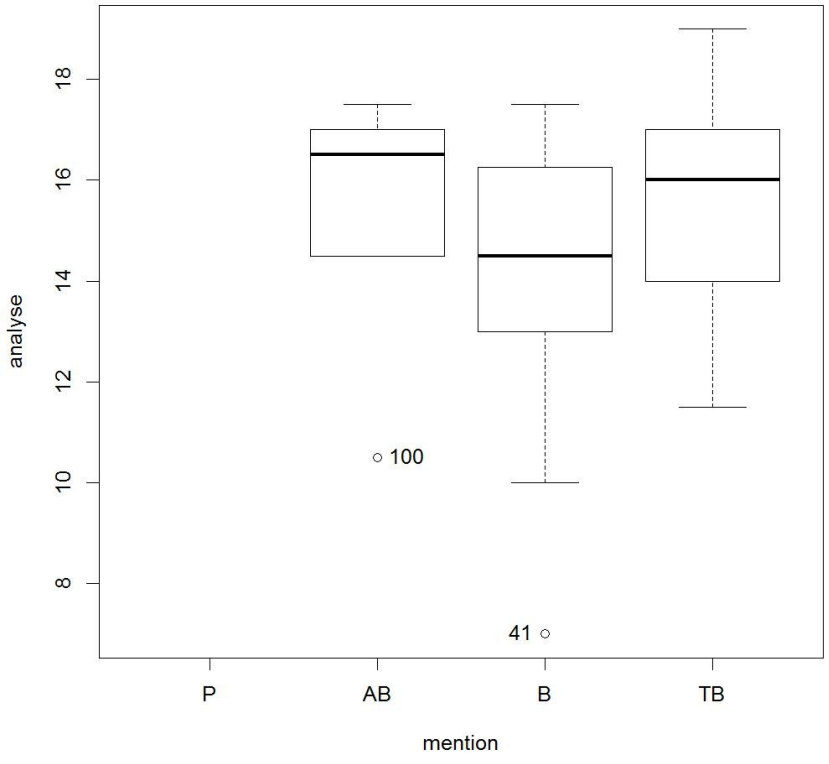
Analyse-concours : dispertion



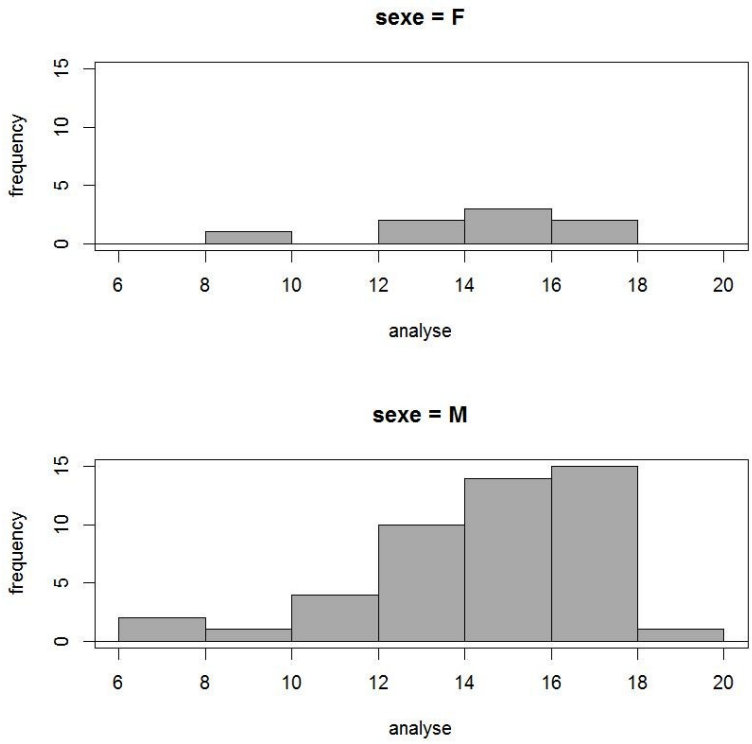
Analyse-filiere : histogramme



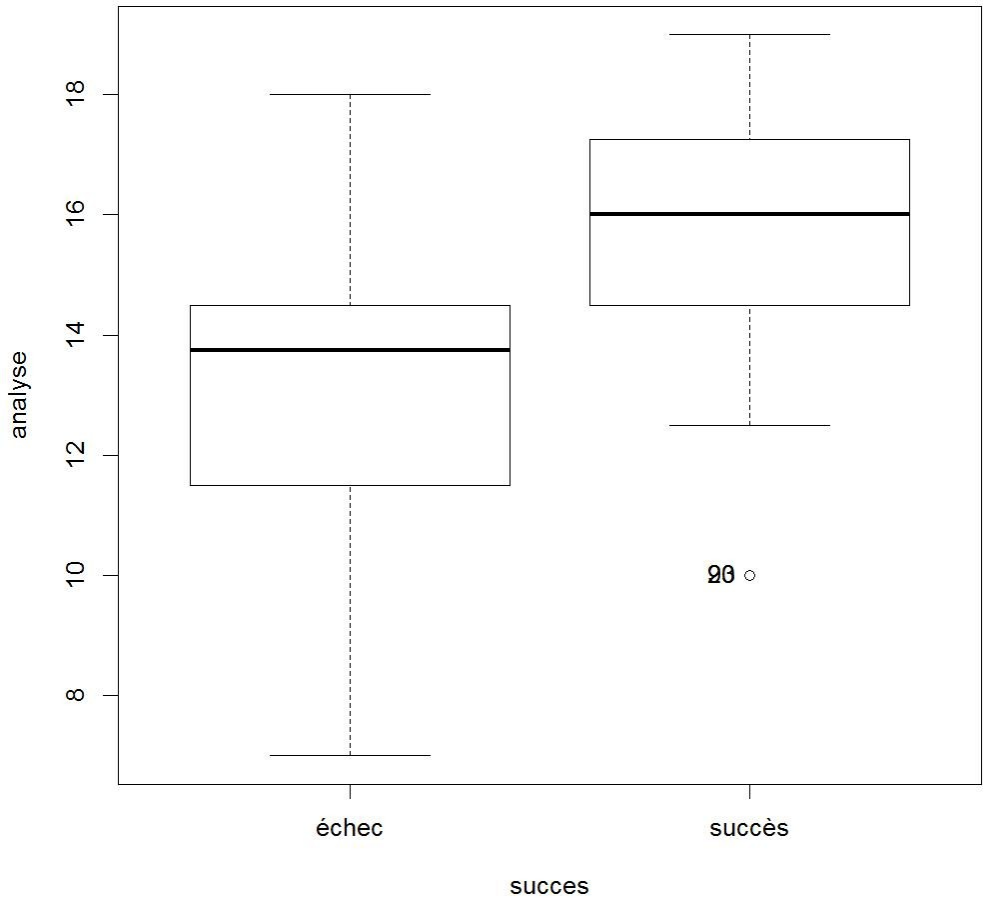
Analyse-mention : histogrammes



Analyse-mention : dispertion

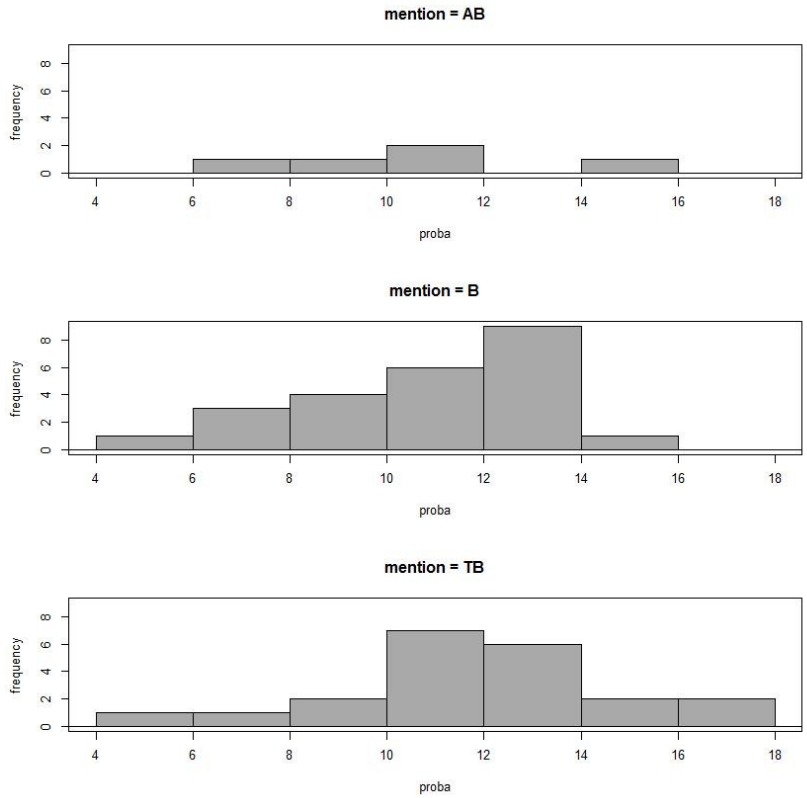


Analyse-sexe : histogrammes

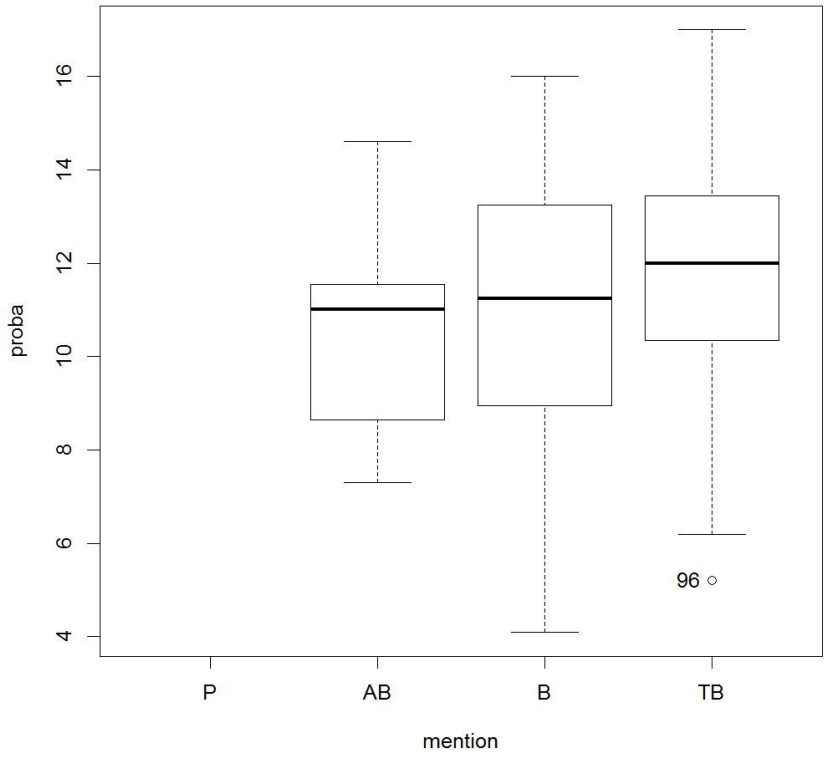


Analyse-succes : dispertion

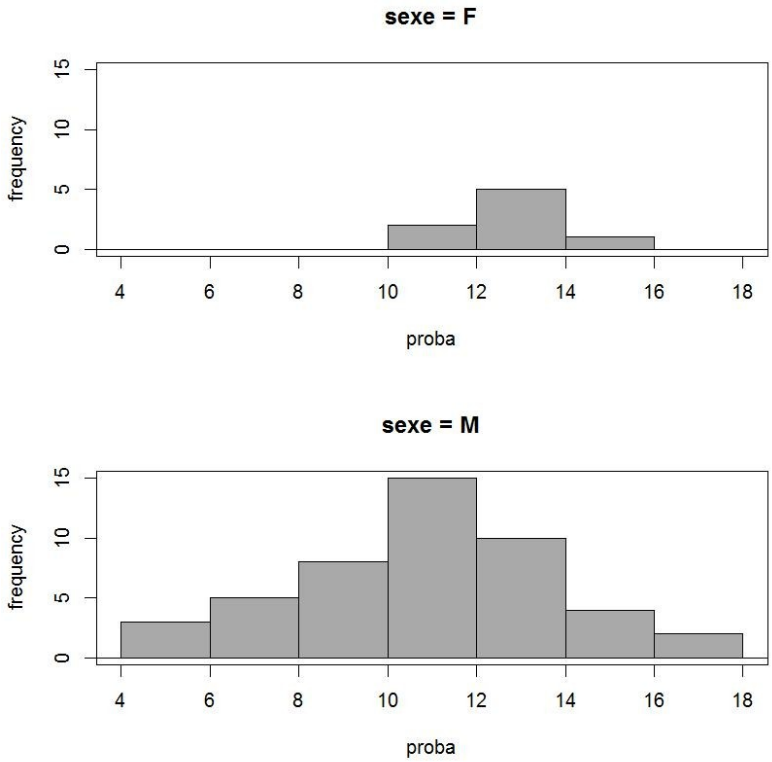
Variable *proba* :



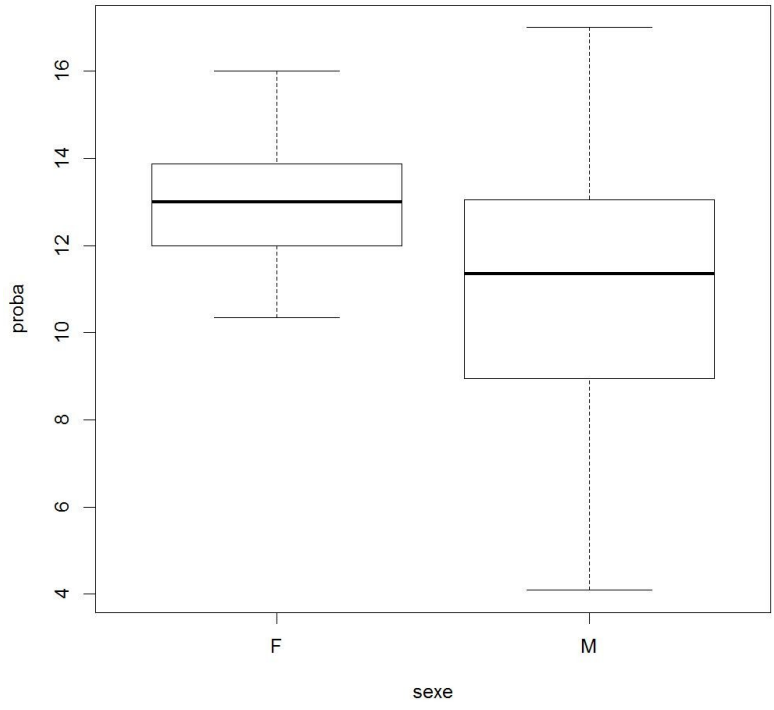
Proba-mention : histogrammes



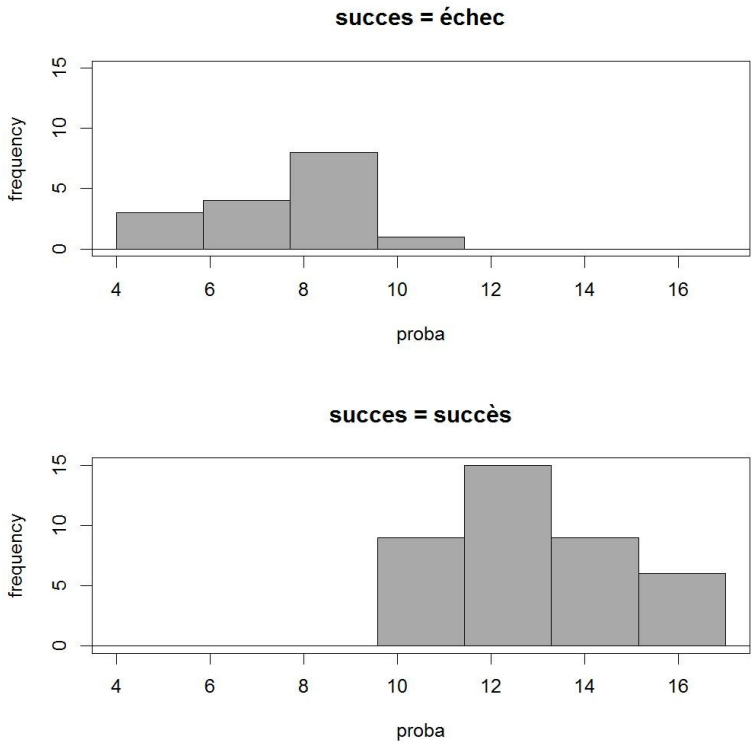
Proba-mention : dispertion



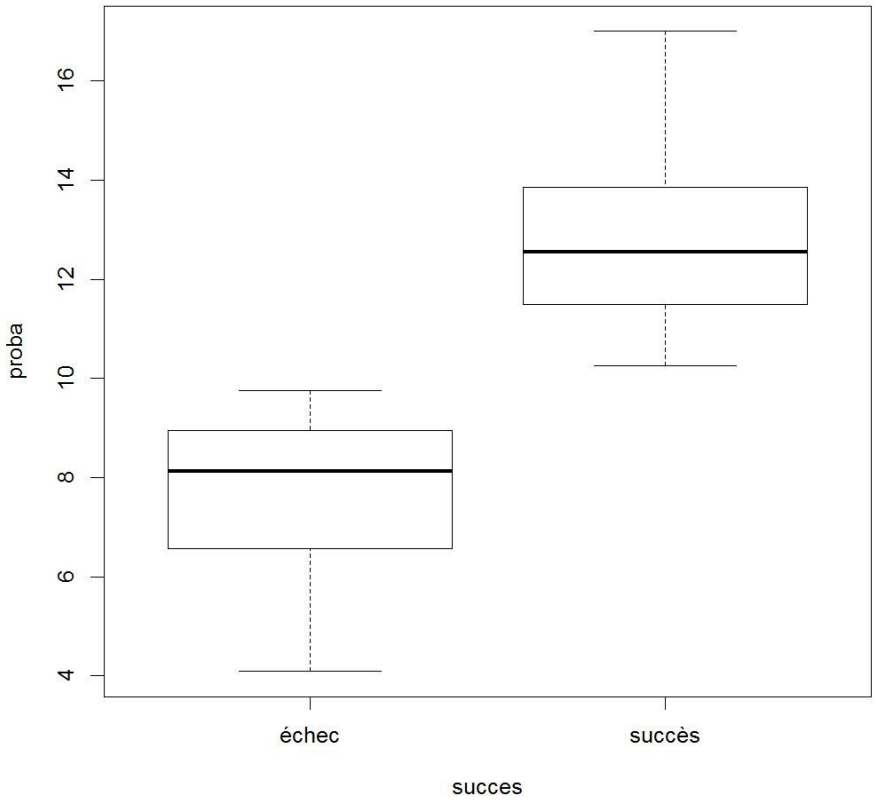
Proba-sexe : histogrammes



Proba-sexe : dispertion



Proba-succes : histogrammes



Proba-succes : dispertion

Question 3.2.1.1.c. L'analyse bidimensionnelle permet d'analyser les données différemment pour prendre en compte plus de paramètres, ainsi on observe l'influence de certaines variables qualitatives sur les notes d'analyses et de proba.

On observe les proportions d'étudiants en fonction de leur note en analyse, triés par succès à l'examen de probabilités . Une majorité d'étudiants ont une note plus haute dans la partie succès, et les notes les plus faibles dans la partie échec.

On peut aussi se demander s'il n'y pas une corrélation entre les notes d'analyse et celles de probabilités.

Question 3.2.1.3

Variable *rang* croisée avec *filiere*

Filière	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*
rang	NaN	NaN	1074.9048	688.5000	558.2857	617.0000	640.8333	515.1111

Résumé numérique entre les variables rang et filière (moyenne)

Filière	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*
rang	NA	NA	1104.0	688.5	644.0	757.0	631.5	473.0

Résumé numérique entre les variables rang et filière (médiane)

Filière	CI	CPP	MP	MP*	PC	PC*	PSI
rang	NA	NA	262.79077	590.43416	281.03067	307.66459	63.26584

Résumé numérique entre les variables rang et filière (écart type)

Ce graphe permet de résoudre visuellement le problème soulevé précédemment, celui du rang qui avait un sens différent suivant la filière. Désormais la distinction est marquée et l'analyse des données peut se faire de façon plus complète. Ainsi, de par la forte dépendance avec la variable *filiere*, la variable *rang* est très difficilement utilisable sans prendre en compte la filière.

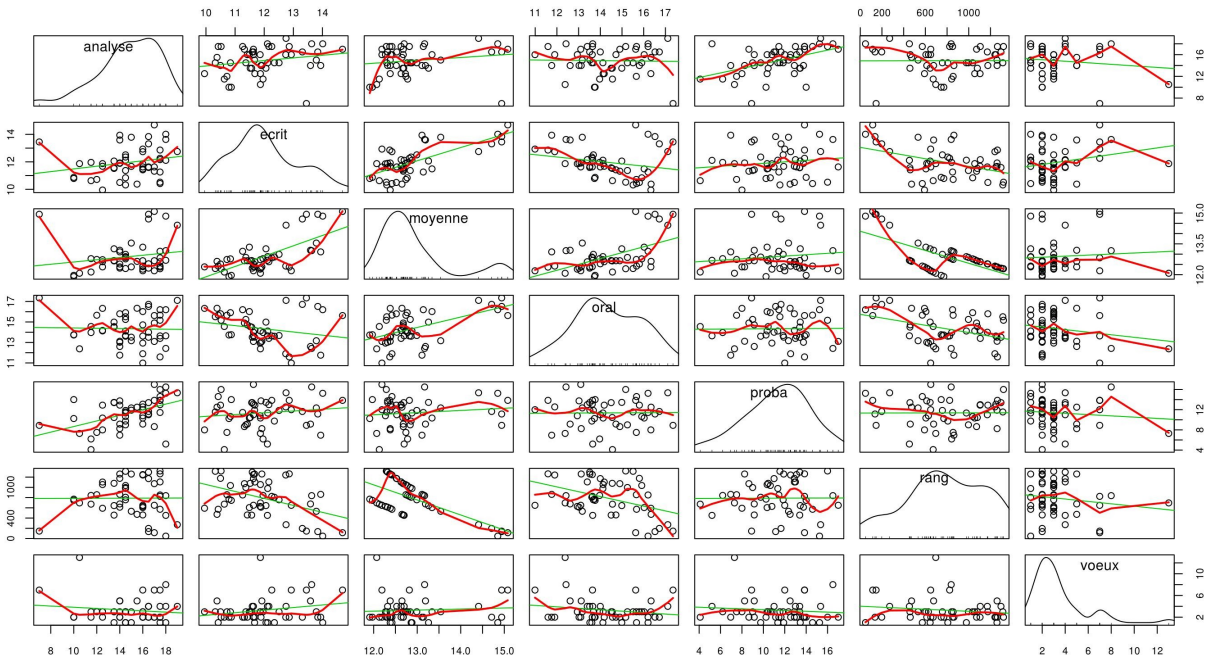
II.2.2 Croisement entre variables quantitatives

Question 3.2.2.1

	Analyse	Ecrit	Moyenne	Oral	Proba	Rang	Voeux
Analyse	1	0.2204279	0.16938465	-0.02439720	0.482642513	0.0044748	-0.125937
Ecrit		1	0.69603362	-0.22737578	0.137168956	-0.4337386	0.2390189
Moyenne			1	0.53989504	0.120149583	-0.6776006	0.0699118
Oral				1	0.011427337	-0.4085150	-0.182343
Proba					1	0.00594790	-0.097613
Rang						1	-0.148303
Voeux							1

Matrice de corrélation

Question 3.2.2.2



Nuages de points

Question 3.2.2.3 La matrice de corrélation permet d'observer les coefficients de corrélation entre deux variables deux à deux. On observe un fort taux de corrélation entre la *moyenne* et l'*écrit* (p=0,70), un taux élevé entre moyenne et oral (p=0,54) et pour confirmer l'intuition de la corrélation entre *analyse* et *proba*, on a : (p=0,48) ce qui dénote un certain lien entre les deux variables, qui n'est pas toujours respecté pour tous les résultats.

Les graphes en nuages de points permettent d'observer des relations existantes entre les variables, et peuvent apporter plus d'informations que les coefficients de corrélation de la matrice : on peut observer quatre populations différentes dans la superposition de *rang* et de *moyenne* (les quatre principales filières) avec une relation très linéaire (le *rang* augmente quand la *moyenne* diminue).

II.2.3 Croisement entre variables qualitatives

Question 3.2.3.1.a)

Succes/bac	A	M	PC	SI	SVT
échec	0	8	1	0	5
succès	0	22	2	4	8

Succes-Bac : table de contingence

Succes/concours	Civil	Fonctionnaire
Échec	15	1
Succès	33	6

Succes-concours : table de contingence

Succes/filiere	CI	CPP	MP	MP*	PC	PC*	PSI	PSI*
Échec	0	1	5	0	3	1	2	3
Succès	1	0	16	2	4	4	4	6

Succes-filiere : table de contingence

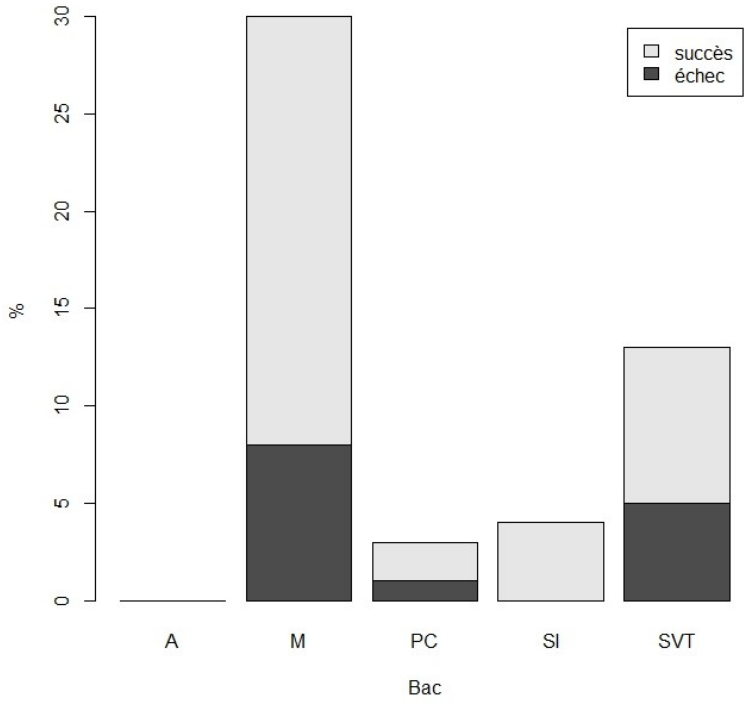
Succes/mention	P	AB	B	TB
Échec	0	2	8	4
Succès	0	3	16	17

Succes-mention : table de contingence

Succes/sexe	F	M
Échec	0	16
Succès	8	31

Succes-sexe : table de contingence

Question 3.2.3.1.b)



Succes-bac : profil colonnes

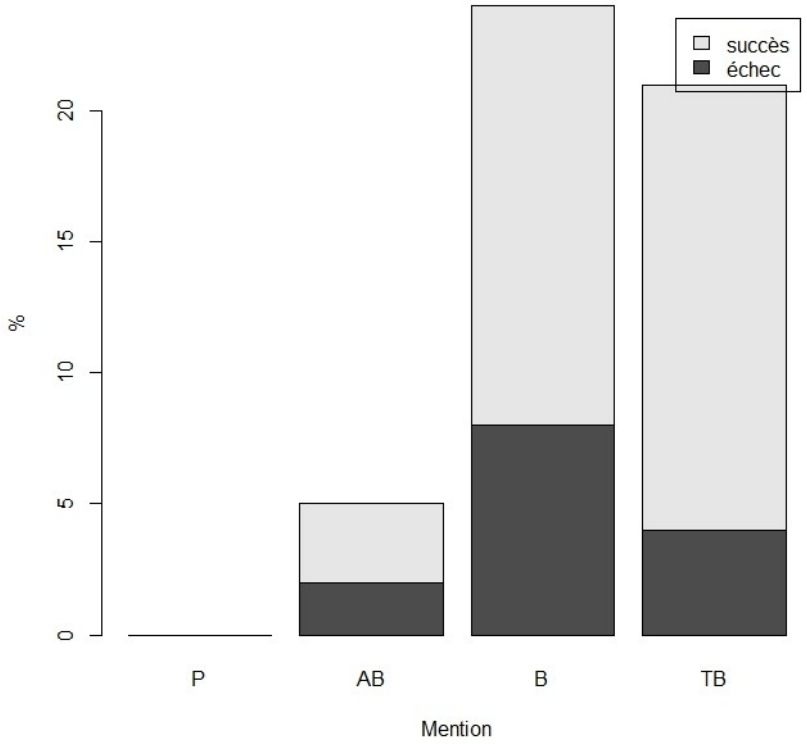
Étonnamment, un fort pourcentage des étudiants ayant fait spé maths au lycée n'ont pas validé leur examen de probabilités

Succes/bac	A	M	PC	SI	SVT
échec	NaN	26.7	33.3	0	38.5
succès	NaN	73.3	66.7	100	61.5

Succes-bac : table de contingence (pourcents)

Succes/concours	Civil	Fonctionnaire
Échec	31.2	14.3
Succès	68.8	85.7

Succes-concours : table de contingence (pourcents)

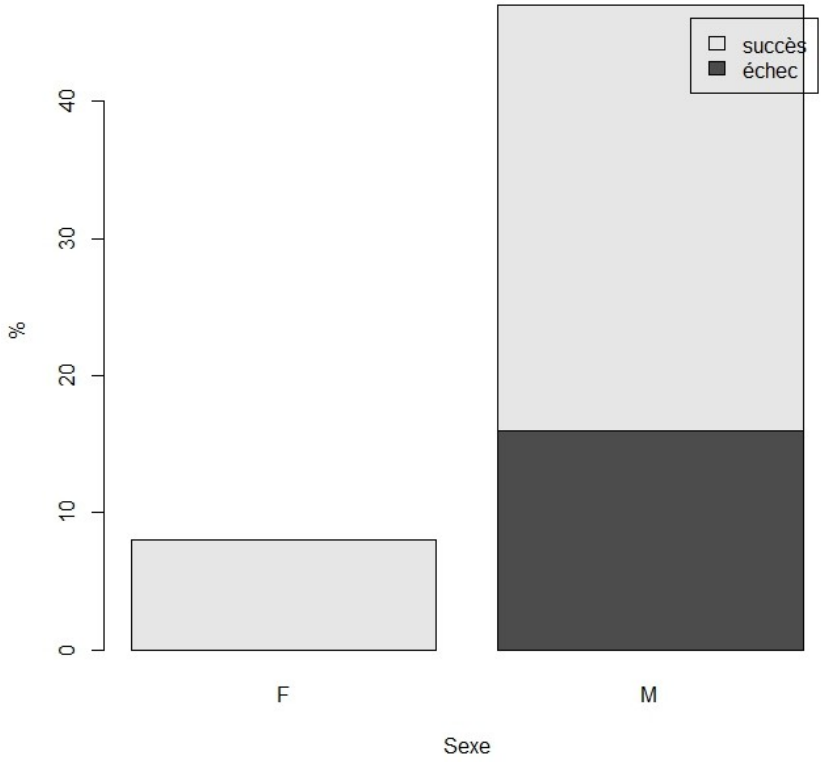


Succes-mention : profil colonne

Succes/mention	P	AB	B	TB
Échec	NaN	40	33.3	19
Succès	NaN	60	66.7	81

Succes-mention : table de contingence (pourcents)

On retrouve un pourcentage de succès qui augmente avec le prestige de la mention au bac



Succes-sexe : profil colonnes

Succes/sexe	F	M
Échec	0	34
Succès	100	66

Succes-sexe : table de contingence (pourcents)

Question 3.2.3.2

Croiser la variable *succes* avec les autres variables qualitatives nous a permis d'observer les proportions d'étudiants ayant réussi ou échoué à l'examen de probabilités suivant les critères des autres variables, comme le sexe, la filière en classes prépa, la mention au bac...

On peut alors observer des liens logiques : en proportion les étudiants qui ont eu mention TB au bac ont plus réussi que ceux qui ont eu B à 81% de succès contre 66.7%

On retrouve les résultats dans les diagrammes en barre pour une visualisation facile.

Il y a tout de même un effet taille très important: on pourrait comprendre que les étudiants de la filière CI réussissent énormément bien (100% de réussite) alors que ceux de la filière CPP ont de grosses difficultés (100 % d'échec). Mais l'échantillon sur ces deux filières est de taille 1 ! Ce qui n'est pas suffisant pour se ramener à la population.

Ainsi, l'étude descriptive bidimensionelle noua a permis dans un premier temps d'observer les moyennes, variances, et autre indicateurs statistiques des variables quantitatives en les croisant avec les variables qualitatives.

Ensuite nous avons pu tester la corrélation entre variables quantitatives (rand, moyenne, ...)

Enfin, nous avons pu observer les proportions des étudiants entre variables qualitatives (succès, ...) ce qui permet de trouver des tendances.

Il faut effectuer des tests afin de confirmer ou d'infirmer les conjectures émises précédemment.

III. Étude inférentielle

III.1. Tests d'hypothèses pour un échantillon

Question 4.1.1

Test d'adéquation à une loi normale pour la variable *analyse* :

Hypothèses maintenues : $(X_1,\ldots,X_n)$ indépendant identiquement distribué (iid) de fonction de répartition F inconnue.

En effet, la note obtenue par un étudiant i n'influe normalement pas sur la note obtenue par un autre étudiant (sauf éventuelle triche). De plus, on admet que toutes les notes suivent la même loi.

Test de Shapiro-Wilk : **H0 : F = Normale** contre **H1 : F ≠ Normale**

On a : W = 0.933, p-value = 0.004373

Pour un risque de 5%, alpha > p-value donc **on rejette la normalité avec un risque de 5%**

Test d'adéquation à une loi normale pour la variable *proba :*

Hypothèses maintenues : $(X_1,\ldots,X_n)$ iid de fonction de répartition F inconnue.

Test de Shapiro-Wilk : **H0 : F = Normale** contre **H1 : F ≠ Normale**

On a : W = 0.9804, p-value = 0.5065

Pour un risque de 5 %, alpha < p-value donc on ne rejette pas la normalité avec un risque bêta non contrôlé.

Si la distribution des notes d'*analyses* ne ressemble pas à une loi normale, on voit que le test de Shapiro-Wilk est incapable de rejeter la loi normale pour la variable *proba*. Pourtant, ces deux variables représentent des notes obtenues dans deux matières différentes, on aurait pu s'attendre à ce qu'elles suivent la même loi.

Question 4.1.2.a

Hypothèses maintenues : $(X_1,\ldots,X_n)$ iid, de loi normale par test de Shapiro-Wilk $N(\mu,\sigma^2)$, variance σ^2 inconnue.

H0 : μ > 12 contre **H1 : μ ≤ 12**

Test paramétrique de Student pour une moyenne (t-test univarié)

Question 4.1.2.b

Valeur de la statistique observée : 11.27636

Région critique : Wα = $\left\{\sqrt{n}\,\frac{\bar{x}_n - 12}{s'_n} < -t_{n-1;2\alpha}\right\}$

On a : $W_{0,05}$ = $\{-\infty\,;11.94633\}$

Alors, 11.27636 ϵ $W_{0,05}$, donc on rejette H0 avec un risque controlé de 5 %. La moyenne est inférieure à 12.

Question 4.1.2.c

On a la p-valeur α_c = 0.03812. Alors α = 0,05 > α_c. Par la p-valeur on rejette aussi H0 avec un risque contrôlé de 5 %.

Question 4.1.2.d

Le test nous a permis de confirmer à 95% que la moyenne de proba était inférieure à 12 sur toute la promo grâce à cet échantillon seulement. En revanche, l'absence de loi normale par Shapiro-Wilk laisse douter de la justesse de la conclusion du test (bien qu'on n'ait pas coché « loi normale »).

III.2. Tests d'hypothèses pour deux échantillons

Question 4.2.1.a

Pour répondre à la question « peut-on affirmer affirmer que les femmes réussissent mieux que les hommes en probabilités ?», il faut effectuer un **test de Student de comparaison des moyennes**.

Question 4.2.1.b

Pour répondre à la question « peut-on dire qu'il existe une différence entre les moyennes en analyse et en probabilités ? », il faut effectuer un **test de comparaison de deux moyennes-observations appariées.**

Question 4.2.1.c

Pour répondre à la question «peut-on dire que la variable *succes* dépend des autres variables qualitatives ? », il faut effectuer un **test d'indépendance du Khi-deux.**

Question 4.2.2.a

Pour la question « peut-on affirmer affirmer que les femmes réussissent mieux que les hommes en probabilités ?», le test de Student de comparaison des moyennes requiert les hypothèses maintenues :

- $(X_1, \ldots, X_{n1})$ i.i.d de loi normale $N(\mu_1, \sigma_1^2)$, μ_1 et σ_1^2 inconnues. (Les notes obtenues par les femmes en probabilités).
- $(Y_1, \ldots, Y_{n2})$ i.i.d de loi normale $N(\mu_2, \sigma_2^2)$, μ_2 et σ_2^2 inconnues. (Les notes obtenues par les hommes en probabilités).
- Variances supposées égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$**).** (Pas obligatoirement avec le test proposé dans R)
- $(X_1, \ldots, X_{n1})$ et $(Y_1, \ldots, Y_{n2})$ deux échantillons indépendants. (Les notes des femmes n'ont aucune raison d'être influencées ou d'influencer celles des hommes).

Hypothèses de test : **H0 : $\mu_1 - \mu_2 \leq 0$ contre H1 : $\mu_1 - \mu_2 > 0$**

Pour la question « peut-on dire qu'il existe une différence entre les moyennes en analyse et en probabilités ? »,le test de comparaison de deux moyennes-observations appariées requiert les hypothèses maintenues :

- $(X_1, \ldots, X_n)$ i.i.d de loi X, de moyenne μ_1 inconnue. (Les notes obtenues en analyse, indépendantes d'un individu à l'autre sauf cas de triche).
- $(Y_1, \ldots, Y_n)$ i.i.d de loi Y, de moyenne μ_2 inconnue. (Les notes obtenues en probabilités, indépendantes d'un individu à l'autre sauf cas de triche).
- $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ appariées. (En effet, les notes ont été obtenues par le même individu et ne sont donc pas indépendantes.)
- $(D_1, \ldots, D_n)$ échantillon i.i.d tel que $D_i = X_i - Y_i$ suive une loi normale $N(\mu_D, \sigma_D^2)$, $\mu_D = \mu_1 - \mu_2$ et σ_D^2 inconnues.

Hypothèses de test : **H0 : $\mu_D = 0$ contre H1 : $\mu_D \neq 0$**

Pour la question «peut-on dire que la variable *succes* dépend des autres variables qualitatives ? »,le test d'indépendance du Khi-deux requiert les hypothèses maintenues :

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ échantillon i.i.d, issu du couple (X, Y) (X variable *succes*, Y les autres variables qualitatives).
  - X une v.a. à p modalités $c_1, \ldots, c_p$. Ici p = 2 et c_1 = succès et c_2 = échec

- Y une v.a. à q modalités $c'_1, \ldots, c'_q$.

Hypothèses de test : **H0 : X et Y indépendantes contre H1 : X et Y liées**

Question 4.2.2.b

Les femmes réussissent-elles mieux que les hommes ? :

Test de Shapiro-Wilk : p-value = 0,5065, on ne peut pas rejeter la normalité, donc on a bien une distribution normale pour la variable probabilité das l'échantillon.

Script :
t.test(proba~sexe, alternative='two.sided', conf.level=.95, var.equal=FALSE, data=ienac20)

Sortie :
data: proba by sexe
t = 2.7172, df = 16.028, p-value = 0.01521
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.4468159 3.6164820
sample estimates:
mean in group F mean in group M
13.01250 10.98085

Le test de Student de comparaison des moyennes rejette l'hypothèse d'égalité des moyennes en probabilité pour les hommes et les femmes avec une p-valeur de 0,01521 < 0,05
On peut donc affirmer à 95 % que les femmes réussissent mieux en moyenne que les hommes.
Cependant, avec le faible nombre de filles étudiantes, on peut se poser des question sur la validité de la normalité de la distribution sur un si faible échantillon, et ainsi remettre en cause le résultat.

Peut-on dire qu'il existe une différence entre les moyennes en analyse et en probabilités ?:

Les 3 premières hypothèses sont vérifiées pour le test de comparaison de deux moyennes-observations appariées.
La dernière hypothèse maintenue, D = X-Y suit une loi normale est difficile à vérifier en revanche.
En la supposant vraie, effectuons le test :

Script :
t.test(ienac20$analyse, ienac20$proba, alternative='two.sided', conf.level=.95, paired=TRUE)

Sortie :
Paired t-test

data: ienac20$analyse and ienac20$proba
t = 9.3204, df = 54, p-value = 7.771e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.751412 4.259497
sample estimates:
mean of the differences
3.505455

Avec une telle p-valeur de 8e-13, on rejette aisément à 95 % l'hypothèse H0 : « Il n'y a pas de différence entre les moyennes d'analyse et de probabilités ». On peut par contre discuter de la validité de la dernière hypothèse maintenue, on devrait effectuer un test de normalité sur D pour savoir s'il est honnête d'effectuer le test sur échantillons appariés.

Peut-on dire que la variable *succes* dépend des autres variables qualitatives ? :

Les hypothèses du test d'indépendance du Chi-Deux citées plus haut sont bien vérifiées ici pour la variable succes comme pour toutes les autres variables qualitatives.

Script :
chisq.test(xtabs(~succes+concours, data=ienac20), correct=FALSE)

Sortie :

```
        Pearson's Chi-squared test

data:  xtabs(~succes + concours, data = ienac20)
X-squared = 0.8523, df = 1, p-value = 0.3559
```

Le test d'indépendance du Chi-Deux n'arrive pas à montrer que les variables succès et concours ne sont pas indépendantes, on acceptera alors l'indépendance du statut Fonctionnaire/Civil avec le succès à l'examen de probabilités, bien que l'on attendait que les fonctionnaires soient meilleurs que les civils.

De la même sorte, on montre que le test ne peut pas rejeter l'indépendance de succès avec la variable mention (au bac), avec une p-valeur de 0,66. On les considère donc indépendantes (avec un risque bêta). On pouvait s'attendre à de meilleurs résultats de la part des hautes mentions, mais le test manque de puissance (ou le bac remonte à trop loin et les étudiants ont tout oubliés).

En revanche, on peut observer que le test rejette l'indépendance entre succès et sexe. Le test estime en effet que les deux sont liés. On aurait pu s'en douter, vu que dans l'échantillon 100 % des filles ont réussi, tandis que 16 garçons sur 47 ont échoué.

Synthèse :

L'étude inférentielle nous a permis d'effectuer et d'interpréter des tests d'hypothèses sur les variables de notre échantillon afin d'obtenir des informations sur la population générale des étudiants IENAC. Cependant, pour réaliser ces tests nous avons du nous accepter des hypothèses pas toujours réalistes, et devons donc manipuler les conclusions avec des pincettes.

Conclusion

Cette étude statistique des étudiants de l'ENAC en formation ingénieur nous a permis d'observer les répartitions des étudiants par rapport aux différentes variables proposées, et ainsi nous avons pu observer des liens entre variables, mais également réaliser des test d'hypothèse pour pouvoir généraliser l'observation de l'échantillon des 55 étudiants à toute la population du cursus.

Cette étude nous a confirmé quelques évidences (ceux qui ont eu mention TB au baccalauréat sont en moyenne meilleurs aux examens que les autres), mais nous a donné parfois des résultats surprenants, certains dénués de sens : Par exemple, en étudiant la variable succès et la varialbe sexe, on voit que 8 filles sur les 8 de l'échantillon ont réussi l'examen de probabilités, et de ce fait les tests trouvent un lien entre les 2 variables, ce qui paraît incohérent a priori.

Annexes

Script :

3.2.1.a)

```
# Table for analyse:
tapply(ienac20$analyse, list(bac=ienac20$bac), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(bac=ienac20$bac), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(concours=ienac20$concours), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(concours=ienac20$concours), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(filiere=ienac20$filiere), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(filiere=ienac20$filiere), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(mention=ienac20$mention), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(mention=ienac20$mention), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(sexe=ienac20$sexe), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(sexe=ienac20$sexe), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(succes=ienac20$succes), mean, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(succes=ienac20$succes), mean, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(bac=ienac20$bac), median, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(bac=ienac20$bac), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(concours=ienac20$concours), median, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(concours=ienac20$concours), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(filiere=ienac20$filiere), median, na.rm=TRUE)
```

```
# Table for proba:
tapply(ienac20$proba, list(filiere=ienac20$filiere), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(mention=ienac20$mention), median, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(mention=ienac20$mention), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(sexe=ienac20$sexe), median, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(sexe=ienac20$sexe), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(succes=ienac20$succes), median, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(succes=ienac20$succes), median, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(bac=ienac20$bac), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(bac=ienac20$bac), sd, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(concours=ienac20$concours), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(concours=ienac20$concours), sd, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(filiere=ienac20$filiere), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(filiere=ienac20$filiere), sd, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(mention=ienac20$mention), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(mention=ienac20$mention), sd, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(sexe=ienac20$sexe), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(sexe=ienac20$sexe), sd, na.rm=TRUE)
# Table for analyse:
tapply(ienac20$analyse, list(succes=ienac20$succes), sd, na.rm=TRUE)
# Table for proba:
tapply(ienac20$proba, list(succes=ienac20$succes), sd, na.rm=TRUE)
```

3.2.1.1.b)

```
with(ienac20, Hist(analyse, scale="frequency", breaks="Sturges", col="darkgray"))
with(ienac20, Hist(ecrit, scale="frequency", breaks="Sturges", col="darkgray"))
with(ienac20, Hist(moyenne, scale="frequency", breaks="Sturges", col="darkgray"))
with(ienac20, Hist(oral, scale="frequency", breaks="Sturges", col="darkgray"))
```

```
with(ienac20, Hist(proba, scale="frequency", breaks="Sturges", col="darkgray"))
with(ienac20, Hist(rang, scale="frequency", breaks="Sturges", col="darkgray"))
with(ienac20, Hist(voeux, scale="frequency", breaks="Sturges", col="darkgray"))

Boxplot( ~ analyse, data=ienac20, id.method="y")
Boxplot( ~ ecrit, data=ienac20, id.method="y")
Boxplot( ~ moyenne, data=ienac20, id.method="y")
Boxplot( ~ oral, data=ienac20, id.method="y")
Boxplot( ~ proba, data=ienac20, id.method="y")
Boxplot( ~ rang, data=ienac20, id.method="y")
Boxplot( ~ voeux, data=ienac20, id.method="y"
```

3.2.1.3)

```
tapply(ienac20$rang, list(filiere=ienac20$filiere), mean, na.rm=TRUE)
tapply(ienac20$rang, list(filiere=ienac20$filiere), median, na.rm=TRUE)
tapply(ienac20$rang, list(filiere=ienac20$filiere), sd, na.rm=TRUE)
Boxplot(rang~filiere, data=ienac20, id.method="y")
```

3.2.2.1)

```
cor(ienac20[,c("analyse","ecrit","moyenne","oral","proba","rang","voeux")],
use="complete")
```

3.2.2.2)

```
tapply(ienac20$proba, list(succes=ienac20$succes), sd, na.rm=TRUE)
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux, reg.line=lm,
smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal = 'density', data=ienac20)

```{r}
scatterplotMatrix(~proba+rang+voeux | concours, reg.line=lm, smooth=TRUE,
spread=FALSE, span=0.5,
  id.n=0, diagonal= 'density', by.groups=FALSE, data=ienac20)
```

```{r}
scatterplotMatrix(~proba+rang+voeux | concours, reg.line=lm, smooth=TRUE,
spread=FALSE, span=0.5,
  id.n=0, diagonal= 'density', by.groups=FALSE, data=ienac20)
```
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | bac,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | concours,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | concours,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | filiere,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | mention,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | sexe,
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

```{r}
scatterplotMatrix(~analyse+ecrit+moyenne+oral+proba+rang+voeux | succes,
```

```
reg.line=lm, smooth=TRUE,
  spread=FALSE, span=0.5, id.n=0, diagonal= 'density', by.groups=FALSE,
data=ienac20)
```

3.2.3)

1.a)
```
xtabs(~succes+bac, data=ienac20)
xtabs(~succes+concours, data=ienac20)
xtabs(~succes+filiere, data=ienac20)
xtabs(~succes+mention, data=ienac20)
xtabs(~succes+sexe, data=ienac20)
```

1.b)
```
colPercents(xtabs(~succes+bac, data=ienac20))
colPercents(xtabs(~succes+concours, data=ienac20))
colPercents(xtabs(~succes+filiere, data=ienac20))
colPercents(xtabs(~succes+mention, data=ienac20))
colPercents(xtabs(~succes+sexe, data=ienac20))
```

1.c)
```
barplot(xtabs(~succes+bac, data=ienac20), legend.text=T,xlab="bac",ylab="%")
barplot(xtabs(~succes+concours, data=ienac20),
legend.text=T,xlab="concours",ylab="%")
barplot(xtabs(~succes+filiere, data=ienac20), legend.text=T,xlab="filiere",ylab="%")
barplot(xtabs(~succes+mention, data=ienac20),
legend.text=T,xlab="mention",ylab="%")
barplot(xtabs(~succes+sexe, data=ienac20), legend.text=T,xlab="sexe",ylab="%")
```

Graphes

Vous trouverez tous les graphes dans le dossier compressé ci-joint