

Corrigé de l'examen CCF2 de statistiques

Correction de l'exercice 1.

- (1) • On étudie la variable qualitative (ou catégorielle) 'piece'. Pour les manipulations avec \mathbb{R} , on renvoie donc aux sections 2.4 et 2.5 du document de cours.
 • Les effectifs et les pourcentages déterminés par \mathbb{R} sont donnés dans le tableau suivant

	effectifs	pourcentages
f	8	32.000
p	17	68.000

•



Voir les deux graphiques ci-dessus pour la variable 'piece'. On constate que les 'pile' représente la moitié environ des 'face'.

- (2) Normalement, ces données ont été choisies au hasard par les étudiants et chacune des modalités 'pile' ou 'face' devrait représenter environ la moitié des effectifs, ce qui n'est pas le cas! soit, les étudiants n'ont pas déterminé ces valeurs au hasard, soit, ce qui est plus vraisemblable, ils ne sont pas assez nombreux pour que cette répartition uniforme des valeurs 'pile' ou 'face' apparaisse clairement.

Remarque 1. La séquence suivante

```
n <- 1000
aleat <- sample(as.factor(c("f", "p")), replace = T, size = n)
```

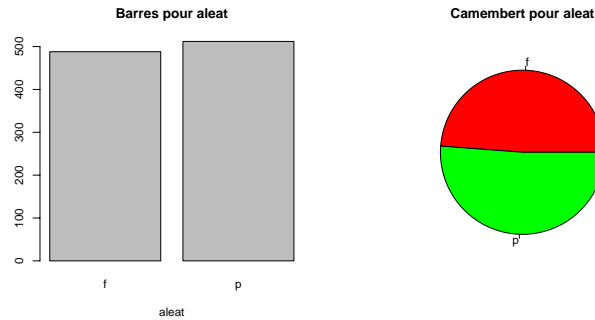
créé un tableau de type catégoriel, avec deux modalités 'f' et 'p' grâce à 1000 tirages aléatoires dans l'ensemble {'f', 'p'}.

- On étudie la variable qualitative (ou catégorielle) 'aleat'.

	effectifs	pourcentages
f	488	48.800
p	512	51.200

- Les effectifs et les pourcentages déterminés par \mathcal{R} sont donnés dans le tableau suivant

-



Voir les deux graphiques ci-dessus pour la variable 'aleat'. Ici, les proportions observées pour chacune des modalités 'f' et 'p' sont proches de la moitié. Les effectifs sont proches de 500, ce qui correspond à $1000/2$. En théorie des probabilités, chaque modalité 'f' ou 'p' "a autant de chance de sortir", ce qui justifie la valeur de $1/2$, probabilité d'apparition de chacune des modalités.

Correction de l'exercice 2.

- On étudie le croisement de la variable qualitative (ou catégorielle) 'sexe' et de la variable qualitative (ou catégorielle) 'baccalaureat'. Pour les manipulations avec \mathcal{R} , on renvoie donc à la section 5.5 du document de cours.
 - La table de contingence déterminée par \mathcal{R} est donnée dans le tableau suivant

	ES	L	S	SMS	STAE
féminin	10	6	14	1	0
masculin	6	0	20	0	1

Les autres résultats donnés par \mathcal{R} sont les suivants :

Noms des indicateurs	Valeurs
χ^2	9.829714
coefficient de Cramer V	0.411677
taille d'effet w	0.411677
probabilité critique p_c	0.0433959

On compare la taille d'effet $w=0.411677$ aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique $p_c=0.0433959$ à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	forte
significativité statistique	oui

- On peut donc affirmer qu'il existe une relation entre les variables 'sexe' et 'baccalaureat'.
- (2) Il apparaît, au vu de la table de contingence, que les filles s'orientent plus vers les bacs 'ES' et 'L' que les garçons, qui eux, choisissent plus le bac 'S'. Cela est confirmé par les calculs de χ^2 .

Correction de l'exercice 3.

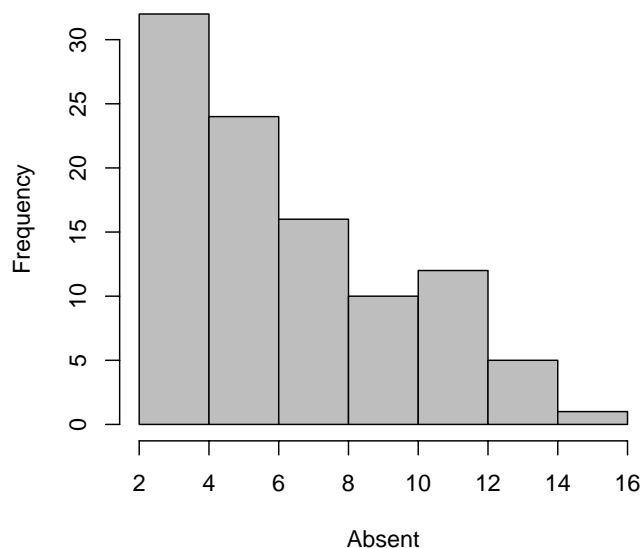
Cet exercice provient d'un examen L3MOS précédent, donné par Stéphane Champely (Printemps 2009).

- (1) (a) • On étudie la variable quantitative (ou numérique) 'Absent'. Pour les manipulations avec \mathcal{R} , on renvoie donc aux sections 3.2 et 3.3 du document de cours.
- Les différents résultats déterminés par \mathcal{R} sont donnés dans le tableau suivant

noms	valeurs
moyenne	6.232
sd	3.356913
Q_1 (quartile à 25 %)	3.45
médiane	5.65
Q_3 (quartile à 75 %)	8.6
minimum	2.1
maximum	14.8
nombre	100

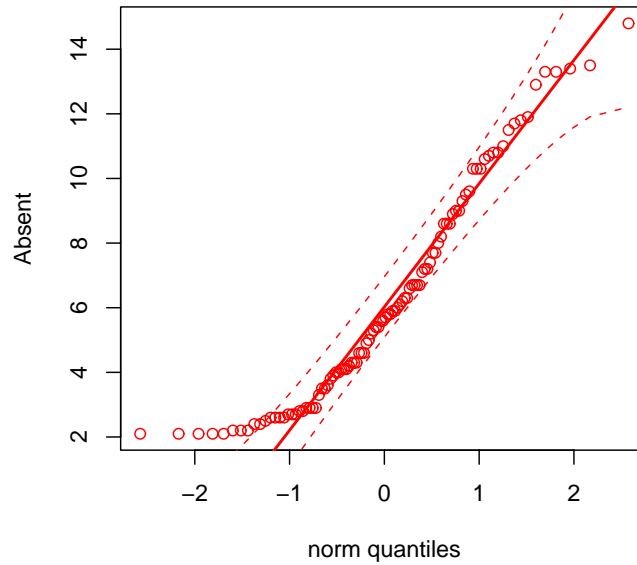
•

Histogramme pour Absent



Voir l'histogramme ci-dessus pour la variable 'Absent'.

- Le nombre de données est assez important pour réaliser un histogramme avec confiance. On peut voir sur ce graphique que la distribution est clairement dissymétrique.



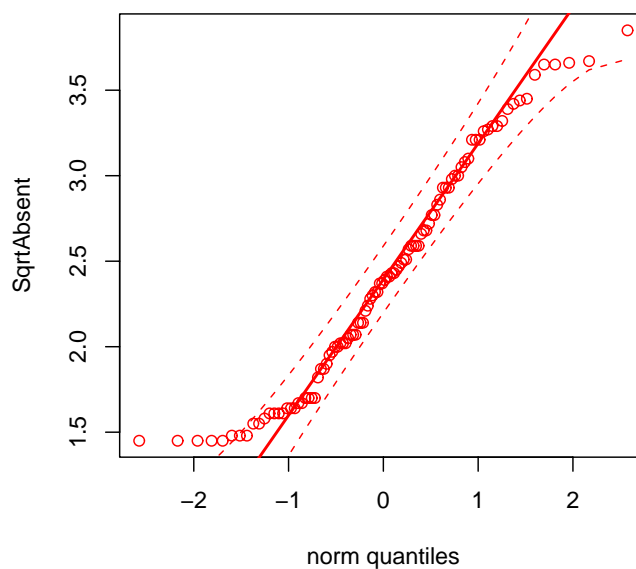
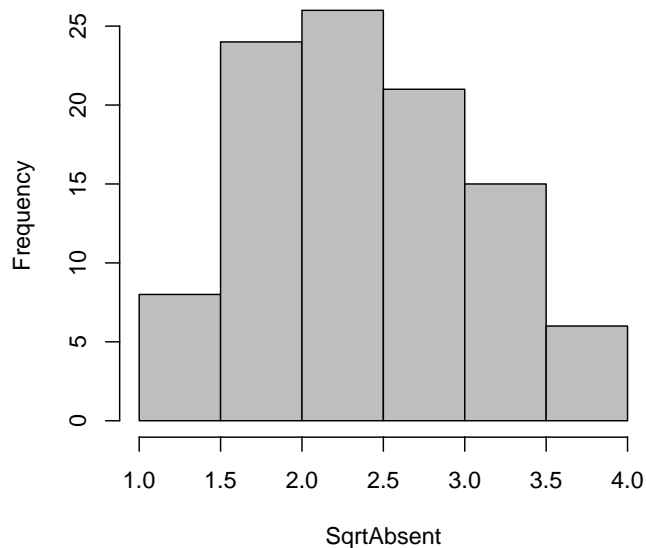
Ceci peut être confirmé par d'autres graphiques (graphe quantile-quantile en particulier, voir figure ci-dessus). Il semble que le nombre typique de jours d'absentéisme soit de l'ordre de 6 à 8, ce qui est confirmé par la valeur de la médiane : $Q_2 = 5.65$. Notons aussi que toutes les entreprises ont au minimum 2.1 jours d'absentéisme (avec les conséquences financières que cela implique).

- (b) • On étudie la variable quantitative (ou numérique) 'SqrtAbsent'.
 • Les différents résultats déterminés par \mathcal{R} sont donnés dans le tableau suivant

noms	valeurs
moyenne	2.4063
sd	0.665847
Q_1 (quartile à 25 %)	1.8575
médiane	2.38
Q_3 (quartile à 75 %)	2.93
minimum	1.45
maximum	3.85
nombre	100

•

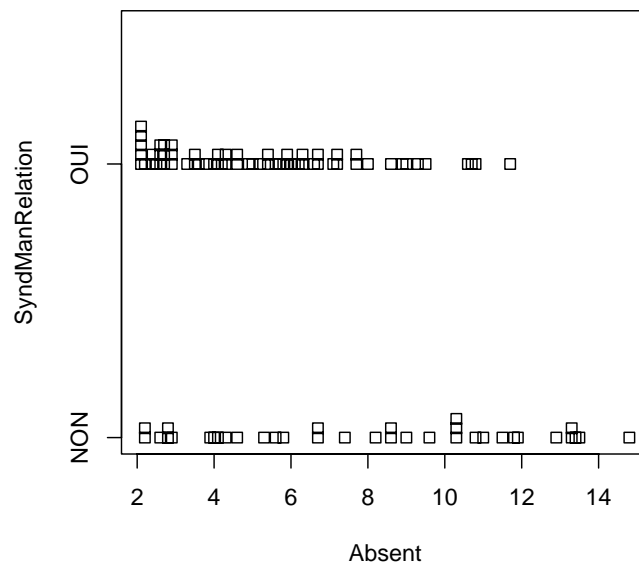
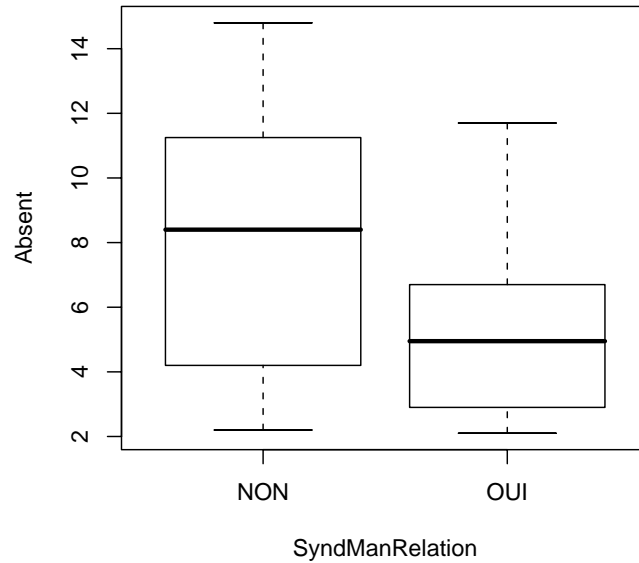
Histogramme pour SqrtAbsent



Voir les graphiques ci-dessus pour la variable 'SqrtAbsent'.

- En ce qui concerne la racine carrée, on va se contenter d'une analyse graphique, à nouveau basée sur l'histogramme et complétée par un graphe quantile-quantile. On voit que la distribution des données est beaucoup plus symétrique. Les données ne sont pas tout à fait normales, elles sont un peu moins dispersées. L'intérêt est que sa description statistique est facilitée (Moyenne et écart-type suffisent), en revanche en termes d'interprétation c'est plus difficile car l'unité de mesure devient inhabituelle.

- (2) • On étudie le croisement de la variable qualitative (ou catégorielle) 'SyndManRelation' et de la variable quantitative (ou numérique) 'Absent'. Pour les manipulations avec \mathbb{R} , on renvoie donc aux sections 6.2 et 6.3 du document de cours.
-



Voir la figure ci-dessous.

- Avec \mathbb{R} , on obtient les statistiques par groupes données dans le tableau suivant ;

	moyenne	écart-type (sd)	0%	25%	50%	75%	100%	n
NON	7.97	3.91	2.20	4.25	8.40	11.12	14.80	36
OUI	5.25	2.55	2.10	2.90	4.95	6.70	11.70	64

On rappelle que :

- le quartile à 0 % correspond au minimum ;
- le quartile à 25 % correspond à Q_1 ;
- le quartile à 50 % correspond à la médiane ;
- le quartile à 75 % correspond à Q_3 ;
- le quartile à 100 % correspond au maximum.

Les graphiques par groupes montrent que l'absentéisme est plus bas lorsqu'il y a de bonnes relations entre les syndicats et le management. On peut également distinguer que dans les entreprises où les relations ne sont pas bonnes, la situation est plus contrastée (plus de dispersion). De plus, les statistiques descriptives par groupes mettent en évidence une différence de 2.7 jours d'absentéisme entre ces deux types d'entreprises.

Confirmons cela grâce à \mathcal{R} .

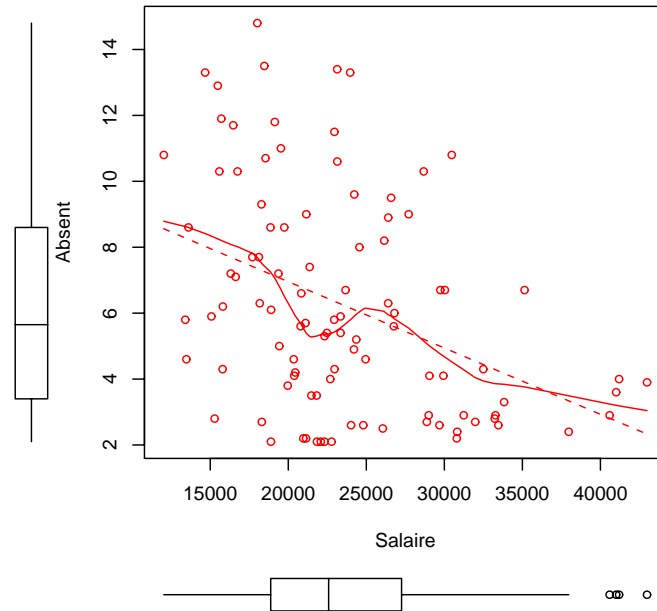
Les autres résultats donnés par \mathcal{R} sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.152692
probabilité critique p_c	5.8326e-05

On compare le rapport de corrélation $RC=0.152692$ aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique $p_c=5.8326e-05$ à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

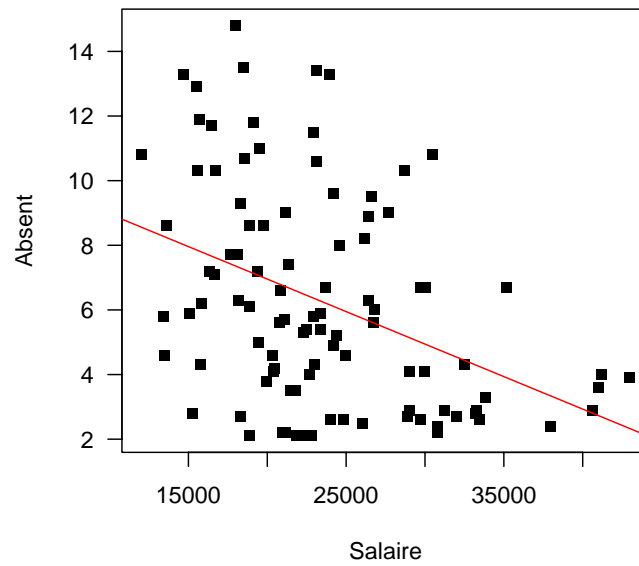
significativité pratique	très forte
significativité statistique	oui

- On peut donc affirmer qu'il existe une relation entre les variables 'SyndManRelation' et 'Absent'.
- (3)
- On étudie le croisement de la variable quantitative (ou numérique) 'Salaire' et de la variable quantitative (ou numérique) 'Absent'.
 - Voir la figure ci-dessous.
 - Avec *Recmdr* :



- *Sans Rcmdr* :

Absent en fonction de Salaire



Sur cette figure, les points semblent alignés.

- Confirmons cela grâce à \mathbb{R} .
Les résultats donnés par \mathbb{R} sont les suivants :

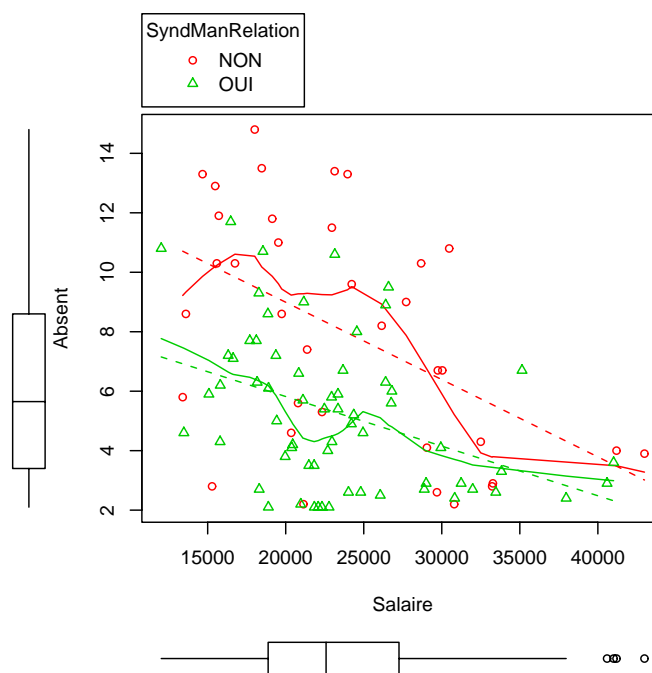
Noms des indicateurs	Valeurs
pende a	-0.000201
ordonnée à l'origine b	10.977259
corrélation linéaire r	-0.398894
probabilité critique p_c	3.94348e-05

On compare la valeur absolue de la corrélation linéaire $r = -0.398894$ aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique $p_c = 3.94348e-05$ à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	forte
significativité statistique	oui

- On peut donc affirmer il existe une relation faible entre les variables 'Salaire' et 'Absent'. De plus, le coefficient directeur (-2e-04) est négatif : plus le salaire moyen augmente, moins il y a d'absentéisme. Enfin, sa valeur nous indique que pour 10000 \$ (annuel) de différence on peut avoir $0.0002012 \times 10000 = 2.01$ jours d'absentéisme en moins.

(4) Corrigions les deux questions simultanément.



Voir la figure ci-dessus.

- relations entre l'absentéisme et le salaire dans les entreprises où les relations syndicats-management sont bonnes :

Les résultats donnés par \mathcal{R} sont les suivants :

Noms des indicateurs	Valeurs
pente a	-0.000167
ordonnée à l'origine b	9.15791
corrélation linéaire r	-0.404617
probabilité critique p_c	0.00091258

On compare la valeur absolue de la corrélation linéaire $r = -0.404617$ aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique $p_c = 0.00091258$ à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	forte
significativité statistique	oui

Remarque 2. Ces résultats peuvent être obtenus en tapant par exemple dans "Rgui"
`determin.quantiquanti(absenteisme$Salaire[absenteisme$SyndManRelation == "OUI"], absenteisme$Absent[absenteisme$SyndManRelation == "OUI"])`

- relations entre l'absentéisme et le salaire dans les entreprises où les relations syndicats-management sont mauvaises :

Les résultats donnés par \mathcal{R} sont les suivants :

Noms des indicateurs	Valeurs
pente a	-0.00026
ordonnée à l'origine b	14.193252
corrélation linéaire r	-0.499093
probabilité critique p_c	0.00194431

On compare la valeur absolue de la corrélation linéaire $r = -0.499093$ aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique $p_c = 0.00194431$ à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	forte
significativité statistique	oui

Remarque 3. Ces résultats peuvent être obtenus en tapant par exemple dans "Rgui"
`determin.quantiquanti(absenteisme$Salaire[absenteisme$SyndManRelation == "NON"], absenteisme$Absent[absenteisme$SyndManRelation == "NON"])`

Ainsi, par rapport à la corrélation linéaire où l'on ne précise pas la nature des relations syndicats-management, les corrélations sont plus grandes, donc la qualité de la liaison meilleure (cela peut être aussi tout simplement dû au fait que, par groupe, le nuage de point contient moins de points!). Sur le graphique, on peut constater que les deux droites sont que la droite verte (correspondant aux bonnes relations) est "sous" la rouge. Par le calcul, cela est confirmé par les deux valeurs de pentes, du même ordre : -0.0001669 pour les bonnes relations et -0.0002601 pour les mauvaises. Mais surtout, regardons la valeurs des ordonnées à l'origine : 9.1579 pour les bonnes relations et 14.19 pour les mauvaises. Ainsi,

à salaire égal, l'absentéisme sera plus faible dans les entreprises avec de bonnes relations syndicats-management que dans celles où elles sont mauvaises ; cet écart diminue quand les salaires s'élèvent. Pour le management, on peut donc compenser de faibles salaires par de bonnes relations avec les syndicats, on peut en quelque sorte acheter *la paix sociale*.

Références

[Coh92] J Cohen. A power primer. *Psychological bulletin*, 112(1) :155–159, 1992.