Année universitaire 2019-2020

TECHNIQUES STATISTIQUES POUR L'EXPÉRIMENTATION

Semestre 7

Auteur : Florent ARNAL
Adresse électronique : florent.arnal@u-bordeaux.fr
Site : http://flarnal.e-monsite.com

Les exercices sont classés suivant différentes catégories présentées ci-dessous :

* R identifie un exercice de révision en lien avec le programme de 1ère année

* A identifie un exercice à réaliser en autonomie (TPNE).

* CR identifie un compte-rendu à rédiger suivant un cahier des charges précisé dans l'énoncé.

* E identifie un exercice lié au programme de S7 autour des techniques statistiques pour l'expérimentation.

* Pour aller plus loin identifie un exercice proposant des prolongements de notions abordées dans le cours de Techniques Statistiques pour l'Expérimentation.

Application 1 : R A

Pour cet exercice, il peut être utile de consulter la partie *Calculs de probabilités* du groupe Begin'R
http://beginr.u-bordeaux.fr/part4-proba.html

1. On suppose que X est distribuée suivant la loi normale d'espérance 100 et d'écart-type 5.
   (a) Calculer $\mathbb{P}(X \leq 110)$, $\mathbb{P}(X > 105)$ et $\mathbb{P}(95 < X < 105)$.
   (b) Déterminer la valeur u telle que :
   $$\mathbb{P}(X < u) = 0.95$$
2. Lors d'un test de comparaison de deux proportions p_1 et p_2, avec de grands échantillons, on obtient comme réalisation de la statistique de test (la loi associée étant la loi normale centrée réduite)
   $$u_{obs} = 2$$
   On teste l'hypothèse selon laquelle la proportion p_1 est supérieure à p_2.
   Formuler les hypothèses de ce test et conclure après avoir calculé la p-valeur.
3. On s'intéresse à un caractère quantitatif X distribué suivant une loi normale de moyenne μ et d'écart-type $\sigma = 2$.
   On souhaite tester :
   $$\mathcal{H}_0 : \mu = 0$$
   $$\mathcal{H}_1 : \mu > 0$$
   On prélève des échantillons de taille $n = 10$.
   (a) Les résultats observés sur un échantillon sont les suivants :

1.40	1.21	1.35	1.38	0.95	1.35	1.55	0.96	0.70	0.39

   La moyenne observée est-elle significativement supérieure à 0 ?
   (b) Plus généralement, à partir de quelle valeur $\bar{x}$ va-t-on conclure que la moyenne est significativement supérieure à 0 ?

Application 2 : R A

Le titre alcoométrique volumique acquis, en % vol, doit être mentionné sur l'étiquette d'un vin avec une tolérance de $0,5\%$ par volume (art. 54 du règlement CE numéro 607/2009 du 14.7.09).
On considère un vin pour lequel est annoncé sur l'étiquette

13 % vol

On prélève un échantillon de 10 bouteilles et on obtient les teneurs en alcool suivantes :

```
[1] 12.5 12.6 13.0 13.3 12.6 13.1 12.7 13.1 12.7 12.9
```

1. Proposer une représentation graphique.
2. Peut-on considérer que la production dont est extrait cet échantillon est conforme au cahier des charges ?

Application 3 : R A

On considère que, pour un cépage donné, un vin rouge peut paraître acide si le pH est inférieur à 3,4.
On prélève un échantillon de 10 bouteilles dont les pH sont les suivants :

```
[1] 3.30 3.37 3.43 3.28 3.42 3.40 3.41 3.51 3.28 3.53 3.33 3.29
```

Peut-on considérer le pH moyen des vins de la production inférieur à 3,4 ?

Application 4 : A R
Comparer les pH de ces deux cépages au vu des résultats observés récapitulés ci-dessous :
Merlot :

```
[1] 3.56 3.48 3.57 3.56 3.52 3.42 3.45 3.49 3.50
```

Cabernet Franc :

```
[1] 3.57 3.61 3.56 3.68 3.62 3.56 3.62 3.64 3.63
```

Application 5 : A E
On considère deux populations gaussiennes de même écart-type $\sigma = 10$.

1. On fait l'hypothèse que l'écart entre les deux moyennes est tel que

$$\mu_1 - \mu_2 = 5$$

   Quelle taille d'échantillons préconiseriez-vous pour comparer ces deux populations ?
2. Quelle serait votre conclusion si l'on envisageait que

$$\mu_1 - \mu_2 = 1$$

Application 6 : R E
On souhaite comparer les pH associés à trois cépages : Merlot, Cabernet Sauvignon et Cabernet Franc.
Les résultats sont donnés ci-dessous :

	Merlot	Cabernet_Sauvignon	cabernet_Franc
1	3.32	3.53	3.48
2	3.15	3.63	3.50
3	3.19	3.54	3.52
4	3.24	3.57	3.55
5	3.30	3.60	3.57

1. Importer les données dans R à partir d'un fichier csv ne comportant que deux colonnes correspondant aux cépages et aux pH.
2. Déterminer des résumés paramétriques et graphiques de ces données.
3. Les conditions d'utilisation d'une ANOVA sont-elles vérifiées ?
4. Formuler les hypothèses du test associé à la détection d'un effet factoriel, préciser la statistique de test et la loi associée ainsi que la nature du test.
5. Calcul de résidus
   (a) Calculer les résidus associés à la modalité "Merlot".
   (b) En utilisant la fonction resid( ), déterminer les résidus associés à cette analyse de variance.
   (c) Déterminer la SCE résiduelle.
6. Calculer la SCE totale et en déduire la SCE factorielle.
7. (a) Compléter le tableau d'anova et conclure.
   (b) Retrouver ces résultats avec la fonction *summary()*.
8. Donner une estimation ponctuelle des effets, les représenter et proposer la constitution de différents groupes.

Application 7 : E CR

On souhaite comparer l'effet de différents types de levure sur les concentrations (en ng/l) d'un arôme d'un vin : l'acétate de 3-mercaptohexyle (3MHA).
Les résultats observés sont les suivants :

```
       levures concentrations
1      pure Sc          589.6
2      pure Sc          486.7
3      pure Sc          515.4
4      pure Sc          457.6
5      pure Sc          404.5
6  mixed Sc/Cz          223.4
7  mixed Sc/Cz          172.0
8  mixed Sc/Cz          304.5
9  mixed Sc/Cz          327.3
10 mixed Sc/Cz          234.2
11 mixed Sc/Hu          270.7
12 mixed Sc/Hu          233.3
13 mixed Sc/Hu          223.4
14 mixed Sc/Hu          167.8
15 mixed Sc/Hu          180.9
16 mixed Sc/Mp          306.5
17 mixed Sc/Mp          255.1
18 mixed Sc/Mp          394.8
19 mixed Sc/Mp          319.4
20 mixed Sc/Mp          384.9
21 mixed Sc/Pk          473.5
22 mixed Sc/Pk          444.4
23 mixed Sc/Pk          389.3
24 mixed Sc/Pk          537.0
25 mixed Sc/Pk          481.5
```

Proposer, sur 1 page maximum, un résumé de cette analyse de résultats (en vous appuyant sur des graphiques) à destination d'un public de statisticiens.

Application 8 : Pour aller plus loin

On se propose de comparer les rendements moyens en q/ha de blé de variété "étoile de Choisy" pour une petite région agricole de la vallée de la Marne. On monte un plan d'expérience afin de tester si l'apport d'un autre élément fertilisant que l'azote a un effet significatif sur les rendements pour cette céréale et dans cette région.
Pour information, le plan d'expérience est représenté ci-dessous avec les rendements obtenus :

N	NK	NPK	N	NP
NK	NPK	NK	NP	N
NP	NP	N	NK	NPK
NPK	N	NP	NPK	NK

78	73,2	78,5	62,5	72,6
82,8	83,6	73	75,4	56,5
82,8	81,1	63,3	72,4	80,3
85,4	65,2	80,2	72,4	72,3

1. Déterminer des résumés paramétriques et graphiques de ces données.
2. Déterminer les résidus et les représenter.
3. A l'aide des graphiques diagnostiques, les conditions d'application d'une ANOVA semblent-elles vérifiées ?
4. Réaliser une cartographie des résidus à l'aide de R, en utilisant la fonction levelplot du package Lattice.

Application 9 : E

On souhaite comparer les effets de trois traitements notés T_1, T_2 et T_3.

1. Une étude précédente a permis d'évaluer l'écart-type des traitements à 0,1.
   L'expérimentateur envisage que le traitement T_1 devrait générer une augmentation de la moyenne de 0,1 par rapport aux traitements T_2 et T_3 (vraisemblablement identiques).
   On souhaite mettre en place une ANOVA avec un plan équilibré avec 6 répétitions.
   En utilisant la fonction *power.anova.test()*, déterminer la puissance a priori d'une anova adaptée à cette étude.
2. Lors de cette expérimentation, on a obtenu les résultats suivants :

T_1	1,56	1,46	1,75	1,56	1,65	1,6
T_2	1,5	1,57	1,45	1,56	1,46	1,21
T_3	1,53	1,4	1,6	1,59	1,57	1,47

   (a) Réaliser l'ANOVA associée à ces données. Conclure.
   (b) Déterminer la puissance a posteriori de ce test.

Application 10 : E

Une étude a été menée sur l'influence de la situation géographique de placettes sur une parcelle (Est ou Nord) et du Cépage (Merlot ou Cabernet-Franc) sur le pH.
Les résultats obtenus sont récapitulés dans le tableau ci-dessous :

Cépages	Position	pH	Cépages	Position	pH
Merlot	E	3,46	CF	E	3,53
Merlot	E	3,44	CF	E	3,5
Merlot	E	3,46	CF	E	3,63
Merlot	E	3,46	CF	E	3,56
Merlot	E	3,43	CF	E	3,53
Merlot	E	3,51	CF	E	3,54
Merlot	NE	3,48	CF	NE	3,55
Merlot	NE	3,45	CF	NE	3,6
Merlot	NE	3,47	CF	NE	3,57
Merlot	NE	3,49	CF	NE	3,61
Merlot	NE	3,5	CF	NE	3,63
Merlot	NE	3,51	CF	NE	3,5

1. (a) Représenter, à l'aide de box-plot, les pH en fonction de la situation géographique puis du Cépage.
   (b) Représenter, à l'aide d'un box-plot, les pH en fonction de la situation géographique pour le Merlot uniquement.
2. Conjecturer, à l'aide de graphiques, d'éventuelles interactions.
3. Analyser les résultats de cette expérimentation.

Application 11 : CR

L'objectif est de traiter des données liées à une étude réalisée par la Chambre d'Agriculture de Haute-Garonne. Cet essai avait pour but d'optimiser la production de blé dur (rendement, gestion des maladies, apport fongicide) en utilisant 2 leviers agronomiques : la variété (Isildur et Miradoux) et la date de semis (Octobre et Novembre). Pour mesurer l'effet de fongicides, on a utilisé un témoin et deux associations caractérisées ci-dessous :

TNT : témoin traité avec 1 fongicide à la floraison

Piloté : 1 fongicide à la floraison + 2 fongicides

Assurance : 1 fongicide à la floraison + 3 fongicides

Une première étude a montré que la date de semis a une influence sur le rendement suivant la variété. En conséquence, l'expérimentateur a préconisé de se concentrer principalement sur les résultats obtenus pour des semis effectués au mois d'octobre.
Les résultats obtenus pour le mois d'octobre (rendements exprimés en q/ha) sont accessibles sur Moodle.

Proposer un compte-rendu (à destination d'un public de techniciens agricoles) lié au traitement de ces données. La démarche statistique sera explicitée en annexe.

Indications :
• On mettra en évidence un problème d'homoscédasticité à l'aide d'une représentation graphique.
• On rappelle qu'un test non paramétrique peut parfois être utilisé lorsque les conditions d'application d'un test paramétrique ne sont pas vérifiées.
• On pourra proposer une hypothèse liée au choix du mois d'Octobre (plutôt que Novembre).

variétés	fongicide	rendement	variétés	fongicide	rendement
Miradoux	Assurance	56,19	Miradoux	Assurance	47,82
Miradoux	Piloté	45,68	Miradoux	Piloté	58,26
Miradoux	Assurance	58,08	Miradoux	Assurance	56,53
Miradoux	TNT	49,66	Miradoux	TNT	45,58
Miradoux	TNT	44,35	Miradoux	TNT	44,61
Miradoux	TNT	45,78	Miradoux	TNT	43,21
Miradoux	Piloté	50,38	Miradoux	Piloté	44,42
Miradoux	Assurance	45,85	Miradoux	Assurance	54,23
Miradoux	Piloté	56,75	Miradoux	Piloté	51,77
Isildur	Assurance	40,22	Isildur	Assurance	41,77
Isildur	Piloté	34,22	Isildur	Piloté	45,51
Isildur	Assurance	41,57	Isildur	Assurance	43,33
Isildur	TNT	40,8	Isildur	TNT	38,97
Isildur	TNT	39,68	Isildur	TNT	49,34
Isildur	TNT	40,9	Isildur	TNT	39,78
Isildur	Piloté	46,3	Isildur	Piloté	49,58
Isildur	Assurance	41,92	Isildur	Assurance	43,41
Isildur	Piloté	53,89	Isildur	Piloté	37,56

Application 12 : Pour aller plus loin

On a constitué un jury de 6 personnes avec 3 juges professionnels (J_1, J_4 et J_6) ainsi que 3 juges non professionnels (J_2, J_3 et J_5) mais formés.
Chaque personne a attribué plusieurs notes (sur une échelle de 1 à 10) à 4 vins différents en réponse à la question suivante :
"Considérez- vous que ce vin est représentatif d'un grand cru classé de Bordeaux ?".
Les notes (correspondent à une moyenne sur 10) sont présentées ci-dessous :

Jurés	$Vin\ 1$	$Vin\ 2$	$Vin\ 3$	$Vin\ 4$
J_1	4,05	6,85	4,21	4,61
J_2	4,54	4,65	2,22	5,01
J_3	3,23	3,84	5	5,72
J_4	5,01	5,36	4,27	3,32
J_5	3,56	5,64	6,86	3,91
J_6	4,39	2,95	4,58	4,05

Interpréter ces résultats.

Application 13 : Pour aller plus loin

Une étude a été menée pour étudier l'effet de différents types de levures (Saccharomyces et non-Saccharomyces) sur la perception des notes fruitées, la force, la complexité, la présence de quelques arômes.
Les données fournies correspondent à :
• la force (notes fournies par un jury) ;
• la concentration d'une molécule, le 4-mercapto-4-méthylpentan-2-one (4MMP), associée à une odeur marquée de buis et de genêt. Les valeurs sont exprimées ng/l (le seuil de perception de cette molécule est de 0,8 ng/l).

```
   Levure Force 4MMP
1       1   8.6 44.5
2       1   6.8 44.1
3       1   7.9 42.1
4       1   7.2 43.7
5       1   6.3 46.8
6       2   3.4 36.0
7       2   2.6 37.5
8       2   4.9 40.2
9       2   4.8 35.9
10      2   3.5 35.8
11      3   4.3 37.0
12      3   3.3 34.1
13      3   3.3 36.7
14      3   2.6 39.5
15      3   2.6 34.2
16      4   4.8 47.2
17      4   3.8 51.6
18      4   5.7 52.7
19      4   4.5 48.6
20      4   5.4 49.8
21      5   7.3 49.8
22      5   6.2 56.0
23      5   5.7 50.8
24      5   7.6 51.5
25      5   6.7 56.2
```

Interpréter ces résultats.

Application 14 : Pour aller plus loin

L'objectif de cet exercice est de travailler sur la puissance d'un test de conformité d'une moyenne dans le cas d'une distribution normale.
On rappelle que lorsqu'on prélève des échantillons de taille n, on considère la statistique de test

$$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$$

On va s'interesser à la capacité d'un test à détecter une différence de moyenne lorsque on fait l'hypothèse

$$\mathcal{H}_0 : \mu = 10$$

1. On souhaite évaluer la capacité d'un test à détecter une différence $\Delta = 1$ lorsque l'écart-type est voisin de 1 (mais inconnu) pour des échantillons de taille $n = 10$. On teste ici $\mathcal{H}_0$ contre

$$\mathcal{H}_1 : \mu = 11$$

   (a) En utilisant le script ci-dessous, déterminer une estimation de la puissance de ce test. Justifier la réponse.

```
> nbech = 1000
> pvalue = matrix(0,nrow=nbech , ncol=1)
> n = 10
> echantillon = matrix(0,nrow=nbech , ncol=n)
> for (i in 1:nbech) {
+   for (j in 1:n){
+     echantillon[i,j] = rnorm(n=1, mean=11 , sd=1)
+   }
+ }
> # echantillon
> for (k in 1:nbech) {
+   pvalue[k] = t.test(x=echantillon[k,], mu=10, alternative="greater")$p.value
+ }
>
```

   (b) Déterminer la puissance de ce test en utilisant la fonction power.t.test(n=... , delta=..., sd=..., type="one.sample", alternative="one.sided", power=NULL).

2. (a) Quelle taille d'échantillon préconiseriez-vous pour détecter un différence de 0,2 (lors d'un test unilatéral à droite) lorsque l'écart-type (estimé) est égal à 0,5 ?

   (b) Que pourriez-vous conclure si vous n'arriviez pas à détecter une telle différence sur un échantillon de taille 20 (dans les conditions décrites précédemment) ?