TD/TP de MSP

Pour traiter ces exercices, on utilisera le RStudio, environnement de développement multiplateforme gratuit et open source pour R, langage de programmation utilisé pour le traitement de données et l'analyse statistique.

I Statistique descriptive

Exercice 1 :
Un étudiant a obtenu les notes suivantes au cours du semestre :

```
[1]  6  9 10 11 11 12 13 14 17
```

1. Calculer la moyenne et la variance de ces notes.
2. Proposer une représentation graphique de ces notes.
3. Déterminer la médiane ainsi que les premier et troisième quartile.
   À partir de quelles valeurs considèrerait-on une note aberrante ?

Exercice 2 :
Cet exercice est lié à des notions utilisées lors d'essais interlaboratoires pour estimer la "qualité" de méthodes d'essais. On s'intéresse à des mesures (analyses) sur un produit en effectuant 2 répétitions dans un laboratoire. On considère donc un échantillon constitué de 2 valeurs y_1 et y_2.

Exprimer la SCE et la variance observée s^2 de cette série de mesures $(y_1; y_2)$ en fonction de son étendue $w = |y_1 - y_2|$.
Remarque : si cette approche théorique vous paraît difficile, vous pouvez commencer par un exemple numérique en prenant, par exemple, $y_1 = 9$ et $y_2 = 12$.

Exercice 3 :
Créer le diagrammme de Pareto associé aux données suivantes :

Machine	1	2	3	4	5	6	7	8
Nombre de pannes	26	13	35	24	20	52	5	2

florent.arnal@u-bordeaux.fr

Exercice 4 :
La nutrition du jeune adulte est à ce jour un sujet largement inexploré dans le monde. La grande cohorte d'étudiants i-Share (Internet-based Students Health and Research Enterprise), en cours de constitution à Bordeaux, offre une opportunité de mieux connaître les habitudes alimentaires de cette population, afin de mieux définir des stratégies de santé publique dans le futur. Les données sont accessibles via le fichier 'DonneesIShare.csv' disponible sur moodle (avec un descriptif de l'enquête).

1. Combien y a-t-il de personnes âgées de 21 ans qui ont participé à l'enquête ?
2. Dans cette partie, on ne considère que les personnes âgées de 21 ans.
   (a) Représenter graphiquement le nombre de fruits mangés quotidienement.
   (b) Déterminer le nombre moyen ainsi que le nombre médian de fruits mangés par ces personnes.
3. Déterminer un résumé paramétrique du nombre de fruits consommés quotidiennement pour chaque catégorie d'âge.
4. Déterminer une représentation graphique (à l'aide de boxplots) du nombre de fruits consommés en fonction de l'âge.

Exercice 5 :
On s'intéresse à une production de bouteilles de jus de fruits (QN = 150 cl) avec un tiers de Mangue, un tiers d'Orange et un tiers d'eau.
On suppose que les distribution des quantités de chaque constituant sont normales, d'espérance 50 et d'écart-type σ.

1. Dans cette partie, on suppose que $\sigma = 0.2$ cl.
   (a) Créer des échantillons de quantités associés à la mangue, l'orange et l'eau.
   Proposer une représentation graphique des quantités associées à ces trois constituants.
   (b) En déduire un échantillon de volumes associés à 50 bouteilles de cette production.
   Donner une estimation de l'écart-type des volumes des bouteilles de cette production.
   (c) La distribution des quantités des bouteilles de la production est-elle normale ?
   (d) La cahier des charges de l'entreprise impose que la quantité de jus de fruit dans une bouteille doit être comprise entre 149,5 et 150,5 cl. La production sera-t-elle jugée conforme ?
   (e) Déterminer la proportion d'éléments présentant un défaut lié à la quantité sur cet échantillon. Proposer une représentation graphique.
2. Quelle est la valeur maximale de σ pour considérer la production conforme ?

Exercice 6 :
Cet exercice a pour objectif d'aborder, au travers d'un exemple, les notions essentielles en statistiques à deux variables.
Afin de procéder à l'étalonnage d'un nouvel appareil de mesures, on effectue 5 mesures (grandeurs obtenues) avec cet appareil associées à 5 valeurs de référence X (grandeurs théoriques).
Pour chaque mesure effectuée, dont on connaît la valeur de référence, on a repertorié la valeur obtenue, noté Y, .
A noter que les valeurs observées ont toujours été supérieures aux valeurs attendues. Les résultats sont consignés dans le tableau suivant.

florent.arnal@u-bordeaux.fr

Valeurs de référence x_i	0,2	0,5	1	2	4
Observations y_i	0,283	0,676	1,311	2,631	5,231

1. Représenter graphiquement cette série de données.
   Un ajustement affine vous paraît-il judicieux ?
2. Déterminer les coefficients de corrélation linéaire et de détermination entre les variables X et Y.
   Interpréter ces valeurs.
3. Déterminer une équation de la droite d'ajustement linéaire de Y en X.
4. Calculer les estimations $\hat{Y}$.
5. Calculer les différents résidus et les représenter en fonctions de X.
6. En déduire la SCE_{res} puis la SCE_{exp}.
7. Retrouver, en utilisant un quotient de SCE, la valeur du coefficient de détermination.
8. Pour une valeur mesurée (observée) de 6,5, donner un ordre de grandeur de la mesure attendue (théorique).

florent.arnal@u-bordeaux.fr

II Calculs de probabilité

Exercice 7 :
Une grande entreprise organise un voyage de groupe à Marrakech pour récompenser certains de ses collaborateurs.
On admet que la probabilité qu'une personne invitée se rende à ce séjour est égale à $0,9$.

1. L'entreprise a décidé de récompenser 10 salariés de la région bordelaise.
   (a) Quelle est la probabilité qu'ils se rendent tous à ce séjour ?
   (b) Quelle est la probabilité qu'au moins 8 personnes se rendent à ce séjour ?
2. L'entreprise a décidé de récompenser 200 salariés sur toute la France.
   (a) Donner une estimation du nombre de salariés qui participeront à ce séjour.
   (b) Déterminer la probabilité qu'il y ait entre 170 et 190 participants à ce séjour.

Exercice 8 :
Une machine calibre des tomates avec une valeur cible de 150 grammes.
On admet que la variable aléatoire Y égale à la masse, en grammes, d'une tomate prise au hasard, est distribuée suivant une loi normale $\mathcal{N}(150; \sigma)$.

1. Déterminer la valeur de σ sachant que 84 % des tomates ont une masse inférieure à 160 g.
2. On prend, pour cette partie, $\sigma = 10$.
   (a) Quelle est la probabilité qu'une tomate prise au hasard pèse moins de 130 g ?
   (b) Quelle est la probabilité qu'une tomate prise au hasard pèse plus de 170 g ?
   (c) Quelle est la probabilité qu'une tomate prise au hasard pèse entre 145 g et 160 g ?
3. Quelle devrait être la valeur de l'écart-type pour que 95 % des tomates aient un poids compris entre 145 g et 155 g ?

Exercice 9 :
On suppose que la longueur d'un haricot vert est une variable aléatoire normale de moyenne 8,1 mm et d'écart type 1,2 mm.
Les haricots verts dont la longueur est comprise entre 5 et 6,5 mm sont dits "extra fins".
Les haricots verts dont la longueur est comprise entre 6,5 et 8 mm sont dits "très fins".
Les haricots verts dont la longueur est comprise entre 8 et 10,5 mm sont dits "fins".
Les autres sont dits hors catégorie.

1. Quelle est la probabilité qu'un haricot vert choisi au hasard :
   (a) soit " extra fin "; (b) soit " très fin "; (c) soit " fin ".
2. On suppose que dans cette récolte 2,8 % des haricots sont hors catégorie.
   (a) Justifier cette affirmation.
   (b) Calculer la probabilité que, sur un échantillon de 10 haricots, il n'y en ait aucun hors catégorie.

florent.arnal@u-bordeaux.fr

III Intervalles de confiance et Tests

Exercice 10 :
On suppose que la masse des pommes d'une production est distribuée suivant une loi normale d'écart-type $\sigma = 6$ g. Dans des conditions données, on a pesé un échantillon de 10 pommes et la moyenne observée est de 185 g.

1. Déterminer une estimation par intervalle de confiance du poids moyen des pommes de la population étudiée (risque 5 %).
2. Combien faudrait-il contrôler de pommes pour que l'intervalle de confiance du poids moyen soit déterminé avec une amplitude de 4 g au maximum ?

Exercice 11 :
Une machine produit des plaquettes de beurre de quantité nominale 250 g.
Afin de vérifier son bon fonctionnement, on prélève un échantillon de 9 plaquettes.
Les masses observées sont les suivants :

```
[1] 252.3 250.7 252.3 252.3 251.4 249.5 250.1 250.7 251.0
```

Proposer un compte-rendu à destination de la direction concernant le fonctionnement de cette machine.
Une annexe avec des résultats R pourra être présentée.

Exercice 12 :
La teneur en alcool du vin doit être mentionnée sur l'étiquette, mais 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 il est indiqué un degré d'alcool de 13%, c'est-à-dire qu'une bouteille peut en contenir entre $12,5\%$ et $13,5\%$.

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
```

Proposer une analyse de ces données au travers d'un compte-rendu, limité à une page, à destination de personnes de l'entreprise ayant des connaissances statistiques.

Exercice 13 :
Le pH est un des éléments parmi les plus importants permettant de caractériser un vin (acidité, vivacité , ...). Afin de contrôler le pH de vins issus d'un même cépage (Merlot), on a effectué des mesures récapitulés ci-dessous :

3,40 3,44 3,46 3,46 3,49 3,54 3,45 3,45 3,47 3,49 3,5 3,51

1. Déterminer une estimation ponctuelle du pH moyen et de la variance associées à ce cépage.
2. Le pH moyen observé est-il significativement inférieur à 3,5 ?
3. Donner une estimation par intervalle de confiance du pH moyen associé à ce cépage.

florent.arnal@u-bordeaux.fr

Exercice 14 :
On considère une population gaussienne de moyenne (ciblée) 1000 et d'écart-type (estimé) $\sigma = 10$. On teste

$$\mathcal{H}_0 : \mu = 1000$$

contre

$$\mathcal{H}_1 : \mu > 1000$$

1. On fait l'hypothèse que la moyenne de la population est, en réalité, égale à 1010.
   Quelle taille d'échantillon préconiseriez-vous pour détecter une telle dérive ?
2. Quelle serait votre conclusion si l'on envisageait que la moyenne de la production est, en réalité, égale à 1001 ?

Exercice 15 :
Un technicien a répété pendant 10 jours le dosage de NTK (dosage de l'azote total Kjeldhal), exprimé en mg/L N, sur des prélèvements d'eau au même point de prélèvement, à la même heure et selon la même technique.
On considère que la teneur en Azote de la source est constante sur la période. Les fluctuations ne sont dues qu'aux manipulations.
On souhaite étudier la reproductibilité de la méthode de dosage.
La reproductibilité est considérée comme suffisante si la variance de la variable aléatoire X prenant pour valeur le dosage de NTK est inférieure à 0,02.
Les résultats obtenus pendant les dix jours consécutifs (exprimés en mg/L N) sont consignés ci-dessous :

```
[1] 1.22 0.78 0.83 1.15 0.94 0.97 0.83 1.10 0.99 0.88
```

À l'aide des données précédentes, tester la reproductibilité de la méthode.

Exercice 16 :
On souhaite comparer deux conditionneuses en prélevant des échantillons de taille 50.
Sur la conditionneuse 1, on a relevé 2 % de défectueux alors que sur la conditionneuse 2, on a relevé 4 % de défectueux.
Peut-on considérer que ces conditionneuses ont des taux de défectueux similaires ?

Exercice 17 :
Dans le cadre de la fabrication d'un nouveau produit, on souhaite comparer trois compositions.
Les produits ont été jugés par des consommateurs qui les ont jugés Bon ou Très Bon.
Les données recueillies sont les suivantes :

	Produit A	Produit B	Produit C
Bon	39	52	77
Très Bon	61	48	23

Comparer ces trois compositions.

florent.arnal@u-bordeaux.fr

Exercice 18 : Contrôle sur la moyenne de préemballés pour apposer le **e**

Des agents du contrôle qualité effectuent dans une usine un contrôle sur un lot de briques de lait dont la quantité nominale annoncée est de 1000 ml.

Étant dans le cadre de lot d'effectif supérieur à 100, la législation indique, dans le cadre de contrôles destructifs, de prélever des échantillons de taille 20 et d'effectuer un test unilatéral au seuil de risque 0,5 %. On extrait de ce lot un échantillon de 20 préemballés.

Les contenances observées sont les suivantes :

997,8	998,4	1001,3	998,8	1000,2	998,9	1000	998,1	999,9	997,5
997,3	1000,1	997,3	998,3	999,7	999,5	999,8	999,8	1001,5	998

Peut-on considérer que la moyenne des contenus effectifs de tous les préemballages du lot est conforme à la QN annoncée ?

Exercice 19 :

Afin de mesurer le taux d'humidité de farines, on expérimente un nouvel appareil de mesures par infrarouges (IR). On prélève un échantillon de 11 farines et on mesure leur taux d'humidité avec la méthode par IR.

Par la suite, on mesure le taux d'humidité de ces mêmes farines avec la méthode directe (étuve).

Voici les résultats obtenus (exprimés en %) :

Méthode directe	27,85	28,35	28,45	27,9	27,45	27,4	28,15	27,65	28,35	28,15	27,95
Méthode par IR	27,9	28,4	28,4	27,95	27,4	27,5	28,25	27,5	28,5	28,35	27,7

Peut-on considérer les résultats obtenus avec la méthode par IR et ceux obtenus avec la méthode directe similaires ?

Exercice 20 :

Sur une doseuse pondérale de pots de fromage blanc, on désire vérifier le bon réglage de la machine concernant le poids moyen.

1. On prélève un échantillon de taille 7, **avant réglage**.
   Les résultats, exprimés en grammes, sont les suivants :
   406,2 ; 406,5 ; 407 ; 407,4 ; 408,5 ; 409 ; 411,4.
   (a) Déterminer la moyenne et l'écart-type des masses des pots de fromage blanc de cet échantillon.
   (b) Donner une estimation ponctuelle de la variance des masses des pots de fromage blanc de la fabrication dont est extrait cet échantillon.
2. **Après réglage**, un deuxième échantillon est prélevé.
   Les résultats, exprimés en grammes, sont les suivants :
   407 ; 403,2 ; 402,1 ; 407,4 ; 405,2 ; 405,5 ; 407 ; 408,2 ; 405,6 ; 404,4.
   Peut-on décider que la moyenne de fabrication après réglage est inférieure à la moyenne de fabrication avant réglage ?

florent.arnal@u-bordeaux.fr

Exercice 21 :
Dans une entreprise, on a mis à l'essai deux types de fabrication d'un produit.
On a constaté que, sur 200 produits de la fabrication 1, 24 présentent un défaut et que, sur 200 produits de la fabrication 2, 18 présentent un défaut.
Y a-t-il un type de fabrication plus fiable que l'autre ?

Exercice 22 :
L'étude de l'effet de l'acide citrique sur le brunissement (mesuré par un indice de brun) de pâtes alimentaires donne les résultats suivants :

Doses d'acide citrique		
5 ppm	10 ppm	20 ppm
25,2	22,1	18,4
24,3	23,8	19,5
26,8	21,9	18,9
25,9	22,6	19,9

Afin de déterminer si l'acide citrique a une influence sur le brunissement des pâtes, on décide de mettre en place une ANOVA.

1. Calculer les résidus associés à cette analyse de variance et déterminer la valeur de la SCE_{res}.
2. A l'aide de représentations graphiques, vérifier que les conditions d'application d'une Anova sont vérifiées.
3. Formuler les hypothèses de ce test, préciser la variable de décision et la nature du test.
4. Établir le tableau d'analyse de variance.
5. Conclure.

Exercice 23 :
On désire comparer le travail de trois doseuses pour boîtes de haricots verts de quantité nominale égale à 800 g. Les masses obtenues sont les suivantes :

Doseuse 1	Doseuse 2	Doseuse 3
805,2	801,5	798,9
798,5	804,8	802,1
803,6	795,5	805,4
797,1	797,2	807,8
795,7	805,8	813,1
799,2	792,1	811,2
798,3	793,6	802,6

1. Quel est le facteur étudié ? Quelle est la variable étudiée ?
2. Proposer une représentation graphique des masses des boites associées à ces 3 doseuses.
3. Peut-on considérer ces trois doseuses similaires ?
   Dans le cas contraire, faire des groupes homogènes liés aux doseuses.
4. En tant que responsable de ces trois lignes, que préconiseriez-vous ?

florent.arnal@u-bordeaux.fr

Exercice 24 :
Lors de l'évaluation par l'analyse sensorielle des qualités organoleptiques d'anciennes variétés de pommes (Reinette de Blenheim ; Reinette des Capucins ; Reinette des Flandres ; Cabarette), on a demandé à un jury constitué de 9 personnes de noter le *Croquant* (bruit perçu lorsqu'on croque dans le morceau de pomme), le *Sucré* (évaluation du caractère sucré du jus obtenu lorsqu'on mange le fruit), l'*Acidité* (évaluation du caractère acide du jus obtenu lorsqu'on mange le fruit), la *Jutosité* (perception du jus lorsqu'on a croqué le fruit, qu'on le mâche et qu'on en extrait le jus) et la *Fermeté* (résistance à la mastication.
Les résultats ci-contre donnent pour chaque membre du jury la note globale attribuée à chaque espèce de pomme.

1. On désire mettre en place une anova avec 4 modalités et un plan équilibré avec n répétitions. On estime que l'écart-type (résiduel) est égal à 0,6 et que les effets seront égaux à

$$-0,5 \quad ; \quad 0 \quad ; \quad 0 \quad ; \quad 0,5$$

Combien de répétitions n préconiseriez-vous pour cette expérimentation ?

2. Les résultats présentés dans la suite donnent pour chaque membre d'un jury constitué de 9 personnes la note globale attribuée à chaque espèce de pomme.

Reinette de Blenheim	Reinette des Capucins	Reinette des Flandres	Cabarette
7,6	6,9	6,1	7,1
7,3	6,7	5,8	6,8
7,6	6,4	5,4	6,9
6,2	7,3	5,7	7,0
7,2	7,2	6,0	7,0
7,9	7,1	5,5	6,7
7,4	6,5	6,1	7,0
7,7	6,7	6,2	6,2
7,6	6,9	6,5	6,6

(a) Faire une représentation graphique de ces données avec des couleurs et un titre.
Que peut-on conjecturer ?

(b) Pourquoi ne peut-on pas envisager de réaliser une ANOVA ?

(c) Y a-t-il une différence significative entre les qualités organoleptiques de ces anciennes variétés de pommes ?

(d) Proposer un "classement" de ces variétés.
On présentera une analyse critique des résultats obtenus.

(e) Représenter graphiquement les effets pour ces 4 variétés avec des couleurs et un titre.

Exercice 25 :
On a soumis un vin à deux méthodes d'élevage.
Afin de les comparer, des échantillons de vins ont été soumis à une analyse sensorielle auprès d'un jury d'experts.
Interpréter les résultats ci-dessous.

Méthode 1	13.5	13.5	13.5	14	16	13	14	15
Méthode 2	14	14	14	14.5	14.5	13.5	15	15

florent.arnal@u-bordeaux.fr

IV Carte de contrôle et capabilité

Exercice 26 :
On a prélevé 5 échantillons de 4 fromages.
Les données observées sont les suivantes :

```
   [,1]  [,2]  [,3]  [,4]
1 102.4 100.8  99.2  98.9
2  99.7  99.7  99.6 100.3
3  99.1 100.4  98.8  99.8
4 100.4 100.1 100.8  99.9
5 100.5 101.1  99.3  98.7
```

1. La distribution des masses des fromages de la production est-elle normale ?
2. Représenter graphiquement les données de ces 5 échantillons.
3. Déterminer une estimation de la variance de la production à l'aide du carré moyen.
4. Déterminer, par calcul, une estimation de l'écart-type de la production en utilisant les étendues.
5. Déterminer les limites de contrôle de la carte de Shewart des moyennes.
6. Construire la carte associée aux données en faisant apparaître les limites de surveillance.
7. L'entreprise a fixé une quantité nominale $QN = 100$ g avec des tolérances de 99 à 101 g. Déterminer les indices de capabilité C_p et C_{pk}.

florent.arnal@u-bordeaux.fr