

# Séance de travaux dirigés 2

## Démarche statistique

### Exercice 1 : Caractérisation du lieu de production d'un café

Le problème qui suit est inspiré d'un stage de fin d'études réalisé par une étudiante de la spécialisation *Science des données* du cursus d'ingénieur agro-alimentaire d'Agrocampus.

Un groupe industriel commercialisant du café souhaite comparer les cafés provenant de différents lieux de production à partir de leur profil de composition physico-chimique, dont une des composantes importantes est le taux de matière sèche (DM). Pour cela, il s'appuie sur des données contenant le lieu de production, codé par un entier allant de 1 à 7, de 240 échantillons de café.

Les statistiques descriptives élémentaires pour ces données sont fournies dans le tableau 1.

---

R script

```
> numSummary(dta$DM, statistics=c("mean", "sd", "quantiles"))
```

| mean     | sd        | 0%       | 25%      | 50%      | 75%      | 100%     | n   |
|----------|-----------|----------|----------|----------|----------|----------|-----|
| 90.19918 | 0.5299891 | 88.57223 | 89.84318 | 90.13754 | 90.47799 | 91.93607 | 240 |

Table 1: Statistiques descriptives élémentaires pour les données **café**.

---

Le tableau 2 fournit quelques statistiques descriptives par lieu de production des taux de matière sèche de l'ensemble de l'échantillon.

---

R script

```
> numSummary(dta$DM, statistics=c("mean", "sd", "quantiles"), groups=dta$Localisation)
```

|   | mean     | sd        | 0%       | 25%      | 50%      | 75%      | 100%     | data:n |
|---|----------|-----------|----------|----------|----------|----------|----------|--------|
| 1 | 90.39711 | 0.4377907 | 89.38683 | 90.07922 | 90.50817 | 90.69827 | 91.17519 | 50     |
| 2 | 90.26092 | 0.2480545 | 89.58015 | 90.10104 | 90.28563 | 90.41711 | 90.70547 | 26     |
| 3 | 89.68936 | 0.4082911 | 88.98607 | 89.27865 | 89.79238 | 89.99135 | 90.40554 | 26     |
| 4 | 91.36635 | 0.4547603 | 90.48211 | 91.09946 | 91.31974 | 91.82588 | 91.93607 | 13     |
| 5 | 90.13393 | 0.4137080 | 89.11718 | 89.90425 | 90.11694 | 90.34565 | 91.23802 | 22     |
| 6 | 89.98806 | 0.2815068 | 89.39688 | 89.80103 | 89.99477 | 90.18512 | 90.62641 | 84     |
| 7 | 90.50184 | 0.6358823 | 88.57223 | 90.30511 | 90.53441 | 90.88827 | 91.44958 | 19     |

Table 2: Données **café**: statistiques descriptives par lieu de production des taux de matière sèche de l'ensemble de l'échantillon.

---

1. *Donnez l'expression du modèle statistique des taux de matière sèche de café permettant de tester l'existence de différences moyennes entre les lieux de production. Quels sont les paramètres de ce modèle ?*
2. *Quelles sont les hypothèses  $H_0$  et  $H_1$  du test de l'existence de différences moyennes entre les lieux de production ? Exprimez ces hypothèses à partir des paramètres du modèle de la question précédente.*
3. *Quelles sont les estimations des paramètres mesurant les effets dans ce modèle ?*

La somme des carrés des écarts résiduels, notée RSS, est un indicateur de la qualité d'ajustement du modèle.

4. Que vaut RSS ?
5. En déduire la valeur estimée de l'écart-type résiduel du modèle.
6. Donnez l'expression du modèle nul (associé à l'absence de différences moyennes entre les lieux de production) des taux de matière sèche. Quelle est la valeur estimée des paramètres de ce modèle ?
7. Que vaut la somme des carrés des écarts résiduels  $RSS_0$  mesurant la qualité d'ajustement du modèle nul ?
8. En déduire la valeur du coefficient  $R^2$  du modèle de la question 1.
9. Quelle est l'expression de la statistique de test permettant de tester l'existence de différences moyennes entre les lieux de production ? Quelle est la valeur prise par cette statistique de test ?
10. Quelle est la distribution sous l'hypothèse  $H_0$  de la statistique de test introduite à la question précédente ?

Le tableau 3 donne les probabilités de dépassement de différentes valeurs pour une variable aléatoire distribuée selon la loi donnée à la question précédente.

---

R script

```
> x = seq(from=0.1,to=10,by=0.1)
> # Séquence régulière de valeurs: 0.1, 0.2, ..., 9.9, 10
> round(pf(x,df1=6,df2=233,lower.tail=FALSE),digits=3)
> # Fonction de répartition de la loi de Fisher en x

[1] 0.996 0.977 0.936 0.879 0.808 0.730 0.650 0.571 0.496 0.426 0.363 0.307 0.258 0.215
[15] 0.179 0.148 0.122 0.100 0.082 0.067 0.054 0.044 0.035 0.029 0.023 0.019 0.015 0.012
[29] 0.010 0.008 0.006 0.005 0.004 0.003 0.002 0.002 0.002 0.001 0.001 0.001 0.001 0.000
[43] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[57] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[71] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[85] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[99] 0.000 0.000
```

Table 3: Probabilités de dépassement de différentes valeurs pour une variable aléatoire distribuée selon la loi de la statistique de test sous l'hypothèse  $H_0$ .

---

11. Que vaut la p-value du test de l'existence de différences moyennes entre les lieux de production ? L'effet est-il significatif au seuil  $\alpha = 0.05$  ?
12. Quelle est la plus grande valeur de la statistique de test qui conduirait à ne pas considérer l'effet comme significatif, toujours au seuil  $\alpha = 0.05$  ?