

Séance de travaux dirigés 1

Démarche statistique

Exercice 1 : Variable quantitative ou catégorielle ? Statistique descriptive ou inférentielle ?

1. Déduire de leur intitulé la nature, quantitative ou catégorielle, des variables suivantes : la température, le sexe, la couleur préférée, le nombre de frères, le département de naissance.
2. Pour chacune des problématiques suivantes, dire si l'analyse statistique relève ou non de l'inférence (i.e. une généralisation d'un échantillon à une population). Si oui, définir l'individu statistique, la population et l'échantillon.
 - Un mois avant des élections, on souhaite évaluer les intentions de vote pour le candidat A.
 - On veut étudier la répartition du PIB des Pays de l'Union Européenne.
 - On souhaite évaluer les mobilités professionnelles des français entre régions à partir des résultats d'un recensement général de la population.
 - On veut vérifier que le volume contenu dans des bouteilles est conforme à la valeur nominale inscrite sur le produit.

Exercice 2 : 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. Le tableau 2 donne les taux de matière sèche mesurés dans le lieu de production '5', triés par ordre croissant.

R script

```
> summary(dta)
```

	DM	Localisation
Min.	:88.57	1:50
1st Qu.	:89.84	2:26
Median	:90.14	3:26
Mean	:90.20	4:13
3rd Qu.	:90.48	5:22
Max.	:91.94	6:84
		7:19

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

1. Calculer la moyenne, l'écart-type et la variance des taux de matière sèche du tableau 2.
2. Construire une boîte de dispersion des taux de matière sèche du tableau 2.

On souhaite maintenant s'assurer que la distribution des taux de matière sèche du tableau 2 peut être considérée comme normale. Pour cela, on veut construire un graphe, que l'on appelle **graphe quantiles-quantiles**, permettant

```

> dm5 = dta$DM[dta$Localisation=="5"]
>      # Taux de matière sèche mesurés sur les échantillons
>      # de café provenant du lieu '5'
> sort(round(dm5,digits=2))
>      # Affichage des valeurs arrondies à 2 décimales (fonction round)
>      # et triées par ordre croissant (fonction sort)

[1] 89.12 89.70 89.73 89.85 89.88 89.90 89.92 89.96 90.04 90.07 90.09 90.14
[13] 90.17 90.19 90.24 90.25 90.38 90.46 90.48 90.57 90.57 91.24

```

Table 2: Données **café**: taux de matière sèche mesurés sur les échantillons provenant du lieu de production '5', triés par ordre croissant.

de comparer les quantiles observés des taux de matière sèche aux quantiles théoriques de la loi normale de moyenne nulle et d'écart-type 1, dite aussi **loi normale standard**.

3. Si Z suit une loi normale standard, et u_p est le quantile de cette loi normale standard d'ordre $0 < p < 1$, que vaut $\mathbb{P}(Z \leq u_p)$?

Le tableau 3 donne les quantiles de la loi normale standard pour différentes valeurs de p .

```

> proba = seq(0.05,0.95,0.05) # Séquence : 0.05, 0.10, 0.15, ..., 0.90, 0.95
> qtheo = qnorm(proba)         # Quantiles de la loi normale standard
> names(qtheo) = proba        # Affecte un nom à chaque quantile (nom = ordre du quantile)
> round(qtheo,2)              # Valeurs arrondies à 2 décimales

 0.05   0.1   0.15   0.2   0.25   0.3   0.35   0.4   0.45   0.5   0.55   0.6
-1.64 -1.28 -1.04 -0.84 -0.67 -0.52 -0.39 -0.25 -0.13  0.00  0.13  0.25

 0.65   0.7   0.75   0.8   0.85   0.9   0.95
 0.39   0.52   0.67   0.84   1.04   1.28   1.64

```

Table 3: Quantiles d'ordre $p = 0.05, 0.10, \dots, 0.95$ de la loi normale standard.

4. D'après le tableau 3, quelle est la médiane de la loi normale standard ?

5. Si X suit une loi normale d'espérance μ et d'écart-type $\sigma > 0$, et u_p^* est le quantile de cette loi normale d'ordre $0 < p < 1$, montrer qu'il existe une relation linéaire entre u_p^* et u_p : en d'autres termes, montrer qu'il existe des coefficients a et b tels que $u_p^* = a + bu_p$. Donner les coefficients a et b de cette relation linéaire.

6. D'après la question précédente, si les taux de matière sèche sont distribués selon une loi normale, quelle forme doit prendre le graphe des points d'abscisse u_p et d'ordonnée q_p , où q_p est le quantile observé d'ordre p des taux de matière sèche ?

La figure 1 reproduit le graphe quantiles-quantiles pour les données du tableau 2.

7. D'après le graphe de la Figure 1, peut-on considérer que les taux de matière sèche des cafés du lieu de production '5' sont distribués selon une loi normale ?

Si x_i désigne le taux de matière sèche mesuré sur le i ème café de l'échantillon, on appelle *ième donnée centrée-réduite* la valeur $x_i^* = (x_i - \bar{x})/s_x$, où \bar{x} et s_x sont respectivement la moyenne et l'écart-type de la série (x_1, \dots, x_n) et n est la taille de l'échantillon. Le tableau 4 donne les valeurs centrées réduites des données du tableau 2.

```
> qobs = quantile(dta$DM[dta$Localisation=="5"],probs=proba)
> # Quantiles observés
> plot(qtheo,qobs,pch=16,bty="1",xlab="Quantiles théoriques",
+ ylab="Quantiles observés",main="Graphe Quantile-quantile (loi normale)")
```

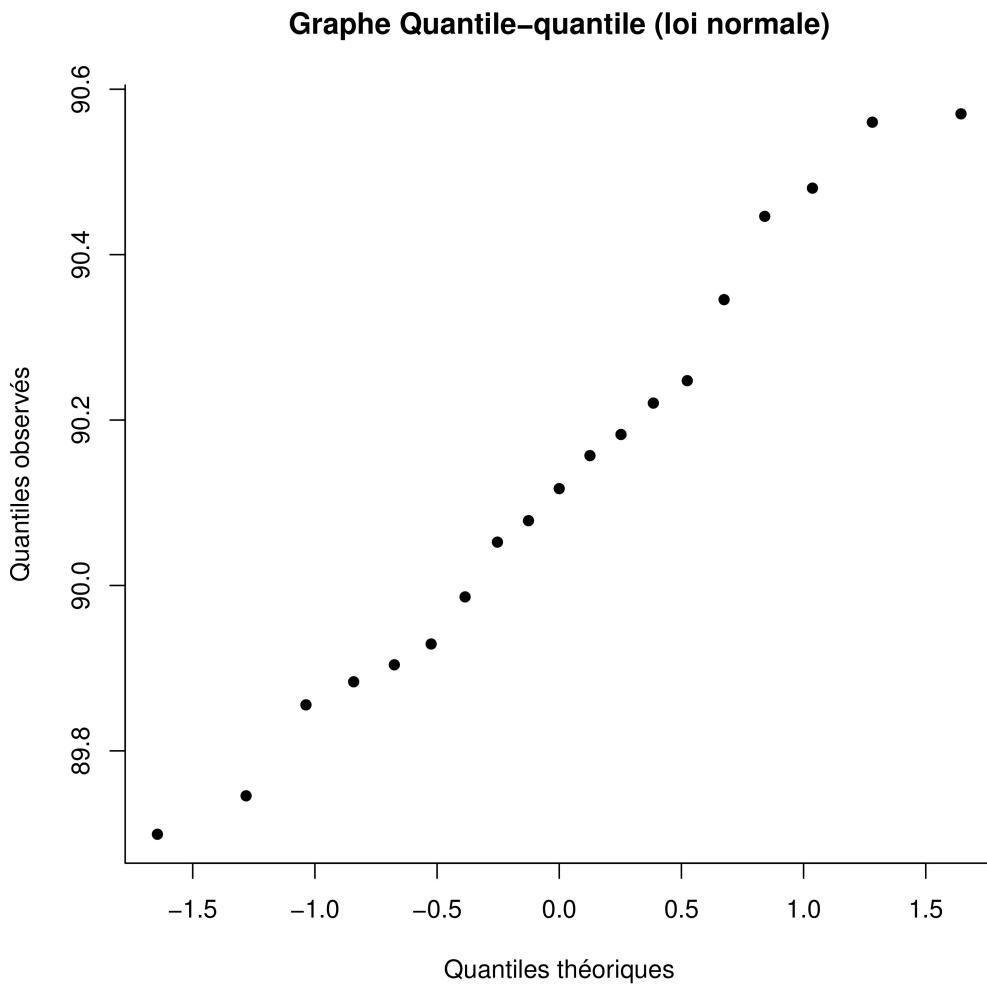


Figure 1: Graphe quantiles-quantiles pour le diagnostic graphique de normalité des données du tableau 2.

```
> dm5.scaled = scale(dm5)[,1] # Taux de matière sèche centrées-réduites
> round(dm5.scaled,2)

[1] -0.62 -0.22  0.08  0.59 -0.11 -0.16 -0.97  0.25  1.05  0.84 -0.68 -1.06 -0.52 -2.46
[15]  0.80  0.14 -0.43  0.29  0.03 -0.57  1.05  2.67
```

Table 4: Données du tableau 2 centrées-réduites

8. Quelle est la moyenne des données du tableau 4 ? Que vaut leur écart-type ?
9. D’après le tableau 4, y-a-t-il un ou plusieurs échantillons de café dont le taux de matière sèche peut paraître anormalement élevé, ou au contraire anormalement bas ? Si oui, lesquels ?

