# TP n° 12 : Introduction aux tests statistiques

**Exercice 1.** Le Bureau Communautaire de Référence de la Communauté Européenne établit des références pour les traces inorganiques dans les denrées alimentaires. Ainsi, pour la quantité de zinc présent dans le pain complet, la moyenne de référence est de 19.5 microgrammes par gramme. Un expérimentateur prélève 8 échantillons de pain complet dans une production et mesure la teneur en zinc dans chacun d'entre eux. Les résultats sont :

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21.1 | 21.5 | 21 | 21.6 | 19.1 | 20.3 | 22.2 | 21.5 |

On suppose que la teneur de zinc en microgrammes présent dans le pain complet peut être modélisée par une *var* $X$ suivant une loi normale.

La problématique est la suivante : Peut-on affirmer, avec un faible risque de se tromper, que la production de pain complet n'est pas conforme aux règles européennes concernant la teneur en zinc ?

1. Expliciter les hypothèses à tester.
2. Déterminer la p-valeur associée en utilisant la commande `t.test`.
3. Répondre à la problématique (*on précisera le degré de significativité s'il y a rejet de* $H_0$*).*
4. *Questions complémentaires.* Que vaut la statistique de test observée notée $t_{obs}$ ? Que vaut $\mathbb{P}(|T| \geq |t_{obs}|)$, où $T$ est une *var* suivant loi de Student à $\nu = 7$ degrés de liberté ?

**Exercice 2.** Un fabriquant de gâteaux commercialise ses produits avec sur l'emballage la mention "Teneur moyenne en lipides inférieure ou égale à 20 grammes". Le fabriquant veut contrôler sa production car il pense que ses gâteaux ne respectent plus l'indication notée sur l'emballage. Il prélève au hasard dans sa production 7 gâteaux. Les résultats sont :

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | 23 | 23 | 23 | 22 | 20 | 23 |

On suppose que la teneur en lipides en grammes d'un gâteau peut être modélisée par une *var* $X$ suivant une loi normale.

La problématique est la suivante : Peut-on affirmer, avec un faible risque de se tromper, que le fabriquant à raison ?

1. Expliciter les hypothèses à tester.
2. Déterminer la p-valeur associée.
3. Répondre à la problématique (*on précisera le degré de significativité s'il y a rejet de* $H_0$*).*
4. *Questions complémentaires.* Que vaut la statistique de test observée notée $t_{obs}$ ? Que vaut $\mathbb{P}(T \geq t_{obs})$, où $T$ est une *var* suivant loi de Student à $\nu = 6$ degrés de liberté ?

**Exercice 3.** On dit qu'un citron est de calibre C si son diamètre est compris entre 6.50 et 7.30 centimètres. Un producteur affirme que la majorité des citrons sortant de sa production sont de calibre $C$. Pour vérifier cette affirmation, un contrôleur extrait au hasard 40 citrons sortant de cette production et mesure leurs diamètres.

Les résultats, en centimètres, sont :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6.71 | 6.43 | 7.74 | 6.51 | 6.79 | 7.32 | 6.65 | 6.78 | 7.22 | 6.42 |
| 5.82 | 6.18 | 6.45 | 5.95 | 7.86 | 7.11 | 7.02 | 6.80 | 6.93 | 6.90 |
| 6.16 | 6.79 | 6.82 | 6.13 | 5.93 | 6.95 | 6.63 | 6.44 | 6.18 | 7.03 |
| 7.48 | 7.21 | 6.05 | 6.66 | 6.77 | 6.08 | 6.75 | 6.74 | 6.25 | 7.24 |

La problématique est la suivante : Est-ce que le contrôleur peut affirmer, avec un faible risque de se tromper, que le producteur a raison ?

1. Expliciter les hypothèses à tester.
2. Déterminer la p-valeur associée en utilisant la commande `prop.test` (sans correction de Yates).
3. Répondre à la problématique (*on précisera le degré de significativité s'il y a rejet de* $H_0$*).*

**Exercice 4.** L'objectif de cet exercice est d'introduire en douceur les principales commandes R sur la régression linéaire simple, ainsi que les tests statistiques élémentaires associés. On considère le jeu de données `cars` disponible dans R.

1. Charger les données, comprendre d'où elles émanent et afficher les noms des variables considérées, à savoir, `speed` et `dist`.
2. Tracer le nuage de points constitué des points de coordonnées (`speed`, `dist`).
3. Comme le nuage de points est étiré dans une direction, on peut envisager un lien linéaire entre `dist` et `speed`. Ainsi, afin d'expliquer les valeurs de `dist` en fonction de celles de `speed`, on considère le modèle de régression linéaire simple :

$$\texttt{dist} = \beta_0 + \beta_1 \texttt{speed} + \epsilon.$$

   Dans cette écriture, $\beta_0$ et $\beta_1$ sont 2 réels inconnus que l'on cherche à estimer avec les données, et $\epsilon$ est une *var* modélisant une somme d'erreurs naturelles.

   La méthode des moindres carrés ordinaires nous permet d'estimer $\beta_0$ et $\beta_1$, et aussi de mesurer l'impact de `dist` sur `speed` par le biais de tests statistiques.

   Reproduire les commandes suivantes :

```
reg = lm(cars$dist ~ cars$speed)
summary(reg)
```

   Cela renvoie, entre autre, l'estimation (ponctuelle) de $\beta_0$: $-17.5791$, l'estimation (ponctuelle) de $\beta_1$: 3.9324 et on constate que le rejet de l'hypothèse $H_0$: $\beta_1 = 0$ est "hautement significatif" $(***)$, entraînant que l'impact de `speed` sur `dist` est aussi "hautement significatif" $(***)$ (la p-valeur associée vérifie $< 0.001$).

4. La droite qui ajuste au mieux le nuage de points est celle d'équation : $y = -17.5791 + -17.5791x$. Cette droite nous permet de faire des prévisions: si on se fixe une valeur $x_0$ pour `speed`, la valeur moyenne pour `dist` sera $y_0 = -17.5791 + -17.5791x_0$. Visualiser cette droite à l'aide de la commande :

```
abline(reg, col = "blue")
```