

Séance de travaux dirigés 5

Démarche statistique

Exercice 1 : Test de dégustation de bières

Par un test de dégustation, on a recueilli les notes de perception d'acidité pour 4 bières blanches (voir figure 1) par 32 juges (chaque juge n'a rendu qu'une seule évaluation). Ces notes sont données sur une échelle allant de 0 à 10, 0 traduisant une absence totale d'acidité et 10 traduisant au contraire une acidité extrême.

R script

```
> with(boxplot(Acidite~Biere,cex.lab=1.4,cex.axis=1.4,xlab="Bière",
+ ylab="Note d'acidité"),data=dta)
```

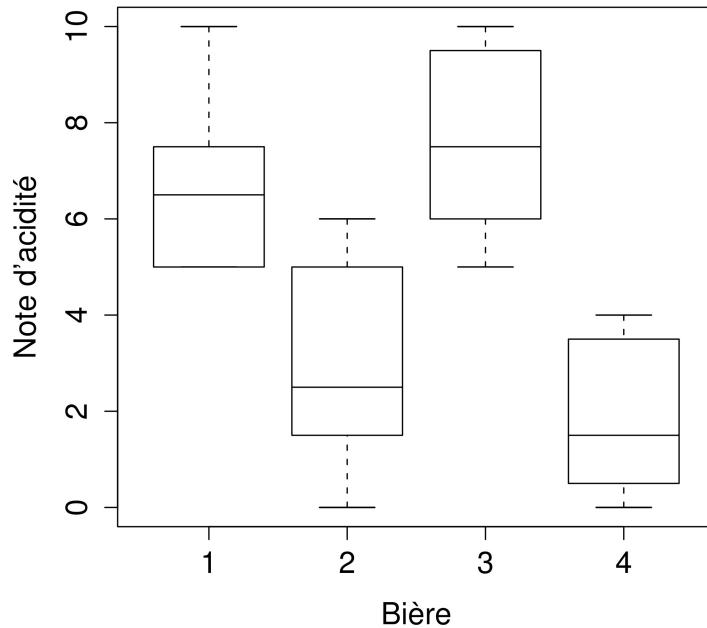


Figure 1: Lien entre l'acidité perçue et la bière.

L'objectif de cette dégustation est de savoir si les différences de perception d'acidité de ces bières doivent être considérées comme significatives.

1. Quel modèle statistique des notes d'acidité permet de répondre à cette question ?²
2. Complétez les colonnes *Df*, *Mean Sq* et *F value* de la table d'analyse de variance 1 du modèle de la question 1 en justifiant brièvement vos réponses.
3. Quelle est la valeur du coefficient R^2 de ce modèle ?

Le tableau 2 donne les probabilités de dépassement des entiers de 1 à 30 pour une variable aléatoire distribuée selon une loi de Fisher dont les degrés de liberté $df1$ et $df2$ et sont donnés dans le tableau 1.

4. D'après le tableau 2, complétez la colonne *Pr(>F)* de la table d'analyse de variance 1 par une valeur approchée de la *p-value* du test. Peut-on considérer que la différence entre les perceptions d'acidité de ces bières est significative au seuil de 5% ?

Le tableau 3 donne les coefficients estimés du modèle de la question 1, pour trois choix possibles de la modalité de référence du facteur *Biere*. A partir de ce tableau, on cherche à tester les différences d'acidité perçue entre deux bières, pour tous les couples possibles de bières.

```
> acid.lm = lm(Acidite~Biere,data=dta)
> anova(acid.lm)
```

Analysis of Variance Table

Response: Acidite

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Biere	??	184.84	??	??	??
Residuals	??	100.62	??		

Table 1: Table d'analyse de la variance pour le test de comparaison des bières selon leur acidité perçue.

```
> vecF = seq(from=1,to=30,by=1) # Séquence des valeurs entières entre 1 et 30
> vecF
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
[30] 30

> pf(vecF,df1,df2,lower.tail=FALSE) # Probabilités de dépassement de vecF pour une
> # loi de Fisher à df1 et df2 degrés de liberté
[1] 4.073603e-01 1.368282e-01 4.730923e-02 1.727495e-02 6.679959e-03 2.727557e-03
[7] 1.170989e-03 5.262362e-04 2.465153e-04 1.199225e-04 6.037902e-05 3.136910e-05
[13] 1.677253e-05 9.207823e-06 5.179311e-06 2.979463e-06 1.749988e-06 1.047897e-06
[19] 6.388638e-07 3.960791e-07 2.494404e-07 1.594167e-07 1.032983e-07 6.780928e-08
[25] 4.506054e-08 3.029123e-08 2.058622e-08 1.413590e-08 9.802228e-09 6.860649e-09
```

Table 2: Probabilités de dépassement des entiers de 0 à 30 pour une variable aléatoire distribuée selon une loi de Fisher dont les degrés de liberté sont donnés dans le tableau 1.

5. Pour comparer les acidités perçues entre deux bières, pour tous les couples possibles de bière, combien de tests doit on réaliser ?

6. Quel seuil peut-on définir pour le risque de l'erreur de 1ère espèce de chacun de ces tests pour garantir un risque inférieur à 5% de commettre plus d'une erreur sur l'ensemble des tests en rejetant l'hypothèse nulle ? Finalement, quelles sont les bières perçues différemment de manière significative ?

Exercice 2 : Garantir une teneur en matière grasse minimum

Une entreprise agro-alimentaire qui fabrique des produits laitiers souhaite vérifier que la crème qu'elle produit a bien, au minimum, les 15 % de matière grasse annoncés sur l'étiquette. En effet, lors de la fabrication de cette crème, les quantités d'eau et de lait mélangées sont de plusieurs dizaines de milliers de litres et le débit et l'arrêt des vannes sont gérés manuellement. Ainsi, un retard dans la fermeture des vannes entraîne un surplus de l'un ou l'autre des ingrédients.

Vingt échantillons de crème sont prélevés et la teneur en matière grasse est mesurée. La moyenne des 20 teneurs est $\bar{x} = 14.91$ et l'écart-type $s_x = 0.59$.

1. Donnez une estimation de l'écart-type de la moyenne des 20 teneurs en matières grasses.

²Donnez l'expression mathématique de ce modèle, ses paramètres et les postulats sur ce modèle.

```

> summary(acid.lm)$coefficients

            Estimate Std. Error    t value    Pr(>|t|)    
(Intercept)   6.625   0.6702378  9.884551 1.242940e-10  
Biere2        -3.625   0.9478594 -3.824407 6.715252e-04  
Biere3         1.000   0.9478594  1.055009 3.004401e-01  
Biere4        -4.750   0.9478594 -5.011292 2.689024e-05  

> tmp = dta                                # tmp = copie de dta
> biere = relevel(dta$Biere, "2")           # biere = copie de dta$Biere
> tmp$Biere = biere                         # Biere2 = modalité de référence de tmp$Biere
> acid.lm = lm(Acidite~Biere,data=tmp) # Ajustement du modèle à partir des données tmp
> summary(acid.lm)$coefficients

            Estimate Std. Error    t value    Pr(>|t|)    
(Intercept)   3.000   0.6702378  4.476023 1.159853e-04  
Biere1        3.625   0.9478594  3.824407 6.715252e-04  
Biere3        4.625   0.9478594  4.879416 3.856548e-05  
Biere4       -1.125   0.9478594 -1.186885 2.452457e-01  

> biere = relevel(dta$Biere, "3")           # biere = copie de dta$Biere
> tmp$Biere = biere                         # Biere3 = modalité de référence de tmp$Biere
> acid.lm = lm(Acidite~Biere,data=tmp) # Ajustement du modèle à partir des données tmp
> summary(acid.lm)$coefficients

            Estimate Std. Error    t value    Pr(>|t|)    
(Intercept)   7.625   0.6702378 11.376559 5.188826e-12  
Biere1        -1.000   0.9478594 -1.055009 3.004401e-01  
Biere2        -4.625   0.9478594 -4.879416 3.856548e-05  
Biere4        -5.750   0.9478594 -6.066300 1.531370e-06

```

Table 3: Coefficients du modèle de la question 1, pour trois choix possibles de la modalité de référence du facteur Biere..

On cherche à savoir si la teneur en matière grasse moyenne sur l'ensemble de la production, notée μ , est au minimum de 15 %.

2. Exprimez cette problématique sous la forme d'un test d'hypothèses dont vous donnerez les hypothèses nulle et alternative.

3. Donnez l'expression, en fonction de \bar{x} et s_x , de la statistique de test T à utiliser pour répondre à cette question.

La loi sous l'hypothèse nulle de la statistique de test de la question précédente est une loi de Student à k degrés de liberté dont les quantiles sont donnés dans le tableau 4.

4. Que vaut k ?

5. Quelle est la plus grande valeur t^* telle que, si $T \leq t^*$, alors on rejette l'hypothèse nulle du test de la question 1 au seuil de 5% (choisir parmi les valeurs ci-après) ?

- $t^* = 0$
- $t^* = -2.093$
- $t^* = -1.729$

```

> proba = seq(from=0.005,to=0.500,by=0.005)
> # proba = séquence régulière de valeurs dans [0.005;0.500]
> proba

[1] 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050 0.055 0.060 0.065
[14] 0.070 0.075 0.080 0.085 0.090 0.095 0.100 0.105 0.110 0.115 0.120 0.125 0.130
[27] 0.135 0.140 0.145 0.150 0.155 0.160 0.165 0.170 0.175 0.180 0.185 0.190 0.195
[40] 0.200 0.205 0.210 0.215 0.220 0.225 0.230 0.235 0.240 0.245 0.250 0.255 0.260
[53] 0.265 0.270 0.275 0.280 0.285 0.290 0.295 0.300 0.305 0.310 0.315 0.320 0.325
[66] 0.330 0.335 0.340 0.345 0.350 0.355 0.360 0.365 0.370 0.375 0.380 0.385 0.390
[79] 0.395 0.400 0.405 0.410 0.415 0.420 0.425 0.430 0.435 0.440 0.445 0.450 0.455
[92] 0.460 0.465 0.470 0.475 0.480 0.485 0.490 0.495 0.500

> round(qt(proba,df=k),digits=3)
> # Quantiles associés de la loi de Student à k ddl
> # Arrondis à 3 décimales

[1] -2.861 -2.539 -2.346 -2.205 -2.093 -2.000 -1.920 -1.850 -1.786 -1.729 -1.677
[12] -1.628 -1.583 -1.540 -1.500 -1.462 -1.426 -1.392 -1.359 -1.328 -1.297 -1.268
[23] -1.240 -1.213 -1.187 -1.161 -1.136 -1.112 -1.088 -1.066 -1.043 -1.021 -1.000
[34] -0.979 -0.958 -0.938 -0.918 -0.899 -0.880 -0.861 -0.842 -0.824 -0.806 -0.789
[45] -0.771 -0.754 -0.737 -0.720 -0.704 -0.688 -0.671 -0.656 -0.640 -0.624 -0.609
[56] -0.593 -0.578 -0.563 -0.548 -0.533 -0.519 -0.504 -0.490 -0.475 -0.461 -0.447
[67] -0.433 -0.419 -0.405 -0.391 -0.377 -0.364 -0.350 -0.337 -0.323 -0.310 -0.297
[78] -0.283 -0.270 -0.257 -0.244 -0.231 -0.218 -0.205 -0.192 -0.179 -0.166 -0.153
[89] -0.140 -0.127 -0.115 -0.102 -0.089 -0.076 -0.064 -0.051 -0.038 -0.025 -0.013
[100] 0.000

```

Table 4: Quantiles de la loi de Student à k dégrés de liberté.

- $t^* = 1.729$
- $t^* = 2.093$

6. Finalement, les données incitent-elles à suspecter le process de ne pas fournir en moyenne des produits ayant une teneur en matière grasse d'au moins 15 % ? Que vaut la p-value de ce test (choisir parmi les valeurs ci-après) ?

- approximativement 0.250
- 0.05
- approximativement 0.682
- 0.025

La réglementation permet d'écrire sur l'emballage *teneur minimale de 15 %* à condition que l'entreprise puisse démontrer que, si la teneur moyenne μ sur l'ensemble de la production valait en réalité 14.95%, alors le test mis en place permettrait de détecter ce non-respect de la valeur nominale avec une probabilité, notée π , au moins égale à 0.90.

7. Parmi les formulations suivantes de π , laquelle est exacte ?

- $\pi = \mathbb{P}_{\mu=14.95}(T \leq t^*)$

- $\pi = \mathbb{P}_{\mu=14.95}(|T| \geq |t^*|)$
- $\pi = \mathbb{P}_{\mu=15}(T \leq t^*)$
- $\pi = \mathbb{P}_{\mu=15}(|T| \geq |t^*|)$

Lorsque $\mu = 14.95$, la loi de la statistique de test est une loi de Student à k degrés de liberté non-centrée, dont le paramètre de non-centralité peut être estimé par $\lambda = \sqrt{20}(14.95 - 15)/s_x$. Le tableau 5 donne les quantiles de cette loi.

R script

```
> lambda = sqrt(20)*(-0.05/0.59)
> round(qt(proba,df=k,ncp=lambda),digits=3)

[1] -3.326 -2.989 -2.786 -2.639 -2.523 -2.426 -2.343 -2.270 -2.205 -2.146 -2.091 -2.041
[13] -1.994 -1.951 -1.909 -1.870 -1.833 -1.798 -1.764 -1.732 -1.701 -1.671 -1.642 -1.614
[25] -1.587 -1.561 -1.535 -1.511 -1.486 -1.463 -1.440 -1.418 -1.396 -1.374 -1.353 -1.333
[37] -1.312 -1.293 -1.273 -1.254 -1.235 -1.217 -1.198 -1.181 -1.163 -1.145 -1.128 -1.111
[49] -1.094 -1.078 -1.061 -1.045 -1.029 -1.013 -0.998 -0.982 -0.967 -0.951 -0.936 -0.921
[61] -0.906 -0.892 -0.877 -0.863 -0.848 -0.834 -0.820 -0.805 -0.791 -0.778 -0.764 -0.750
[73] -0.736 -0.723 -0.709 -0.696 -0.682 -0.669 -0.655 -0.642 -0.629 -0.616 -0.603 -0.590
[85] -0.576 -0.563 -0.551 -0.538 -0.525 -0.512 -0.499 -0.486 -0.473 -0.460 -0.448 -0.435
[97] -0.422 -0.409 -0.397 -0.384
```

Table 5: Quantiles de la loi de Student à k degrés de liberté non-centrée de paramètre de non-centralité λ .

8. Quelle est la valeur de π (choisir parmi les valeurs ci-après) ?

- $\pi = 0.05$
- $\pi = 0.10$
- $\pi = 0.90$
- $\pi = 0.95$

