Département
Licence

MOSE2014 Probabilités et statistiques
Mathématiques TP machine 4
Ph. Thieullen

TP machine IV
(Statistique inférentielle : tests d'hypothèse)

Le TP en entier est à rendre. N'oubliez pas de commencer votre script par `# TP4.R` *et les noms de chaque étudiant :* `# noms, prenoms` *formant le binôme ou le trinôme.*

– Récupérer un fichier de données sur internet par la commande suivante (sans ajouter de retour à la ligne à la chaîne de caractères)

```
adresse_fichier <-
      "http://www.math.u-bordeaux1.fr/
      ~thieulle/Data/2012_001.txt"
poids <- read.table(adresse_fichier, header=TRUE,
      sep="", quote="", dec=",")
```

Bien comprendre l'aide en ligne de `?read.table` et ses paramètres `header` pour une première ligne de texte, `sep` pour la séparation des colonnes, `quote` pour les données non numériques, `dec` pour la convention des nombres décimaux. Le résultat `poids` est un `data.frame` ou une liste de vecteurs de même longueur.

– Le fichier donne le poids de petits pots de caramel ou de chocolat. Extraire d'abord quelques informations

```
cat("affichage des 10 premières lignes : \n")
print(poids[1:10,])
cat("noms des colonnes : \n")
print(names(poids))
cat("nombre de petits pots : \n")
print(length(poids[,1]))
cat("résumé statistique : \n")
print(summary(poids))
```

– Afficher sur une même figure les deux histogrammes (on rappelle qu'on accède aux objets d'une liste par `$`)

```
hist(poids$caramels, ...)
```

```
hist(poids$chocolats, ...)
```

On devra trouver la figure 1 et on n'oubliera pas d'annoter les graphiques.

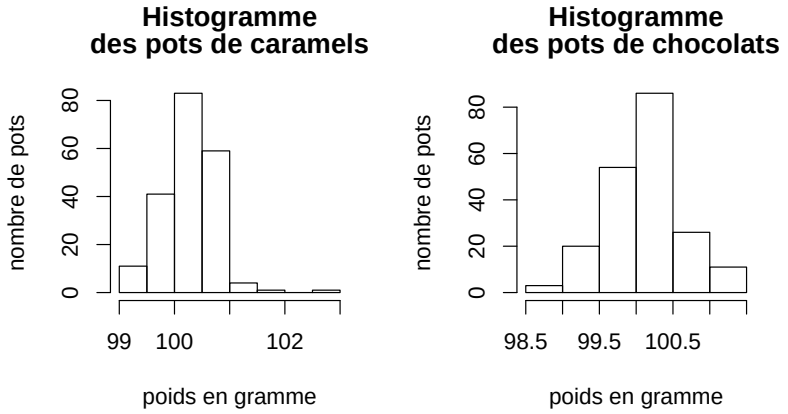

FIGURE 1 – Histogramme des poids de petits pots lactés

Remarquer comme les deux histogrammes sont différents.

– Le fabriquant produit par heure 200 pots et garantit que le poids dans chacun des cas est de 100 g. Faire un test sur l'hypothèse nulle « le poids est inférieur à 100 g » d'abord sur tous les pots produits pendant une heure (tout le fichier). Utiliser `t.test` en allant voir l'aide en ligne et formater les réponses sous la forme

```
alpha <- 0.005
test_caramels <- t.test(poids$caramels, ...)
cat("test sur 200 pots de caramels : \n")
print(test_caramels)
```

Quelle est la p-valeur du test ?

– Il est possible que le contrôle de ces pots soit destructif (le fabriquant contrôle la composition en même temps). Le fabriquant ne peut pas se permettre alors de détruire la production d'une heure. Il choisit au hasard 20 pots parmis ces 200 et fait le test sur cet échantillon. Pour choisir au hasard, utiliser `sample` et son aide en ligne.

```
poids_caramels <- sample(poids$caramels, ...)
alpha <- 0.005
test_caramels <- t.test(poids_caramels, ...)
cat("Test sur 20 pots de caramels : \n")
print(test_caramels)
```

Remarquer maintenant comme la p-valeur a augmenté.

– Pour bien comprendre ce que fournit `t.test`, calculer la p-valeur de l'échantillon de 20 pots en utilisant la formule du cours

$$t_{crit} = \sqrt{n}\,\frac{\bar{x} - 100}{s_{n-1}}, \quad p_{valeur} = \mathbb{P}(\mathcal{T} > t_{crit}).$$

où $\mathcal{T}$ est une loi de Student de $ddl = n - 1$. On écrira les calculs sous la forme suivante et on comparera avec le dernier `print(test_caramels)`.

```
t_crit <- ...
p_valeur <- ...
cat("p-valeur du cours : \n")
cat("t_crit : ", t_crit, "\n")
cat("p-valeurs : ", p_valeur, "\n")
```

Faire tourner cette dernière partie du programme (y compris le choix des 20 pots de caramels) plusieurs fois en remarquant que t_{crit} et p_{valeur} changent à chaque fois. Comment l'expliquez-vous ?

– On réalise maintenant un test d'hypothèse sur la différence des moyennes des poids. On prendra là aussi, d'abord les 200 pots, puis deux échantillons de 20 pots de caramel et chocolat pris au hasard. On choisira entre le cas apparié et le cas indépendant. Formuler par exemple les réponses sous la forme

```
poids_chocolats <- sample(poids$chocolats, ...)
alpha <- 0.005
test_poids <- t.test(poids$caramels, poids$chocolats, ...)
cat("Test de différence de moyenne sur 200 pots : \n")
print(test_poids)
test_poids <- t.test(poids_caramels, poids_chocolats, ...)
cat("Test de différence de moyenne sur 20 pots : \n")
print(test_poids)
```

On notera que les moyennes des poids des 200 pots diffèrent à une erreur de 1.4‰, alors que pour 20 pots, ces moyennes peuvent différer avec une grande p_{valeur} (par exemple de 70%) montrant qu'en fait elles coïncident.

– L'histogramme des poids des 200 pots de chocolats suggère que la répartition est gaussienne. On cherche à réaliser un test du chi-deux d'ajustement. On commence par créer un tableau de poids par classes. On prendra

$$]98.5, 99], \quad]99, 99.5], \quad \ldots \quad ,]101, 101.5].$$

On utilisera `seq` pour générer les classes, et `cut` pour récupérer les effectifs par classes. Le tableau des effectifs par classes est donné par la nouvelle structure `table`. La fonction `levels` récupère le nom des classes. On rédigera la réponse suivant le modèle suivant.

```
cat("Test d'ajustement avec la loi gaussien \n")
mu <- mean(...) # moyenne des 200 pots de chocolat
sigma <- sd(...) # écart-type
cat("moyenne : ",mu,"\n")
cat("sigma : ", sigma,"\n")
# Dans la commande suivante, créer 7 valeurs limites
# de 98.5 à 101.5 pour obtenir 6 classes
classes <- seq(...)
facteurs <- cut(poids$chocolats, classes)
nom_classes <- levels(facteurs)
dist_chocolats <- table(facteurs)
print(dist_chocolats)
```

On doit obtenir le tableau suivant

$]98.5, 99]$	$]99, 99.5]$	$]99.5, 100]$	$]100, 100.5]$	$]100.5, 101]$	$]101, 101.5]$
3	20	54	86	26	11

On crée ensuite un vecteur de distribution des 200 poids comme si ces poids suivaient la loi normale $\mathcal{N}(\mu, \sigma)$. On complètera le code

```
dist_normale = vector(mode="numeric", length=6)
dist_normale[1] = pnorm(99,mu,sigma)
dist_normale[2] = ...
dist_normale[3] = ...
dist_normale[4] = ...
dist_normale[5] = ...
dist_normale[6] = 1 - pnorm(101,mu,sigma)
names(dist_normale) <- nom_classes
print(length(poids$chocolats)*dist_normale)
```

On trouvera (après arondi à la première décimale)

$]-\infty, 99]$	$]99, 99.5]$	$]99.5, 100]$	$]100, 100.5]$	$]100.5, 101]$	$]101, +\infty]$
2.1	16.8	55.6	74.8	40.9	9.9

Faire enfin un test d'ajustement du chi-deux

```
test_ajust <- chisq.test(dist_chocolats, p=dist_normale)
print(test_ajust)
```

Refaire ces calculs en utilisant les formules du cours

$$d_{crit} = \sum_{i=1}^{r} \frac{(N_i - np_i^0)^2}{np_i^0}, \quad p_{valeur} = \mathbb{P}(\chi^2 > d_{crit}),$$

où χ^2 est un chi-deux à $r-1$ degrés de liberté. On complètera le code

```
cat("p-valeur du cours : \n")
d_crit <- ...
ddl <- ...
p_valeur <- 1 - pchisq(..., ...)
cat("d_crit : ", d_crit, "\n")
cat("p_valeur : ", p_valeur, "\n")
```

On comparera avec les valeurs données par `chisq.test`.