Tests statistiques

1 Intervalle de confiance

1.1 Pour une proportion

On est ici dans le cas d'une population répartie en deux catégories. On a vu qu'à partir d'un échantillon de n individus l'intervalle de confiance d'une proportion p se calcule approximativement grâce à la loi normale, et donne approximativement :

$$i_p = \left[f - \sqrt{\frac{f(1-f)}{n-1}} u_\alpha,\ f + \sqrt{\frac{f(1-f)}{n-1}} u_\alpha \right].$$

Prenons un exemple : sur 20 individus, il y a 5 mâles (soit $f = 0.25$). Calculons l'intervalle de confiance pour $\alpha = 5\%$:

$$i_p = \left[0.25 - \sqrt{\frac{0.25 \times 0.75}{19}} \times 1.96,\ 0.25 + \sqrt{\frac{0.25 \times 0.75}{19}} \times 1.96 \right] \simeq [0.055,\ 0.445].$$

Ne nous trompons pas, ce calcul est très approximatif du fait de la faible taille de l'échantillon. Voyons plutôt une méthode infaillible, que l'on peut utiliser quelque soit l'effectif. Il s'agit simplement du calcul de la probabilité exacte basé sur la loi binomiale. En effet, c'est ce que fait R. Pour calculer l'intervalle, on utilise la commande `binom.test(x,n)$conf.int` où `x` est le nombre d'individus de la catégorie qui nous intéresse et `n` l'effectif total. Voici ce qui se passe dans le cas de notre exemple :

```
> binom.test(5,20)$conf.int
[1] 0.08657147 0.49104587
attr(,"conf.level")
[1] 0.95
```

On trouve donc ici $i_p = [8.7\%,\ 49.1\%]$. On obtient un résultat différent de notre calcul approximatif.
Par défaut, R effectue les tests à 5%. On peut modifier ceci avec l'option `conf.level`.
Remarque : Dans le cas de grands échantillons, on peut retrouver la valeur approximative donnée par la loi normale à l'aide de la fonction `binom.approx` du package `epitools` (on le charge à l'aide de la commande `require(epitools)`).

1.2 Pour une moyenne

Dans le cours, on a considéré deux cas :

- la variable étudiée suit une loi normale.
- l'échantillon est de grande taille ($n > 30$).

Dans le premier cas, on a

$$i_\mu = \left[\bar{x} - \frac{s_c}{\sqrt{n}} t_\alpha,\ \bar{x} + \frac{s_c}{\sqrt{n}} t_\alpha \right].$$

Dans le second cas, quelle que soit la loi suivie par la variable étudiée, la moyenne calculée sur notre échantillon de grande taille suit approximativement une loi normale (d'après le théorème central limite). On avait ainsi comme intervalle de confiance :

$$i_\mu = \left[\bar{x} - \frac{s_c}{\sqrt{n}} u_\alpha,\ \bar{x} + \frac{s_c}{\sqrt{n}} u_\alpha \right].$$

On considère un échantillon de taille 10 représentant la taille d'étudiants de sexe masculin :

```
> men = c(172.5,175,176,177,177,178.5,179,179,179.5,180)
```

On a un petit échantillon mais la taille de l'homme suit une loi normale. Pour calculer l'intervalle de confiance, on utilise la fonction `t.test(men)$conf.int` :

```
> t.test(men)$conf.int
[1] 175.6719 179.0281
attr(,"conf.level")
[1] 0.95
```

On obtient $i_\mu = [175.7,\ 179.0]$.

1.3 Pour une variance

Dans le cas de la variance, si la variable aléatoire suit une loi normale, on a l'intervalle de confiance suivant :

$$i_{\sigma^2} = \left[\frac{n-1}{b_\alpha}s_c^2,\ \frac{n-1}{a_\alpha}s_c^2\right].$$

Contrairement à la moyenne, il n'y a pas de fonctions qui permettent de calculer directement l'intervalle de confiance. Il faut donc "faire le calcul à la main" en calculant les deux quantiles d'ordre 0.975 et 0.025 à l'aide de la fonction `qchisq`. On peut également créer sa propre fonction, par exemple :

```
conf.int.sigma2 = function(x,sc2=0,alpha=0.05){
  n = length(x)
  if (n!=1) {
    sc2 = var(x)
  }else{
    n=x
  }
  b_inf = (n-1)/qchisq(1-alpha/2,n-1)*sc2
  b_sup = (n-1)/qchisq(alpha/2,n-1)*sc2
  ic_sigma2 = c(b_inf,b_sup)
  return(ic_sigma2)
}
```

2 Comparaisons

On va résumer sous forme de tableau les principales commandes de tests :

Commande	Description
`t.test(x)`	Permet de réaliser un test de Student de comparaison d'une espérance inconnue à une valeur de référence, où la variance est inconnue.
`binom.test(x)`	Permet de réaliser un test exact de comparaison d'une proportion à une valeur de référence.
`prop.test(x)`	• Permet de réaliser un test de comparaison d'une proportion à une valeur de référence pour de grands échantillons ($np > 10$ et $n(1-p) > 10$). • Permet un test de comparaison de deux proportions inconnues pour de grands échantillons.
`t.test(x,y)`	Permet de réaliser un test de comparaison de deux espérances inconnues où les variances sont égales (Test de Student) ou inégales (Test de Welsh) sur des populations indépendantes ou non.

Commande	Description
`var.test(x,y)`	Permet de réaliser un test de comparaison de deux variances inconnues (Test de Fisher -Snedecor).
`fisher.test(x,y)`	Permet de réaliser un test de comparaison de deux proportions inconnues.

Notons que l'on a toujours besoin d'une des hypothèses suivantes :

- la variable étudiée suit une loi normale.
- l'échantillon est de grande taille ($n > 30$).

Remarques importantes:

- On précise quelle comparaison on effectue ($\neq$, $<$ ou $>$) à l'aide de l'option `alternative`. Les valeurs possibles sont `"two.sided"` (valeur par défaut), `"less"` ou `"greater"`.
- On n'a pas de fonction pour la comparaison d'une variance inconnue à une variance de référence. Comme pour l'intervalle de confiance, il est nécessaire de faire les calculs "à la main".
- Lors d'un test de Student pour comparer deux espérances, on précise que les variances sont égales à l'aide du paramètre `var.equal=T` .
- Dans le cas d'un test de Student pour séries appariées, on utilise le paramètre `paired=T`.
- Il est possible de tester la normalité d'une distribution avec la fonction `shapiro.test`.

R ne donne pas les résultats comme dans le cours. Il renvoie en sortie la `p-value` (notons la p). On interprète ensuite les résultats de la façon suivante (avec les notations du cours) :

- si $p \leq \alpha$, on rejette $\mathcal{H}_0$.
- sinon on ne rejette pas $\mathcal{H}_0$.

3 Tests du Khi-deux

3.1 Test de conformité

On se placera sous les mêmes hypothèses que dans le cours, càd $np_i \geq 5$. Si ce n'est pas le cas, il faudra regrouper les classes. Pour effectuer le test, on utilise la fonction `chisq.test`. Attention, R exige que les valeurs théoriques soient données sous forme des proportions attendues (entre 0 et 1) alors que les valeurs observées doivent être données sous forme d'effectifs (entre 0 et l'infini). Prenons un exemple : Parmi les 89 individus obtenus lors d'un croisement, la répartition des homozygotes et hétérozygotes est 25 **AA**, 59 **Aa** et 5 **aa**. Les proportions attendues sont respectivement 0.25, 0.5 et 0.25. On a bien ici $np_i \geq 5$.

```
> observed = c(25,59,5)
> theo = c(0.25,0.5,0.25)
> chisq.test(observed,p=theo)
        Chi-squared test for given probabilities

data:  observed
X-squared = 18.4382, df = 2, p-value = 9.913e-05
```

La mention `for given probabilities` indique que c'est bien un test de conformité à une distribution théorique que l'on a effectué. Ici la p-value est très inférieure à 5% (d'où rejet de $\mathcal{H}_0$). Notre population ne suit donc pas la loi Mendélienne.

3.2 Test d'indépendance

On reste toujours dans le cadre du cours ($np_{i,j} \geq 5$). On utilise la même fonction que précédemment (`chisq.test`) mais on ne rentre que des observations sous la forme d'un tableau. Voyons un exemple (exercice 6 du cours):

```
> be_cool=c(30,15,15)
> be_aware=c(10,5,15)
> be_zen=c(15,10,35)
> table=rbind(be_cool,be_aware,be_zen)
> colnames(table)=c("moins","autant","plus")
> table
         moins autant plus
be_cool     30     15   15
be_aware    10      5   15
be_zen      15     10   35
> chisq.test(table)

        Pearson's Chi-squared test

data:  table
X-squared = 14.5542, df = 4, p-value = 0.005721
```

La p-value est inférieure à 5%, les lignes et colonnes ne sont pas indépendantes, ici le niveau de stress dépend de la méthode de relaxation.

4 Tests non paramétriques

Voici un tableau qui donne les fonctions pour réaliser les tests non paramétriques :

Fonction	Description
`shapiro.test(x)`	Permet de réaliser un test de normalité de Shapiro-Wilk
`wilcox.test(x,y)`	Permet de réaliser un test de comparaison de moyennes de deux populations. Si les populations sont indépendantes : test de Mann et Witney, si elles sont appariées : test de Wilcoxon (avec le paramètre `paired = TRUE`)
`kruskal.test(liste)`	Permet de réaliser un test de comparaison de moyennes de k populations
`cor.test(x,y)`	Permet de réaliser un test d'association/corrélation entre deux échantillons appariés, utilisant la méthode de Pearson, Kendall ou Spearman

5 Exercices

Exercice 1 :

Un étudiant en deuxième année de licence de biologie s'intéresse à un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Il mesure la quantité de toxine par gramme de solution. Il a obtenu les neufs mesures suivantes, exprimées en milligrammes :

1.2 0.8 0.6 1.1 1.2 0.9 1.5 0.9 1.0

Ces mesures seront supposées être les réalisations de variables aléatoires indépendantes et identiquement distribuées suivant la loi normale d'espérance μ et d'écart-type σ.

1. Donnez une estimation ponctuelle de l'espérance μ et de l'écart-type σ de la quantité de toxine par gramme de la solution.
2. Déterminez un intervalle de fonction à 95% pour l'espérance μ de la quantité de toxine par gramme de la solution.
3. Déterminez un intervalle de fonction à 95% pour la variance σ^2 de la quantité de toxine par gramme de solution.

Exercice 2 :

Pour comparer l'effet de la vitamine C du jus d'orange et de l'acide ascorbique de synthèse, on a donné, pendant 6 semaines, du jus d'orange à un groupe de 10 cobayes et de la vitamine de synthèse à un groupe de 10 autres cobayes. Puis on a mesuré la longueur des odontoblastes des incisives. Les résultats suivants ont été relevés :

- jus d'orange :

8.2 9.4 9.6 9.7 10.0 14.5 15.2 16.1 17.6 21.5

- acide ascorbique :

4.2 5.2 5.8 6.4 7.0 7.3 10.1 11.2 11.3 11.5

Testez, au risque $\alpha = 5\%$, l'hypothèse $\mathcal{H}_0$: l'effet des deux produits est le même contre $\mathcal{H}_1$: le jus d'orange accélère plus la croissance que l'acide ascorbique. Pour répondre à cette question vous allez suivre la démarche suivante :

1. Rentrez les données en créant deux vecteurs. Vous nommerez le premier `jus_orange` et le second `acide_ascorbique`.
2. Testez ensuite la normalité de chacun des deux échantillons (fonction `shapiro.test`) et concluez.
3. A partir des conclusions obtenues des deux tests de normalité, pouvez-vous envisager la démarche du cours : à savoir réaliser un test de Student pour répondre à la question qui vous est posée ? Réaliseriez-vous un test unilatéral ou bilatéral ?
4. Dans le cadre du test de Student, il y a deux cas à différencier : soit les variances des deux populations sont égales soit elles ne le sont pas. Pour répondre à cette question, réaliser un test de Fisher-Snedecor.
5. D'après le résultat obtenu à la dernière question, adaptez le test de Student correspondant à cette réponse et concluez.

Exercice 3 :

Chez un groupe de 10 sujets, les effets d'un traitement destiné à diminuer la pression artérielle ont été expérimentés. Les résultats (valeur de la tension artérielle systolique en cmHg) ont été relevés sur les 10 sujets et sont présentés dans le tableau ci-dessous.

Sujet n°	1	2	3	4	5	6	7	8	9	10
Avant traitement	15	18	17	20	21	18	17	15	19	16
Après traitement	12	16	17	18	17	15	18	14	16	18

Le traitement a-t-il une action significative au risque 5% ? Pour répondre à cette question, vous allez suivre la démarche suivante :

1. Rentrez les données en créant deux vecteurs. Vous nommerez le premier `avnt` et le second `aprs`. Quels sont les modes de chacun des deux vecteurs ? Quelles sont les tailles de chacun des deux vecteurs ?

2. Créez un vecteur `diff` en faisant la différence entrez le vecteur `aprs` et le vecteur `avnt` et affichez à l'écran le vecteur `diff`.

3. La variable aléatoire que vous étudierez ici est la différence des tensions artérielles. Savez-vous pourquoi vous allez étudier la variable aléatoire `diff` ?

4. Avant d'appliquer le test paramétrique adéquat, effectuez le test de normalité.

5. Pour répondre à la question initialement posée, quel est le test que vous allez utiliser ? Quel type de test choisissez-vous ? Unilatéral ou bilatéral ?

6. Réalisez le test en précisant les deux hypothèses et la statistique du test. Puis concluez c'est-à-dire répondez à la question "le traitement a-t-il une action significative, au risque $\alpha = 5\%$?"

Exercice 4 :

Le tableau suivant donne la répartition de 10000 personnes en fonction de leur groupe sanguin et de leur facteur Rhésus.

	O	A	B	AB	Totaux
Rh_+	3570	3825	935	170	8500
Rh_-	630	675	165	30	1500
Totaux	4200	4500	1100	200	10000

Les deux caractères, groupe sanguin et facteur Rhésus sont-ils indépendants ?

Remarques :

- On peut convertir un objet de classe `matrix` en un objet de classe `table` à l'aide de la fonction `as.table`.
- On peut représenter les données (en rentrant les données dans un objet appelé `Rhesus`) :

  ```
  plot(Rhesus,main="Dénombrements")
  ```

- On peut calculer les marges (appelées totaux dans le tableau) avec la fonction `margin.table`.

Exercice 5 :

160 boules ont été sélectionnées dans une urne à quatre couleurs dont la répartition est inconnue. Les boules ont été reportées dans le tableau suivant :

Couleurs	Noir	Rouge	Jaune	Vert	Totaux
Proportion théorique	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	1
Effectif observé	100	18	24	18	160

A partir de cet échantillon, vous voulez savoir s'il est possible que la répartition des couleurs dans l'urne soit de 9/16 de boules noires, 3/16 de boules rouges, 3/16 de boules jaunes et 1/16 de boules vertes.

Exercice 6 :

Un "nez" donne une note de qualité à $n = 10$ parfums. Les scores x_i sont notés de 1 à 10 (10 étant la meilleure note) et les prix des parfums correspondants y_i sont présentés dans le tableau suivant :

Parfum	1	2	3	4	5	6	7	8	9	10
Qualité	10	1	2	5	4	3	6	7	9	8
Prix	95	60	52.5	51.5	49.5	47.5	55	48	56	53

Tester le coefficient de Spearman au seuil de 5% (fonction `cor.test()`). Même question avec le coefficient de Kendall.

Exercice 7 :

La concentration d'un produit dans les eaux de deux rivières fait l'objet d'un contrôle journalier :

```
qt1=c(5.34,5.01,5.14,5.02,5.35,5.17,5.11,5.26,5.0)
qt2=c(5.33,5.13,5.16,5.09,5.49,5.32,5.24,5.23,5.6)
```

1. Existe-t-il une différence dans les eaux de ces deux rivières ? Utiliser un test de Mann et Whitney.
2. On considère maintenant que ces données ont été mesurées, le même jour, sur la même rivière, en amont (qt1) et en aval (qt2) d'une usine. La concentration est-elle supérieure en aval de l'usine ? Utiliser un test de Wilcoxon (Ne pas oublier de modifier les options `paired` et `alt`).

Exercice 8 :

Le jeu de données `airquality` de R contient des données relatives à la qualité de l'air à New-York du 1er mai 1973 au 30 septembre 1973. La question que l'on se pose est la suivante : Peut-on considérer que la quantité d'ozone présente dans l'air est distribuée de façon similaire pour les 5 mois de l'étude ?

1. Représenter graphiquement les données relatives à chaque moi de l'étude (diagramme, boîte à moustache...), et interpréter.
2. Que pouvez-vous dire graphiquement de l'égalité des moyennes, médianes et variances ?
3. Afin de répondre à la question, tester l'égalité des moyennes de chaque mois à l'aide du test de Kruskal-Wallis.
4. Si les 5 moyennes ne sont pas toutes égales, effectuer alors les tests 2 à 2.
5. Analyser la normalité des données pour chaque groupes. Aurais-t-on pu utiliser des tests paramétriques pour répondre à la question d'égalité des moyennes ? Si oui, faites-le.