Université de Lausanne — Institut des Sciences Sociales

Statistique I non-psy.
Cours du professeur André Berchtold

Corrigé de la seconde série d'exercices supplémentaires

Exercices 10.3, 11.1, 12.1, 13.6, 14.2, 15.8, 16.4

Exercice 10.3

Les observations sont distribuées selon une loi inconnue, mais comme l'échantillon est de grande taille, le théorème central limite assure que la moyenne de l'échantillon est bien distribuée selon une loi normale de paramètres μ et σ^2/n.
Etant donné que nous supposons $\sigma^2 = s^2$, nous pouvons utiliser la statistique suivante :

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0,1)$$

L'intervalle de confiance s'écrit alors

$$\begin{aligned} \mu &= \bar{x} \pm z_{1-\alpha/2}\, \sigma_{\bar{x}} \\ &= \bar{x} \pm z_{0.975}\, \frac{\sigma}{\sqrt{n}} \\ &= 76400 \pm 1.96 \cdot \frac{5250}{10} \\ &= 76400 \pm 1029 \\ \longrightarrow \quad & [75371; 77429] \end{aligned}$$

Exercice 11.1

1. On teste $H_0 : \mu = \mu_0 = 50$ contre $H_1 : \mu = \mu_1 > 50$. L'écart-type de la population étant supposé connu, on utilise la statistique de test suivante :

   $$Z_0 = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \sim N(0,1)$$

   Il s'agit d'un test unilatéral à droite. Pour un risque $\alpha = 0.05$, le seuil de rejet vaut 1.645. La zone de rejet de l'hypothèse nulle s'écrit alors

   $$R = \{z_0^{calc} | z_0^{calc} \geq 1.645\}$$

D'après l'échantillon,

$$\begin{aligned} \bar{x} &= 646.4/12 = 53.8\bar{6} \\ \sigma_{\bar{x}} &= \sigma/\sqrt{n} = 5/\sqrt{12} = 1.4434 \end{aligned}$$

d'où

$$z_0^{calc} = \frac{53.8\bar{6} - 50}{1.4434} = 2.68$$

Etant donné que la valeur calculée, 2.68, est supérieure au seuil de rejet, 1.645, l'hypothèse nulle du test est rejetée. On peut donc admettre que la moyenne des taux d'acceptation pour l'ensemble des communes de la population est significativement supérieure à 50%.

2. Dans le cas d'un test pour la moyenne avec variance connue, la taille d'échantillon n'intervient pas dans le calcul du seuil de rejet. Celui-ci reste donc fixé à 1.645. En revanche, l'écart-type de la moyenne devient

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 5/\sqrt{100} = 0.5$$

et la valeur calculée devient

$$z_0^{calc} = \frac{53.8\bar{6} - 50}{0.5} = 7.73$$

Dans ce cas, l'hypothèse nulle est encore plus fortement rejetée qu'au point précédent. Ceci est normal, car une même différence entre la moyenne de l'échantillon et l'hypothèse nulle est maintenant établie sur la base d'un plus grand échantillon, donc de plus d'information.

Exercice 12.1

1. Il s'agit d'effectuer un test de Student pour données indépendantes. Afin de savoir quelle version de ce test doit être appliquée, nous commençons par effectuer un test de l'égalité des variances :

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ contre } H_1 : \sigma_X^2 \neq \sigma_Y^2$$

A partir des échantillons, on estime les variances des populations comme suit :

$$\begin{aligned} \hat{\sigma}_X^2 &= \frac{n}{n-1} s_X^2 = \frac{4}{4-1} 2.04 = 2.72 \\ \hat{\sigma}_Y^2 &= 2.45\bar{6} \end{aligned}$$

La statistique de test vaut alors

$$F = \frac{2.72}{2.45\bar{6}} = 1.1072$$

Cette statistique est distribuée selon une loi de Fisher à 3 et 3 degrés de liberté. En prenant un risque $\alpha = 10\%$, le seuil de rejet de H_0 vaut 9.28. L'hypothèse nulle d'égalité des variances est donc acceptée.

On effectue un test de Student pour données indépendantes et variances égales :

$$H_0 : \mu_Y - \mu_X = 0 \text{ contre } H_1 : \mu_Y - \mu_X > 0$$

L'écart-type de la différence des moyennes vaut

$$\begin{aligned}
\hat{\sigma}_{\bar{Y}-\bar{X}} &= \sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right) \frac{n_X s_X^2 + n_Y s_Y^2}{n_X + n_Y - 2}} \\
&= \sqrt{\left(\frac{1}{4} + \frac{1}{4}\right) \frac{4 \cdot 2.04 + 4 \cdot 1.8425}{4 + 4 - 2}} \\
&= 1.1376
\end{aligned}$$

La statistique de test vaut alors

$$t_0^{calc} = \frac{6.25 - 5}{1.1376} = 1.10$$

Cette statistique est distribuée selon une loi de Student à 4+4-2=6 degrés de liberté. Pour un risque $\alpha = 10\%$, le seuil de rejet du test vaut 1.44. L'hypothèse nulle est donc acceptée : Le changement de nom n'a pas entraîné une hausse des ventes du yoghourt.

2. Si nous supposons que les données sont appariées (c'est-à-dire que les mêmes 4 hypermarchés ont été étudiés à deux reprises), nous pouvons construire une variable $D = Y - X$ représentant l'augmentation observée entre X et Y dans chaque hypermarché. Cette variable prend les valeurs 1.5, 0, 2.8 et 0.7 pour une moyenne $\bar{d}$=1.25 et une variance s_d^2=1.0825. On effectue alors le test suivant :

$$H_0 : \mu_D = 0 \text{ contre } H_1 : \mu_D > 0$$

La statistique de test se calcule comme

$$t_0^{calc} = \frac{\bar{d}}{\hat{\sigma}_{\bar{d}}} = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n-1}}} = \frac{1.25}{\sqrt{\frac{1.0825}{3}}} = 2.08$$

Cette statistique est distribuée selon une loi de Student à 4-1= 3 degrés de liberté. Pour un risque $\alpha = 5\%$, le seuil de rejet du test unilatéral à droite vaut 1.638. L'hypothèse nulle est donc rejetée et l'on peut admettre que le changement de nom du yogourt a entraîné une hausse des ventes de celui-ci.

Remarque : Les deux parties de l'exercice amènent à des conclusions inverses, mais cela n'est pas faux, car les données ne sont pas considérées de la même manière (elles sont une fois indépendantes et une fois appariées) et la méthode de test utilisée n'est donc pas exactement la même. Cet exercice montre l'importance de choisir la bonne méthode de test en fonction des données afin d'arriver à une conclusion correcte !

Exercice 13.6

Pour les patients atteints d'un cancer de l'estomac, la moyenne de l'échantillon vaut 283.6 et la variance vaut 140'449. La somme des carrés résiduelles peut être calculée à partir des variances de chaque groupe de patients :

$$SC_{res} = 10 \cdot 1'484'600 + 10 \cdot 62'210 + 10 \cdot 211'080 + 10 \cdot 140'449 = 18'983'390$$

La somme des carrés expliquée est obtenue par soustraction :

$$SC_{exp} = 28'667'000 - 18'983'390 = 9'683'610$$

Nous construisons la table ANOVA suivante :

SC	dℓ	MC	F
SC_{exp} = 9'683'610	3	3'227'870	6.12
SC_{res} = 18'983'390	36	527'316	
SC_{tot} = 28'667'000	39	735'051	

La statistique de test est distribuée selon une loi de Fisher à 3 et 36 degrés de liberté. Pour un risque $\alpha = 5\%$, le seuil de rejet exact ne figure pas dans les tables usuelles. En revanche, on dispose des seuils de rejet 2.84 correspondant à une loi avec 3 et 40 degrés de liberté, et 2.92 correspondant à une loi avec 3 et 30 degrés de liberté. Le vrai seuil de rejet se trouve donc entre ces deux valeurs.
Etant donné que la statistique calculée à partir des données vaut 6.12, elle est de toute façon nettement supérieure au seuil de rejet (même si celui-ci est plus proche de 2.92 que de 2.84) et l'hypothèse nulle est rejetée. On eput donc admettre que le temps de survie n'est pas identique dans les 4 groupes de patients.

L'hypothèse nulle ayant été rejetée, un test de Tukey est effectué pour comparer deux à deux les 4 groupes. Les hypothèses de chaque comparaison s'écrivent

$$H_0 : \mu_i = \mu_j \text{ contre } H_1 : \mu_i \neq \mu_j \,, \text{ pour } i \neq j$$

Tous les groupes étant de la même taille, l'écart-type de chaque comparaison vaut

$$\sigma_q = \sqrt{\frac{527'316}{10}} = 229.63$$

Comme pour l'ANOVA, le seuil de rejet de chaque test, $q_{0.05;36;4}$ ne se trouve pas dans la table usuelle du test de Tukey. En revanche, on trouve les deux valeurs les plus proches suivantes : $q_{0.05;40;4}$=3.79 et $q_{0.05;30;4}$=3.84. Le vrai seuil de rejet se situe donc entre ces deux valeurs et ce degré de précision est suffisant dans le cadre de cet exercice. On peut construire la table de comparaison suivante :

Comparaison	Différence	q	Conclusion
Sein vs Bronches	1207.1	5.26	Sein $\neq$ Bronches
Sein vs Estomac	1180	5.14	Sein $\neq$ Estomac
Sein vs Colon	951.7	4.14	Sein $\neq$ Colon
Colon vs Bronches	255.4	1.11	Colon = Bronches
Colon vs Estomac	228.3	0.99	Colon = Estomac
Estomac vs Bronches	27.1	0.12	Estomac = Bronches

En conclusion, on peut dire que la durée de survie en cas de cancer du sein est en moyenne supérieure aux trois autres durées de survie. En revanche, il n'y a pas de différence significative entre les cancers de l'estomac, du colon et des bronches.

Exercice 14.2

Nous effectuons un test bilatéral de la somme des rangs de Wilcoxon. Les hypothèses s'écrivent

$$H_0 : \text{med(Asie)} = \text{med(Europe)} \text{ contre } H_1 : \text{med(Asie)} \neq \text{med(Europe)}$$

Asie		Europe	
rangs	valeurs	valeurs	rangs
1.5	4		
		1.5	4
3.5	5		
		3.5	5
6	6		
		6	6
		6	6
		8.5	7
		8.5	7
W=11			

Par ordre décroissant, nous avons W_i = 3(3+6+1)-11 = 19. La statistique de test vaut donc 11.
Pour un risque $\alpha = 5\%$, le seuil de rejet du test vaut 7 d'après la table de Wilcoxon. La valeur calculée étant supérieure au seuil de rejet, l'hypothèse nulle est acceptée. On peut donc dire que l'âge médian d'entrée à l'école primaire est similaire en Asie et en Europe.

Exercice 15.8

1. Nous effectuons le test suivant :

   $$H_0 : \text{Indépendance entre sexe et première demande contre } H_1 : \text{Dépendance}$$

   La statistique de test, K^2=6.38, est distribuée selon une loi du chi-2 à (ℓ-1)(c-1)=(3-1)(2-1)=2 degrés de liberté. Pour un risque $\alpha = 10\%$, le seuil de rejet du test vaut 4.61. La valeur calculée étant supérieure à ce seuil, l'hypothèse nulle est rejetée et l'on peut admettre que dans la région Rhône-Alpes, la première demande n'est pas la même selon que l'on est une fille ou un garçon.
2. Soit le tableau de probabilité suivant tel qu'observé dans le Nord-Pas-de-Calais :

0.27	0.4
0.15	0.1
0.05	0.03

Ce tableau est utilisé comme distribution théorique de probabilité. En multipliant ce tableau par l'effectif total de l'échantillon de la région Rhône-Alpes (500), on obtient le tableau suivant des effectifs théoriques :

135	200	335
75	50	125
25	15	40
235	265	500

La statistique du chi-2 calculée entre ce dernier tableau et le tableau observé pour Rhône-Alpes vaut K^2=19.93. Il y a 5 degrés de liberté (6 catégories - 1 contrainte). Pour $\alpha = 10\%$, le seuil de rejet du test vaut alors 9.24. L'hypothèse d'égalité entre les distributions de Rhône-Alpes et du Nord-Pas-de-Calais est donc rejetée.

Remarque : Etant donné que nous n'effectuons pas ici un test d'indépendance, mais un test de comparaison d'une distribution observée (Rhône-Alpes) avec une distribution "théorique" (Nord-Pas-de-Calais), les sommes des lignes et des colonnes ne sont pas identiques entre les deux tableaux et la seule contrainte concerne le nombre total d'observations (500).

Exercice 16.4

Soit X le nombre total de médecins et Y la proportion de médecins généralistes (en %). A partir des informations fournies dans l'énoncé, nous pouvons calculer les éléments suivants qui seront utiles par la suite :

	X	Y
Moyenne	501.46	41.46
Variance	346′269.02	106.30
Covariance	−3′595.15	
Corrélation	−0.5926	

1.

$$\hat{b} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{-3'595.15}{346'269.02} = -0.0104$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 41.46 + 0.0104 \cdot 501.46 = 46.18$$

Le modèle de régression s'écrit alors

$$Y = 46.18 - 0.0104\, X + e$$

2. Le coefficient de détermination se calcule comme

$$R^2 = r_{XY}^2 = (-0.5926)^2 = 0.35$$

Ce modèle ne permet d'expliquer que 35% de la variable dépendante Y, ce qui est insuffisant pour en faire un bon modèle.

3. Nous effectuons le test suivant :

$$H_0 : b = 0 \text{ contre } H_1 : b \neq 0$$

L'écart-type de la pente vaut

$$\hat{\sigma}_{\hat{b}} = \frac{\hat{\sigma}_e}{\sqrt{n \cdot \text{Var}(X)}} = \frac{8.64}{\sqrt{26 \cdot 346'269.02}} = 0.0029$$

La statistique de test se calcule alors comme

$$t_0^{calc} = \frac{\hat{b}}{\hat{\sigma}_{\hat{b}}} = \frac{-0.0104}{0.0029} = -3.59$$

Cette statistique est distribuée selon une loi de Student à n-2=24 degrés de liberté. Pour un risque $\alpha = 5\%$, les seuils de rejet valent -2.064 et 2.064. La statistique calculée étant supérieure au seuil de droite, l'hypothèse nulle du test est rejetée. Nous pouvons donc conclure que le nombre total de médecins d'un canton a bien une relation positive avec la proportion de médecins généralistes de ce même canton.