Exercices

1 Introduction

1.1

Nous nous intéressons aux succursales des différentes chaînes de la grande distribution alimentaire en Suisse. Nous disposons de données récoltées auprès de 8 succursales genevoises de la Migros.

1. Les 8 succursales pour lesquelles nous disposons de données constituent-elles une population ou un échantillon?
2. Peut-on dire que chacune des 8 succursales considérée est une unité statistique?
3. Quelle est la population qui nous intéresse?
4. Dans quelles conditions pourrait-on dire que les 8 succursales considérées sont une population?

2 Lois de probabilité continues

2.1

Soit une variable Z distribuée selon une loi normale centrée-réduite. Calculer la probabilité pour la variable Z de se trouver dans chacun des cas suivants:

1. Entre $z = 0$ et $z = 1.2$.
2. Entre $z = -0.68$ et $z = 0$.
3. Entre $z = -0.46$ et $z = 2.21$.
4. Entre $z = 0.81$ et $z = 1.94$.
5. A gauche de $z = -0.6$.
6. A droite de $z = -1.28$.
7. A droite de $z = 2.05$ et à gauche de $z = -1.44$.
8. A droite de $z = 2.05$ ou à gauche de $z = -1.44$.

2.2

Soit Z, une variable aléatoire suivant une loi normale centrée et réduite $N(0, 1)$.

1. Déterminer la probabilité des événements suivants :

   (a) $Z < 2.14$

   (b) $Z < -2.14$

   (c) $Z \in [-2.14\ ;\ 2.14]$

2. Pour quel seuil a a-t-on $P(Z < a) = p$ lorsque p prend les valeurs suivantes :

   (a) $p = 0.8686$

   (b) $p = 0.9719$

   (c) $p = 0.2912$

Soit X, une variable aléatoire suivant une loi normale $N(2, 1)$.

3. Déterminer les probabilités suivantes :

   (a) $P(X < 3.98)$

   (b) $P(X < 1.5)$

   (c) $P(X \in [2\ ;\ 3])$

4. Pour quelle valeur de a a-t-on :

   (a) $P(X < a) = 0.7454$

   (b) $P(-a + 2 < X < a + 2) = 0.7458$

Soit Y, une variable aléatoire suivant une loi normale $N(2, 9)$.

5. Déterminer les probabilités suivantes:

   (a) $P(Y < 4.25)$

   (b) $P(Y < 1.5)$

   (c) $P(|Y| < 1)$

6. Pour quelle valeur de a a-t-on:
   (a) $P(Y < a) = 0.64$
   (b) $P(a < Y < 2.5) = 0.3788$
   (c) $P(-a < Y < a) = 0.9$

2.3

Un ensemble de scores d'arithmétique obtenus par des enfants de quatrième année présente une moyenne de 30 et un écart type de 6. Un ensemble de scores obtenus par des enfants de neuvième année présente une moyenne de 36 et un écart type de 12. Nous supposons que les distributions sont normales.

1. Faites un schéma approximatif de ces données en représentant les deux groupes sur la même figure.

2. Quel pourcentage des élèves de quatrième année obtiennent-ils de meilleurs scores que la moyenne des élèves de neuvième année?

3. Quel pourcentage des élèves de neuvième année obtiennent-ils de moins bons scores que la moyenne des élèves de quatrième année?

2.4

1. Soit la variable X suivant une loi normale $N(10, 4)$.
   Calculer $P(8 < X < 13)$.

2. Soit la variable Q_{10} suivant une loi du chi-2 à 10 degrés de liberté.
   Déterminer $P(Q_{10} < 12.55)$, $P(Q_{10} > 18.31)$, $P(Q_{10} < 18.31)$ et $P(12.55 < Q_{10} < 18.31)$.
   Trouver la valeur a telle que $P(Q_{10} < a) = 0.975$.

3. Soit la variable T_8 suivant une loi de Student à 8 degrés de liberté.
   Déterminer $P(T_8 < 0.546)$, $P(T_8 < -0.546)$, $P(-0.546 < T_8 < 1.86)$.
   Trouver la valeur b telle que $P(T_8 > b) = 0.6$.

4. Soit la variable $F_{8,5}$ suivant une loi de Fisher-Snedecor à 8 et 5 degrés de liberté.
   Trouver la valeur de c telle que $P(F_{8,5} < c) = 0.95$.

5. Soit la variable $F_{5,8}$ suivant une loi de Fisher-Snedecor à 5 et 8 degrés de liberté.
   Trouver la valeur d telle que $P(F_{5,8} < d) = 0.95$.

6. Soit la variable $F_{12,3}$ suivant une loi de Fisher-Snedecor à 12 et 3 degrés de liberté.
   Trouver la valeur de e telle que $P(F_{12,3} < e) = 0.99$.

7. Soit la variable $F_{3,12}$ suivant une loi de Fisher-Snedecor à 3 et 12 degrés de liberté.
   Trouver la valeur f telle que $P(F_{3,12} < f) = 0.99$.

2.5

Les lois Binomiale, de Poisson, Normale et du Chi-2 sont 4 des lois de distribution statistique les plus utilisées. Pour chacune des 4 situations décrites ci-dessous, choisissez la loi la plus adaptée. Attention: vous devez choisir une et une seule fois chacune des 4 lois proposées.

1. Nous voulons effectuer un test portant sur le rapport entre deux variances.

2. Nous étudions la distribution du nombre mensuel de faillites dans le canton de Genève.

3. Nous nous intéressons au nombre de femmes présentes dans un échantillon de 25 personnes travaillant dans le domaine du graphisme.

4. Nous étudions la distribution du résultat net après impôts en 2005 des sociétés suisses fondées en 2005.

3 Estimation

3.1

Soit Y le nombre de personnes qui assistent à la représentation d'une pièce de théâtre.

1. Pour 5 jours de représentation pris au hasard on a observé:

i	1	2	3	4	5
y_i	300	280	290	310	295

   Estimer l'espérance $(Y) = \mu$ par la médiane de l'échantillon $\dot{Y}$, puis par la moyenne de l'échantillon $\overline{Y}$.

2. Pour 4 nouveaux jours choisis au hasard on a observé :

i	6	7	8	9
y_i	305	318	290	280

   Recalculer les deux estimations de μ en considérant l'échantillon formé par l'ensemble des 9 observations. Commenter.

3. Pour les cinq premières observations on trouve:

$$\sum_{i=1}^{5} y_i^2 = 435'625$$

   et pour l'ensemble des 9 observations:

$$\sum_{i=1}^{9} y_i^2 = 792'274$$

   Donner une estimation non-biaisée de la variance de Y, ainsi qu'une estimation non-biaisée de l'écart type de $\overline{Y}$ dans chacun des deux cas considérés. Commenter.

4. Supposons maintenant qu'à la suite d'une erreur le nombre d'entrées du 9ème jour ait été mal relevé. De ce fait nous avons:

i	1	2	3	4	5	6	7	8	9
y_i	300	280	290	310	295	305	318	290	28

   Recalculer les deux estimations de μ. Lequel des deux estimateurs est-il le plus pertinent?

3.2

Dans une société de services ayant récemment engagé de nombreux nouveaux employés, deux échantillons ont été prélevés, l'un parmi l'ancien personnel et l'autre parmi le nouveau. On a demandé au hasard à 10 anciens employés et à 6 nouveaux d'évaluer sur une échelle allant de 0 à 20 l'efficacité de l'organisation interne de l'entreprise. Les résultats suivants ont été obtenus:

anciens employés	12	4	5	13	9	7	6	12	11	16
nouveaux employés	5	3	4	2	12	7				

Les moyennes respectives des deux échantillons sont 10.2 et 5.5, et les intervalles de confiance correspondants calculés avec un risque α=5% valent $[7.3; 13.1]$ et $[1.7; 9.3]$. Commentez ces données et déterminez si les populations d'anciens et nouveaux employés ont ou non une opinion similaire quant à l'efficacité de l'organisation interne de l'entreprise.

3.3

Les responsables canadiens et français du département vente d'une société produisant des super-ordinateurs désirent comparer leurs résultats mensuels respectifs. Pour ce faire, ils disposent d'informations sur la distribution des variables C: ventes mensuelles au Canada durant les 24 derniers mois d'exploitation et F: ventes mensuelles en France durant les 24 derniers mois d'exploitation:

Intervalle de confiance (α=5%) pour la moyenne: C: $[12; 20]$, F: $[4; 16]$.
Intervalle de confiance (α=5%) pour la variance: C: $[2; 11]$, F: $[12; 28]$.
Intervalle de confiance (α=5%) pour le coefficient de corrélation linéaire: $[0.54; 0.98]$.

A l'aide de ces informations, comparez les variables C et F.

3.4

Dans le but d'estimer le nombre moyen de passagers (conducteur compris) par véhicule automobile circulant sur l'autoroute Genève-Lausanne, un observateur a recueilli les données suivantes:

Nombre de passagers	1	2	3	4	5	6	Total
Effectif	230	248	117	76	14	3	688

Estimer la moyenne μ et la variance σ^2 de la population à l'aide d'intervalles de confiance à 95%.

4 Tests d'hypothèses

4.1

Afin de déterminer si le taux de participation des Suisses aux votations est en augmentation, le test suivant a été réalisé sur la base d'un échantillon représentatif de données:

H_0: le taux de participation est constant sur les 3 dernières années
H_1: le taux de participation a significativement augmenté au cours des 3 dernières années

La p-valeur du test vaut 0.03. Conclure en prenant un risque de première espèce égale à 5%.

4.2

En 2002 et 2003, sur la base de deux échantillons de taille 20, une société d'assurances a signé chaque jour en moyenne respectivement 115 et 108 nouveaux contrats. Un test unilatéral à gauche (α=5%) dont l'hypothèse nulle était: “les nombres moyens de nouveaux contrats signés par jour sont identiques en 2002 et 2003” a conduit à une p-valeur de 0.45. Selon vous, peut-on dès lors raisonnablement admettre que cette société a signé journellement le même nombre de contrats en 2002 et 2003?

4.3

Les données suivantes représentent le taux d'acceptation d'une initiative pour un échantillon de communes en Suisse.

62.3	44.4	49.2	63.3	47.6	60.1
37.4	55.8	57.5	58.3	56.2	54.3

On suppose que l'écart type des taux pour l'ensemble des communes suisses vaut 5.

1. En prenant un risque de première espèce $\alpha = 0.05$, la moyenne μ des taux de la population de communes est-elle significativement supérieure à 50%? Commenter. Formellement, on testera:

$$\begin{aligned} & H_0 : \mu = \mu_0 = 50 \\ \text{contre} \quad & H_1 : \mu = \mu_1 > 50 \end{aligned}$$

2. Soit un échantillon de taille 100 pour lequel on a obtenu la même moyenne que pour l'échantillon précédent de taille 12. Déterminer la région critique du test pour ce nouvel échantillon et comparer avec celle obtenue au point 1. Commenter.

4.4

En 2005, une enquête menée auprès de femmes quant à leur âge x_i à la naissance de leur premier enfant a donné, pour 51 femmes interrogées au hasard, les résultats suivants:

$$\bar{x} = 28 \text{ années}$$

$$ns^2 = \sum(x_i - \bar{x})^2 = 816 \text{ années}^2$$

Les âges x_i représentent les réalisations d'une variable aléatoire X distribuée selon une loi normale (μ, σ^2).

1. On sait qu'en 1990, pour la population, l'âge moyen des femmes à la naissance de leur premier enfant était de 27 ans. Peut-on affirmer, avec un degré de confiance de 95 %, que l'âge moyen a augmenté en 3 ans ? Pour cela, on testera l'hypothèse $H_0 : \mu = 27$ contre l'hypothèse $H_1 : \mu > 27$.

2. Selon certains chercheurs, l'écart type s de l'échantillon sous-estime l'écart type σ de la population, qu'ils évaluent à 5 ans. Refaire le test précédent en supposant que la vraie valeur de σ est égale à 5 et commenter.

3. Tester l'hypothèse $H_0 : \sigma = \sigma_0 = 5$ contre l'hypothèse $H_1 : \sigma = \sigma_1 < \sigma_0$ avec un risque de première espèce de 5%. Commenter.

4.5

Dans une grande entreprise américaine, le salaire annuel moyen des hommes possédant entre 3 et 5 ans d'expérience est de 58'000$. Les salaires (en milliers de dollars) d'un échantillon aléatoire de 10 femmes possédant entre 3 et 5 ans d'expérience sont les suivants:

54 57 61 51 49 56 60 52 45 66

Y a-t-il des preuves attestant de niveaux de salaires différents pour les hommes et les femmes?

4.6

Supposons que

- X suive une distribution normale d'espérance 100 et d'écart type 20 sous H_0;
- X suive une distribution normale d'espérance 80 et d'écart type 20 sous H_1.

Nous optons pour un test unilatéral à gauche.

1. En choisissant une erreur de première espèce $\alpha = 5\%$, que vaut le seuil de rejet r?
2. En choisissant une erreur de première espèce $\alpha = 1\%$, que vaut le seuil de rejet r?
3. Calculer la probabilité de commettre une erreur de seconde espèce en fixant la probabilité de l'erreur de première espèce à $\alpha = 5\%$.
4. Calculer la probabilité de commettre une erreur de seconde espèce en fixant la probabilité de l'erreur de première espèce à $\alpha = 1\%$.

5 Comparaisons paramétriques

5.1

Une firme agro-alimentaire lance sur le marché le nouveau yoghourt “Borc”. Après quelques mois, elle doit se rendre à l'évidence : les ventes sont médiocres. Le service marketing suggère de changer le nom en “Miam”. Quelque temps plus tard, la firme veut savoir si cette modification a porté ses fruits. Elle dispose des données suivantes sur les ventes, en milliers d'unités pour une semaine, relevées dans divers hypermarchés de taille comparable :

Ventes de :	"Borc" (X)	"Miam" (Y)
	7	8.5
	5.6	5.6
	3.2	6
	4.2	4.9
Moyennes	5	6.25
Variances	2.04	1.8425

1. On suppose que les ventes proviennent d'hypermarchés tirés au hasard à chaque période. Tester (avec un risque de première espèce de 10%) si le changement de nom a amené une hausse significative des ventes.

2. Effectuer le même test en considérant que les données sont appariées.

5.2

Le tableau suivant donne les taux de syndicalisation en pourcents d'entreprises choisies au hasard dans l'industrie d'une part et les services d'autre part:

Industrie	50	60	60	50	
Services	30	30	50	80	55

1. Déterminer si la variance des taux de syndicalisation est identique dans les populations "Industrie" et "Services".

2. Tester la différence des moyennes en utilisant un test de Student approprié.

5.3

Nous voulons comparer les effets de deux anesthésiques X et Y qui procurent respectivement des durées de sommeil aléatoires de moyennes μ_X et μ_Y. Notre hypothèse dit que X et Y sont également efficace. Pour vérifier cette hypothèse, nous allons effectuer un test bilatéral comparant les moyennes, en supposant que les variances des populations sont inégales. Nous disposons de 11 observations indépendantes, 6 pour l'échantillon X et 5 pour Y:

X	8.00	8.97	8.32	8.07	7.86	8.37
Y	9.36	8.47	9.04	8.10	8.71	

Conclure pour un degré de confiance de 95%.

5.4

Un analyste au service du personnel d'une société de distribution se demande quels sont les traits de personnalité qui font qu'un vendeur est un bon vendeur. En particulier, il veut savoir si le fait d'être extraverti est un atout. Il choisit alors 20 vendeurs connus comme excellents et 32 vendeurs médiocres, et il leur fait passer à tous un test d'extraversion. Les scores obtenus sont les suivants, les valeurs les plus hautes marquant l'extraversion la plus élevée:

Bons vendeurs: 12, 17, 20, 19, 11, 9, 7, 4, 12, 15, 13, 18, 20, 16, 15, 16, 18, 13, 11, 10.

Vendeurs médiocres: 12, 7, 9, 13, 15, 17, 12, 11, 13, 10, 9, 8, 7, 15, 13, 6, 5, 5, 13, 15, 17, 19, 18, 20, 19, 17, 13, 16, 8, 6, 7, 8.

Utilisez le listing suivant, produit par SPSS, pour tester l'hypothèse selon laquelle l'extraversion n'influence pas la performance des vendeurs ($\alpha = 5\%$).

[width=clip]ex$_{v}$*endeurs*

6 Comparaisons non-paramétriques

6.1

Dans une institution, on a prélevé deux échantillons : l'un parmi l'ancien personnel et l'autre parmi le nouveau. On a demandé au hasard à 10 anciens employés et à 6 nouveaux d'évaluer sur une échelle allant de 0 à 20 l'efficacité de la nouvelle organisation interne.

On a obtenu les résultats suivants:

anciens employés	12; 4; 5; 13; 9; 14; 6; 12; 11; 16
nouveaux employés	5; 3; 4; 2; 12; 7

1. Au vu des données ci-dessus, peut-on admettre que les opinions des nouveaux et des anciens employés se distribuent de façon analogue?
2. Effectuer un test bilatéral de la somme des rangs de Wilcoxon avec un risque α de 5%. Commenter.

6.2

On examine l'âge d'entrée dans l'enseignement primaire pour différents pays. On considère un échantillon de pays asiatiques et un échantillon de pays européens et on arrive aux âges suivants:

Asie	4	6	5			
Europe	4	5	6	6	7	7

Tester, en utilisant la statistique de la somme des rangs de Wilcoxon, si les âges d'entrée diffèrent significativement entre les deux continents (on prendra $\alpha = 5\%$).

6.3

Voici des échantillons des salaires horaires de base pratiqués par des entreprises de deux branches industrielles:

Bâtiment	Automobile
15	12
12	16
16	16
	16

Effectuer le test de Wilcoxon. Peut-on admettre une différence significative à 5% entre les salaires horaires des deux branches?

6.4

On a observé le déficit budgétaire de 10 villes européennes en 2004, 2005, et 2006 (les villes observées restant les mêmes). En faisant la différence "déficit de 2005" moins "déficit de 2004" on a constaté que 9 villes avaient une différence positive et 1 une différence négative. En faisant la différence "déficit de 2006" moins "déficit de 2005" on a constaté que 7 villes avaient une différence positive et 3 une différence négative.

Effectuer deux tests du signe bilatéraux avec un risque $\alpha = 5\%$, le premier entre 2004 et 2005 et le second entre 2005 et 2006. Commenter.