Statistique inférentielle
TD 1 : Estimation

Exercice 1 : Maîtrise Statistique des Procédés

Une entreprise de construction mécanique fabrique de pièces de moteur de voiture pour un grand constructeur automobile. Les exigences du client sont les suivantes : les pièces doivent faire 20cm de diamètre.
Une fois le procédé bien calibré, la fabrication démarre. Le processus est alors supposé, lorsqu'il est en fonctionnement *normal*, fabriquer des pièces dont le diamètre X suit une loi normale de moyenne 20cm, et d'écart-type 0.1mm. Nous dirons que le processus est alors en fonctionnement normal.
Afin de suivre le bon déroulement de la fabrication, le contrôleur qualité prélève 5 pièces toutes les heures, en mesure le diamètre et calcule la moyenne des 5 diamètres. Voici les résultats trouvés sur une journée complète de 24h : 19.993, 19.993, 19.994, 19.995, 20.004, 19.985, 19.990, 19.990, 19.996, 19.993, 20.000, 20.006, 19.991, 19.992, 19.995, 19.992, 20.002, 20.002.

1. Peut-on utiliser ces résultats pour estimer l'espérance et la variance du diamètre des pièces produites par l'entreprise ? Si oui, faites-le.
2. Soit $\bar{X}_i$ la moyenne des 5 mesures au temps i. Quelle devrait être la loi de $\bar{X}_i$ si le processus était en fonctionnement normal ?
3. Donner un intervalle $[b_{inf}, b_{sup}]$, symétrique autour de la valeur cible de 20cm, auquel doit appartenir la variable $\bar{X}_i$ avec une probabilité de 99.7%.
4. Construire une carte de contrôle (cf. votre cours) sur la moyenne de la production, en utilisant les deux bornes b_{inf} et b_{sup} comme limite. Le procédé est-il resté sous contrôle toute la journée ?

Exercice 2 : comparaison des statistiques S^2 et V^2 pour estimer la variance

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires de loi parente d'espérance μ inconnue et de variance σ^2 également inconnue. Considérons les statistiques

$$V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{et} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1. Calculer leur espérances et en déduire le meilleur estimateur pour la variance σ^2.
2. Nous supposons maintenant que l'espérance μ est connue. Soit la statistique

$$V_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

   Calculer son espérance. Comparer alors les deux estimateurs V_μ^2 et S^2.
3. Sachant que la variance de V^2 est

$$V(V^2) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4],$$

   où $\mu_4 = E[(X_i - \mu)^4]$ est le moment centré d'ordre 4, calculer la variance de V_μ^2 et S^2. Conclure quant au choix d'un estimateur pour σ^2 lorsque l'espérance μ est connue.

Exercice 3 : estimation du paramètre d'une loi de Poisson

Une entreprise de vente à distance s'intéresse au nombre de commandes hebdomadaires d'un nouveau modèle de pantalon. On suppose que ce nombre de commandes suit une loi de Poisson de paramètre λ. Un relevé effectué sur 5 semaines choisies au hasard donne les nombres de commandes suivantes : $2, 4, 1, 0, 3$.
On veut estimer le paramètre λ en construisant l'estimateur du maximum de vraisemblance.

1. Que représente λ ?
2. Soit l'échantillon $X_1, \ldots, X_5$ représentant le nombre de commandes pour les 5 semaines d'observation. Écrire la fonction de vraisemblance $L(X_1, \ldots, X_5, \lambda)$.
3. Que valent la vraisemblance et la log-vraisemblance pour les 5 valeurs observées de l'échantillon : $x_1 = 2$, $x_2 = 4, x_3 = 1, x_4 = 0, x_5 = 3$.
4. En déduire une estimation $\hat{\lambda}$ de λ qui maximise la vraisemblance.
5. Généraliser la démarche pour n semaines choisies au hasard et donner l'estimateur T du maximum de vraisemblance.
6. Quels sont l'espérance et la variance de T ? *(Indication : on rappelle que la somme de deux variables aléatoires de loi de Poisson de paramètres λ_1 et λ_2 suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.)*
7. Calculer l'information de Fisher apporté sur le paramètre λ par un $n-$échantillon $X_1, \ldots, X_n$.
8. En déduire que l'estimateur T est un estimateur efficace de λ.

Exercice 4 : Détection de valeurs aberrantes

Soit $X_1, \ldots, X_n$ un échantillon de fonction de répartition $F(x)$ et de densité $f(x)$. Soit $(Y_1, \ldots, Y_n)$ la version ordonnée croissante de l'échantillon $X_1, \ldots, X_n$. Soient $H_k(x)$ et $h_k(x)$ les fonctions de répartitions et de densité de Y_k.
Soit les deux extrêmes $Y_1 = \inf X_i$ et $Y_n = \sup X_i$.

1. Quelle sont leur lois (donner leur fonction de répartition et de densité) ?
2. Quelle est la probabilité qu'une observation d'une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$ dépasse $\mu + 3\sigma$?
3. Et parmi un échantillon de taille 100, quelle est la probabilité d'avoir une telle observation ?
4. Parmi un échantillon de taille 100 de loi $\mathcal{N}(0, 1)$, quelle valeur ne doit pas être dépassée avec une probabilité de 99% ?
5. Une société d'analyse de la qualité de l'environnement effectue un sondage auprès ses différentes laboratoires (50, répartis dans toute la France) afin d'évaluer s'ils effectuent des mesures correctes. Pour cela le service qualité envoie à chaque laboratoire un échantillon d'eau contenant un certaine teneur en chrome, et leur demande de mesurer cette la teneur en chrome. En prenant en compte les fluctuations dans la préparation de la solution, ainsi que les imprécisions des appareils de mesure, la société suppose que la teneur en chrome (mg/l) suit une loi $\mathcal{N}(10, 1)$.
   Parmi les résultats, deux laboratoires ont retourné des mesures plus éloignées que les autres : le laboratoire L_1 a mesuré une teneur de 6 mg/l (plus petite de toutes les mesures), et le laboratoire L_2 a mesuré une teneur de 13 mg/l (plus grande de toutes les mesures).
   Pouvez-vous dire, avec une probabilité de 99%, que ces mesures sont cohérentes où alors s'agit-il de valeurs aberrantes (erreur de saisie, dérèglement de l'appareil de mesure...) ?

Exercice 5 : détermination d'une statistique exhaustive

Soit X une variable aléatoire de loi γ de paramètre θ. La fonction de densité de X est :

$$f_X(x) = \frac{1}{\Gamma(\theta)} e^{-x} x^{\theta - 1}.$$

1. Montrer que la densité de X peut s'écrire sous la forme

   $$f_X(x) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)]$$

   Une telle densité est dite de la *famille exponentielle*.
2. En déduire une statistique exhaustive pour le paramètre θ fonction d'un échantillon $X_1, \ldots, X_n$.

Statistique inférentielle
TD 2 : Estimation par intervalle de confiance

Exercice 1

On a pesé 10 palettes de briques de la même fabrication ; et on a obtenu les résultats suivants (kilogrammes)

$$759, 750, 755, 756, 761, 765, 770, 752, 760, 767$$

On admet que ces résultats sont issus d'une population distribuée selon une loi normale d'espérance μ et de variance σ^2.

1. Donner une estimation ponctuelle de l'espérance et de la variance du poids d'une palette de brique.
2. Construire un intervalle de confiance pour μ avec les niveaux de confiance 0.90 et 0.99.
3. Quel niveau de confiance choisir pour avoir un intervalle de confiance deux fois plus étroit que celui obtenu avec une confiance de 0.9 ?
4. Supposons maintenant que l'on connaisse la variance, donnée par le constructeur : $\sigma^2 = 42$. Que cela change-t-il sur vos intervalles de confiances ? Recalculez-les si besoin.
5. Combien de palettes de briques aurait-on dû mesurer pour que la longueur de l'intervalle de confiance, de niveau de 95%, n'excède pas $0,5$kg (en supposant que les estimations des moyennes et variances ne changent pas).

Exercice 2

Le laboratoire SIMTECH, firme d'expertises en contrôle des matériaux, a été mandaté par une société de gérance de projets de construction pour évaluer la qualité d'un mélange bitumineux provenant de deux usines. Il a été convenu d'effectuer une vérification par 115 mètres cubes de béton et d'évaluer la résistance à la compression, à l'âge de 3 jours, sur des cylindres standards. Les résultats de la résistance à la compression en kg/cm^2 pour les deux usines se résument comme suit.

	Usine 1	Usine 2
Nombres de cylindres	$n_1 = 25$	$n_2 = 23$
Résistance moyenne de l'échantillon	$\bar{x}_1 = 90,6$	$\bar{x}_2 = 94,4$
Variance de l'échantillon	$v_1^2 = 65,42$	$v_2^2 = 58,24$

On suppose que la résistance à la compression est distribuée normalement quelque soit l'usine de fabrication.

1. Construire un intervalle de confiance pour la variabilité de la résistance à la compression du béton provenant de chaque usine, au niveau de confiance $0,95$.
2. Peut-on en déduire que la variabilité de la résistance à la compression du béton provenant de chaque usine est différente ?
3. Déterminer un intervalle de confiance pour le rapport σ_1^2/σ_2^2 des deux variances, avec un niveau de confiance de 95%.

Exercice 3

Lors d'un sondage précédant les élections présidentielles, 500 personnes ont été interrogées. Bien que ce ne soit pas le cas en pratique, on suppose pour simplifier les calculs que les 500 personnes représentent un échantillon indépendant et identiquement distribué de la population française.
Sur les 500 personnes, 150 ont répondu vouloir voter pour le candidat C_1, et 140 pour le candidat C_2.

1. Donner une estimation ponctuelle des intentions de votes, sous la forme d'un pourcentage.
2. Donner un intervalle de confiance à 95% pour chacun des deux intentions de votes.
3. Peut-on prédire l'élection d'un candidat ?

Exercice 4

Pour juger de la teneur en magnésium d'une eau minérale, on a effectué 10 mesures :

248 246 246 247 247 249 247 250 248 245 (mg pour 10 litres).

La teneur étudiée est supposée être une variable aléatoire normale d'espérance μ et de variance σ^2.

1. Déterminez un intervalle de confiance sur μ pour un niveau de confiance de 0.95.
2. Trouver la valeur σ_0 de σ qui n'aurait que 5 chances sur 100 d'être dépassée.

Exercice 5

Une firme nationale de sondages d'opinion a effectué pour le compte d'une compagnie d'assurance, une étude sur les besoins financiers et la satisfaction des clients. Dans la section du questionnaire concernant les fonds communs de placement, on demande aux clients de donner la valeur (en euros) de tous les fonds communs de placement qu'ils possèdent. Voici les résultats pour un échantillon aléatoire de 20 clients :

Fond commun de placement				
93850	121500	166675	173000	81580
172450	80515	191000	105630	192100
151975	148000	173400	138330	142500
149660	120225	149375	131170	85600

On suppose que la valeur actuelle des fonds communs de placement est distribuée normalement.

1. Donner une estimation ponctuelle de la valeur moyenne des fonds communs de placement des clients.
2. On appelle parfois l'*erreur-type* l'écart-type de l'estimateur utilisé. Quelle est-elle ici ?
3. Déterminez un intervalle de confiance ayant une probabilité de 95% de contenir la vraie valeur du montant moyen des fonds communs de placement.

POLYTECH'LILLE
DÉPARTEMENT G.I.S.

Statistiques inférentielles
TD 3 : Tests sur une population

Exercice 1

Une entreprise SupMetal fournit à un client de la région Nord Pas De Calais, l'entreprise LilTech, des supports métalliques. L'entreprise LilTech exige que les supports aient, en moyenne, une longueur de $70mm$. Ce support est fabriqué par une machine, mais il y a des petites variations de longueur dans les pièces qu'elle produit. On admet que la longueur des supports est distribuée normalement et que la dispersion de la fabrication est de $\sigma = 3mm$. Cette entreprise fournit également les mêmes pièces à l'entreprise PariTech, concurrent direct de LilTech, mais qui commande de beaucoup plus grandes quantités, et qui exige elle une longueur de $67mm$.
Les employés de l'entreprise LilTech ayant souvent des problèmes pour monter ces supports, soupçonnent SupMetal de fournir à LilTech les mêmes pièces qu'à PariTech, afin d'éviter d'avoir à régler la machine à chaque commande de PariTech ou de LilTech. Pour vérifier cela, LilTech prélève un échantillon aléatoire de 25 supports. Les mesures obtenues $(x_1, \ldots, x_{25})$ ont pour longueur moyenne de $\bar{x} = 68mm$.

1. Formuler les hypothèses d'un test statistique permettant de tester l'honnêteté de SupMetal.
2. Écrire la probabilité de l'échantillon $(x_1, \ldots, x_{25})$, autrement dit la vraisemblance, sous chaque hypothèse H_0 et H_1.
3. Former le test du rapport de vraisemblance pour un risque $\alpha = 5\%$ et $\alpha = 1\%$.
4. Conclure.
5. Calculer les risques de deuxième espèce correspondant aux deux risques α et en donner une interprétation.

Exercice 2

Un ingénieur risque crédit, employé dans une société spécialisée dans le crédit à la consommation, veut vérifier l'hypothèse selon laquelle la valeur moyenne des mensualités de ses clients est de 200 euros. Un échantillon aléatoire de 144 clients, prélevé aléatoirement dans la base de données, donne une valeur moyenne estimée à 193.74 euros et un écart-type estimé à 48.24 euros.

1. Quelles sont les hypothèses statistiques associées à la problématique du comptable et quel type de test faut-il mettre en oeuvre pour l'aider à prendre une décision statistiquement correcte ?
2. Peut-il conclure, au niveau de confiance 95% , que la valeur moyenne postulée des stocks est correcte ?
3. Faîtes le schéma des régions de rejet et de non rejet de l'hypothèse nulle H_0 en y notant les valeurs critiques calculées à la question précédente.
4. Représenter sur ce schéma la p_{value} associée à ce test. Que vaut-elle ?
5. En utilisant la p_{value}, quelle aurait été la réponse à la question 2 pour un risque de première espèce $\alpha = 10\%$.

Exercice 3

Pour comparer les proportions de personnes atteintes par la grippe en ville et à la campagne, deux échantillons ont été mesuré :
– sur 100 personnes habitant une grande agglomération, on a observé une proportion $f_0 = 0.24$ de sujets ayant eu la grippe,
– sur 80 personnes habitant à la campagne, on a observé une proportion $f_1 = 0.20$ de sujets ayant eu la grippe.
Les citadins sont-ils plus atteints par la maladie que les ruraux ? ($\alpha = 0.05$)

Exercice 4 :

Une machine est réglée pour fabriquer des plaques de chocolats d'un poids 'moyen' de 250g. Soucieux de ce problème, le service de contrôle de qualité demande une vérification de la machine. Le poids de 10 plaques de chocolats est observé. On obtient les mesures suivantes qui vous sont immédiatement transmises :

poids observés	256	245	253	250	295	251	248	247	252	249

Quelle est votre conclusion ?

Exercice 5 :

Une société de vente à distance demande à l'un de ses ingénieurs marketing de modéliser le nombre d'appels téléphoniques par heure reçus sur le standard dédié aux commandes, dans le but d'optimiser la taille de celui-ci. Les nombres d'appels, relevés sur une période de 53 heures, ont été les suivants :

Nb d'appels x_i	0	1	2	3	4	5	6	7	8	9 et plus
Occurence N_i	1	4	7	11	10	9	5	3	2	1

1. Estimer la moyenne et la variance du nombre d'appels. Quelle type de loi semble le mieux décrire ce nombre d'appel ?
2. Tester l'ajustement à cette loi au risque 5%.
3. Sachant qu'une hôtesse d'accueil téléphonique peut traiter jusqu'à 7 appels par heure, combien d'hôtesses doit-on employer pour pouvoir répondre à 95% des appels téléphoniques ?

Exercice 6 :

Sur 2000 personnes interrogées dans le Nord, 1040 disent acheter régulièrement des vêtements sur le site internet de VetiLille. Sur 1500 interrogées dans le reste de la France, 615 disent acheter sur ce site. Est-ce que ces résultats permettent de soutenir que ce site séduit autant les habitants du Nord que du reste de la France (risque de 5%) ?

Exercice 7 :

Un ingénieur statisticien d'une société d'assurance est chargé d'étudier l'impact d'une campagne de publicité réalisée dans 7 régions dans lesquelles la société est déjà implantée. Pour ceci, il a extrait de la base de donnée, pour un certain nombre d'agents généraux de chaque région, le nombre de nouveaux clients récoltés :

Région	1	2	3	4	5	6	7
Nb d'agents généraux	9	7	7	6	7	6	6
Nb moyen de nouveaux clients	26.88	22.34	19.54	18.95	27.17	25.87	25.72
Variance du nb de nouveaux clients	13.54	12.59	12.87	13.42	13.17	12.56	12.64

L'ingénieur statisticien décide alors de réaliser une analyse de variance afin de tester si le facteur région a une influence sur le nombre de nouveaux clients récoltés.
On appelle X_k^i le nombre de nouveaux clients du i-ème agent général de la région k. Soit n_k le nombre d'agents généraux de la région k, et K le nombre de régions ($K = 7$). Nous supposons que les variables aléatoires X_k^i sont normales, de moyenne μ_k et de variance σ.
Le problème consiste donc à tester

$$H_0 : \mu_1 = \ldots = \mu_K = \mu \qquad \text{contre } H_1 : \exists 1 \leq i, j \leq K \text{ t.q. } \mu_i \neq \mu_j.$$

Soient :

$$\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad \bar{X} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} X_k^i \quad \text{où} \quad n = \sum_{k=1}^{K} n_k.$$

1. Interpréter $\bar{X}_k$ et $\bar{X}$.
2. En remarquant que $X_k^i - \bar{X} = X_k^i - \bar{X}_k + \bar{X}_k - \bar{X}$, démontrer la formule d'analyse de variance :

$$\underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X})^2}_{V_T^2} = \underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \bar{X}_k)^2}_{V_R^2} + \underbrace{\frac{1}{n}\sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})^2}_{V_A^2}$$

   qui représente la décomposition de la variance totale V_T^2 en la variance V_A^2 due au facteur A (variance inter-groupe) plus la variance résiduelle V_R^2 (ou variance intra-groupe).
3. Calculer V_T^2, V_A^2 et V_R^2.
4. Finaliser l'analyse de variance pour juger si la campagne de publicité a eu le même impact dans toutes les régions.

Statistiques inférentielles
TD-TP 4 : Tests sur plusieurs populations

Exercice 1

En prélevant un échantillon (supposé représentatif) de 41 étudiants de Polytech-Lille, on constate que la taille moyenne de cet échantillon est de $\bar{x}_1 = 1.7$m avec un ecart-type de $v_1 = 8$cm. En faisant de même pour un échantillon de 61 étudiants d'une école voisine on trouve une taille moyenne de $\bar{x}_2 = 1.68$m avec un ecart-type de $v_1 = 9$cm. En supposant que ces deux échantillons sont distribués normalement, peut-on affirmer que les étudiants de ces deux écoles sont semblables ($\alpha = 5\%$) ?

Exercice 2 (R)

On souhaite mesurer l'influence de l'alcool sur le temps de réaction au volant. Sur un échantillon aléatoire de 30 chauffeurs, le temps de réaction a été observé en laboratoire avec et sans consommation d'alcool (les 30 chauffeurs ont été réparti aléatoirement). Les temps de réactions en secondes ont été rapportés dans le tableau suivant :

Sans	0.68	0.64	0.68	0.82	0.58	0.80	0.72	0.65	0.84	0.73	0.65	0.59	0.78	0.67	0.65
Avec	0.73	0.62	0.66	0.92	0.68	0.87	0.77	0.70	0.88	0.79	0.72	0.60	0.78	0.66	0.68

1. Tracer sur un même graphique les fonctions de répartition emprique correspondant aux deux situations.
2. Peut on affirmer qu'il y a une influence de l'alcool sur le temps de réaction ($\alpha = 5\%$) ? On utilisera trois tests différents.

Exercice 3 (R)

On désire tester l'effet d'un médicament censé réduire le taux de le cholesterol. On a mesuré le taux de cholesterol (g/l) chez 10 patients, avant la prise de ce médicament, et une semaine après l'avoir pris. Voici les taux obtenus :

Avant	0.1	0.2	0.15	0.3	0.34	0.16	0.09	0.24	0.17	0.29
Après	0.8	0.18	0.12	0.2	0.3	0.21	0.12	0.16	0.17	0.22

Le médicament a-t-il un effet ($\alpha = 5\%$) ?

Exercice 4 (R)

Deux populations de 42 et 50 individus sont utilisées pour étudier un traitement dont on ignore a priori l'effet possible (augmentation ou diminution de performances). Les mesures sont faites indépendamment les unes des autres, et sont réparties en quatre classes :

Classement	mauvais	moyen	bon	excellent
Groupe traité	4	6	17	15
Groupe contrôle	10	13	16	11

1. Tracer sur le même graphique les fonctions de répartitions empiriques associées aux deux groupes
2. Peut-on rejeter l'hypothèse que le traitement est sans effet ? Avec quel risque ?

POLYTECH'LILLE
DÉPARTEMENT G.I.S.

Travaux pratiques de Statistiques Inférentielles sous SAS et R - GIS 3
TP 1 : Statistique Exploratoire

1 Préliminaires et indications

Avant tout, **veuillez lire attentivement l'introduction au logiciel SAS** qui vous a été distribuée.

Connexion Connectez-vous sur vos comptes sous environnement LINUX.
Loguez-vous sur weppes par l'instruction : `ssh -X weppes.studserv.deule.net`
Une fois connecté, lancez SAS par l'instruction : `/usr/local/SAS/SASFoundation/9.2/sas`

Répertoires Créer sur votre compte un répertoire `TP_Stat_SAS`.
Dans ce répertoire, créer 3 sous-répertoires : `librairies`, `donnees`, `programmes`. Vous enregistrerez vos programmes SAS en `.sas` dans le dossier `programmes`, vos fichiers de données (`.dat`) dans `donnees`.
Suivez la note d'introduction à SAS pour créer une librairie dans laquelle vous enregistrerez les tables que nous utiliserons dans ce TP.

Exécution différé Il est possible d'exécuter des programmes SAS sans ouvrir le logiciel SAS. Cela peut être utile notamment lorsque les programmes nécessitent un temps d'exécution long. En pratique, cela diminue aussi les ressources demandées à l'ordinateur pour gérer l'affichage graphique des différentes fenêtres SAS.
Pour cela, il suffit d'enregistrer votre programme sous le nom `mon_prog.sas`, et de lancer son exécution à l'aide de la commande suivante dans un terminal :
`/usr/local/SAS/SASFoundation/9.2/sas mon_prog.sas -fsdevice x11.motif`
A noter qu'il est nécessaire de s'être au préalable loguer sur le serveur `weppes`.
Les résultats sont alors regroupés dans un fichier `mon_prog.lst` tandis que le compte-rendu de l'exécution ainsi que les messages d'erreurs se trouvent dans le fichier `mon_prog.log`.

Consignes

- Chaque exercice devra faire l'objet de l'écriture d'un programme SAS. Pensez à toujours avoir un éditeur de texte dans lequel vous écrivez et sauvez votre code, que vous transférez ensuite à l'éditeur SAS par copier/coller.
- Vous rédigerez un compte rendu détaillé de votre TP, sous Open Office, en incluant vos programmes SAS commenté, les résultats, vos interprétations et commentaires.

2 Exercices de statistique exploratoire

Les jeux de données étudiés sont disponibles sur `http ://math.univ-lille1.fr/~jacques/`

Exercice 1 : Manipulation de données

La procédure `sql` en SAS permet de gérer les bases de données à l'aide du langage SQL. Même si ce n'est pas la seule possibilité pour faire cet exercice, son utilisation est conseillée.

1. Créer une table SAS contenant les données suivantes (nom, sexe, taille et date de naissance), et afficher son contenu :
   tutu M 1,70 11/12/82
   toto M 1,82 21/12/82
   titi F 1,57 25/12/83
   Rencontrez-vous des problèmes dans cette étape ? Pourquoi ?
2. Trier cette table suivant la taille décroissante des individus (`proc sort`).

3. En supposant que le poids en kg des hommes est : $poids = (tailleencm)/2 - 10$ et que celui des femmes est $poids = (tailleencm)/2 - 20$, créer une nouvelle table en ajoutant la variable poids.
4. Quel est le poids moyen des hommes ?
5. Afficher uniquement la personne la plus légère.

Exercice 2 : Statistiques descriptives, premiers graphiques

Récupérez le fichier de données `Employes.dat`. Ce fichier contient pour 12 employés d'une entreprise, le numéro d'identification, l'âge, le sexe, le salaire annuel en euro, l'ancienneté et la situation familiale.

1. Après avoir constaté dans ce fichier la nature des variables, chargez le dans une table SAS.
2. Faites une analyse descriptive des variables numériques par rapport aux 12 employés, puis par rapport aux modalités de la variables sexe et enfin par rapport à la variables situation familiale. Interpréter ces résultats.
3. Y-a-t'il une corrélation entre l'âge et le salaire, entre l'ancienneté et le salaire ?
4. Représenter le salaire en fonction de l'ancienneté (`proc plot`) en différenciant les hommes et les femmes, puis les célibataires des mariés. Ce graphique vous suggère-t-il une constatation ?
5. Sur un histogramme (`proc chart`), représenter les fréquences de salaire en 5 classes, en différenciant hommes et femmes.

Exercice 3 : Analyse d'un jeu de données bancaires

Le jeu de données `GermanCredit.data` comporte des renseignements sur 1000 clients d'une banque allemande, chaque client étant décrit par 20 variables.

1. Calculer les indicateurs de tendance centrale, de dispersion et de forme vu en cours pour les variables « durée du crédit »et « montant du crédit ». Interpréter ces valeurs.
2. Représenter graphiquement les distributions de ces deux variables à l'aide de box-plot et d'histogramme. Représenter également les deux variables sous la forme d'un nuage de point.
3. Pouvez-vous mettre en évidence une corrélation entre ces deux variables ?
4. Nous nous intéressons maintenant aux variables « état marital »et « but du crédit ». Représenter graphiquement et interpréter la distribution de ces variables.

Exercice 4 : Simulation de Monte-Carlo (logiciel R)

On cherche dans cet exercice à approcher l'intégrale $I = \int_0^2 e^{-\frac{x^2}{2}} dx$. Pour cela nous utilisons une méthode de Monte-Carlo (vue en TD de probabilité). Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires uniformes sur $[0, 2]$, et soit $Y_i = e^{-\frac{X_i^2}{2}}$ pour tout $i = 1, n$.

1. Quelle est la limite, au sens de la convergence en probabilité, de $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ lorsque $n \to \infty$?
2. Utiliser ce résultat pour approcher l'intégrale I, en simulant n variables aléatoires Y_i ($n = 100, 10^4, 10^6$).
3. Répéter 100 fois ces approximations, et représenter les résultats sous la forme d'une boîte à moustache pour chacune des 3 valeurs de n utilisées. Que constatez-vous ?
4. Représenter cette fois ces résultats sous la forme d'un histogramme (pour chaque valeur de n). Avez-vous une idée de la distribution de ces résultats d'approximation ? Que vous dit le théorème centrale limite ?

Exercice 5 : Calcul de vraisemblance (logiciel R)

1. Simuler 3 échantillons $X_1, \ldots, X_n$ gaussiens centrés réduits (fonction `rnorm`)) de taille 10, 100 et 1000. On oublie désormais que nous avons simulé ces échantillons à partir une loi normale, et nous allons essayer plusieurs modélisation pour cette échantillon.
2. Première hypothèse : nous supposons que l'échantillon suit une loi exponentielle. Estimer le paramètre de cette loi, et calculer la vraisemblance de l'échantillon sous cette hypothèse.
3. Faites de même pour la loi normale. Que concluez-vous ?

POLYTECH'LILLE
DÉPARTEMENT G.I.S.

Travaux pratiques de Statistiques Inférentielles sous SAS et R - GIS 3
TP 2 : Estimation et tests

Les jeux de données étudiés sont disponibles sur `http://math.univ-lille1.fr/~jacques/`

Exercice 1 (SAS): Test sur l'espérance

On cherche à estimer le temps d'attente moyen au guichet d'une grande banque aux heures de forte affluence. On a observé 26 clients choisis au hasard et on a obtenu les temps d'attente suivants: 6,1; 4,7; 5,6; 4,5; 5,5; 6,8; 2,1; 2,1; 3,5; 2,5; 6,7; 4,4; 4,5; 6,5; 4,9; 3,8; 2,5; 4,0; 6,5; 5,6; 2,7; 3,4; 5,6; 3,5; 4,8; 4,4
On suppose que ces temps d'attente sont distribués normalement.
Peut-on affirmer au risque $\alpha = 5\%$ que le temps moyen d'attente au guichet est égal à 4 minutes ?

1. Créer une table SAS qui contient les temps d'attente des 26 clients ?
2. Donner un intervalle de confiance sur la moyenne et la variance de ces temps d'attentes, à l'aide de la procédure `ttest`.
3. Les procédures `means` et `univariate` permettent d'effectuer un test sur la moyenne basé sur la loi de Student. Pourquoi est-il approprié ici ? Dans quel cas ne le serait-il pas ?
4. Effectuez ce test à l'aide de ces deux procédures.
5. Toujours avec le même risque, peut-on affirmer que le temps moyen d'attente au guichet est supérieur à 4 minutes ?

Indication : dans les options de la procédure `means`*, il faut indiquer* `t` *pour indiquer que l'on veut calculer la statistique du test de Student de nullité de la moyenne, et* `prt` *pour calculer la p-value relative à ce test.*

Exercice 2 (R): Estimation de densité

1. Simuler trois séries de données de tailles $n = 10, n = 100$ et $n = 1000$ représentant des observations i.i.d. issues d'une distribution exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda = 0.5$ et $\lambda = 1$
2. Pour chaque valeur de λ et de n, représenter graphiquement la fonction de répartition empirique et théorique (sur le même graphique)
3. Pour chaque valeur de λ et de n, représenter graphiquement sur le même graphique l'estimation non-paramétrique de la fonction de densité de la loi $\mathcal{E}(\lambda)$ en utilisant les observations simulées et la densité théorique. On utilisera un noyau Gaussien et la taille de fenêtre optimale vue en cours.

Exercice 3 (R): Puissance de test

1. Créer une matrice à $N = 100$ lignes et $n = 100$ colonnes, à l'aide de la commande `matrix`.
2. Remplir chaque ligne de la matrice par un échantillon de 100 simulations de loi normale centrée réduite.
3. Créer une fonction permettant d'effectuer le test de nullité de la moyenne. Cette fonction aura en paramètre le risque de première espèce, et retournera 0 si H_0 est rejeté, 1 sinon.
4. Au risque $\alpha = 5\%$, combien de fois parmi les 100 simulations le test a-t-il accepté H_0, rejeté ?
5. Faire de même en simulant cette fois des gaussiennes centrées en 1. En déduire une valeur expérimentale de la puissance de ce test. Tester plusieurs valeurs de n (10, 50 et 100).

6. La puissance du test de nullité de la moyenne, dans les conditions de cet exercice (distribution gaussienne et variance connue égale à 1), définie par $1 - p(\text{accepter}H_0|H_1)$, est donnée par :

$$\begin{aligned} P(\mu_1) &= 1 - P(|\bar{X}| < \frac{u_{1-\alpha/2}}{\sqrt{n}}|H_1 : \bar{X} \sim \mathcal{N}(\mu_1, \frac{1}{\sqrt{n}})) \\ &= 1 - \Phi(u_{1-\alpha/2} - \sqrt{n}\mu_1) + \Phi(-u_{1-\alpha/2} - \sqrt{n}\mu_1) \end{aligned}$$

   Programmer cette fonction puissance.

7. Représenter $P(\mu_1)$ pour $\mu_1 \in [-2, 2]$, en superposant sur un même graphique les courbes de puissance du test pour $n = 10, 50, 100$. Quel test est le plus puissant ?

8. Dans le cas où $H_1 : \mu_1 = 1$, quelle est la puissance de chaque test. Comparer avec les valeurs expérimentales obtenues en 5.

Exercice 4 (R): Calcul du nombre de sujets pour atteindre une puissance de test

On considère le test $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_0 + \delta$. On suppose $\sigma = 0.5$.

1. Pour $n = 100$ et $\alpha = 5\%$, tracer le graphique de la puissance du test $1-\beta$ en fonction de $\delta \in \{0.1, 0.2, \ldots, 1\}$.

2. Si $\alpha = 5\%, \delta = 0.2$, calculer le nombre d'observations nécessaire pour que le risque de seconde espèce ne dépasse pas $\beta = 5\%$.

Exercice 5 (SAS et R): Test de l'aléatoire d'un échantillon et d'adéquation à une loi donnée

Dans l'exercice 1, nous avons supposé que les temps d'attente au guichet de la banque sont distribués normalement.

1. Vérifier sous **R** que l'échantillon est bien aléatoire.

2. Vérifier à la fois avec **R** et **SAS** que l'échantillon est bien distribué suivant une loi normale.

Travaux pratiques de Statistiques Inférentielles sous SAS et R - GIS 3
TP 3 : Tests

Les jeux de données étudiés sont disponibles sur `http://math.univ-lille1.fr/~jacques/`

Exercice 1 (SAS): Tests de comparaisons moyenne et variance

Afin de sélectionner des candidats qui ont postulé à un emploi, le directeur d'une entreprise a fait passer un test d'aptitude aux candidats, et il a noté le temps (en minutes) nécessaire à chacun des candidats pour répondre au test. Parmi les 27 candidats, 15 étaient des hommes et 12 des femmes. Les résultats obtenus sont les suivants:

Hommes	8,6	10,9	7,3	9,2	8,5	9,2	9,1	8,9	10,7	8,2	7,1	9,4	8,3	9,7	9,2
Femmes	8,3	7,2	8,7	6,7	10,3	6,8	9,8	8,9	9,6	8,6	6,7	7,5			

Nous supposons que les temps de réponse sont distribués normalement.

1. Peut-on dire que les variances des temps de réponse des hommes et des femmes sont identiques ?
2. Si la performance des candidats des deux sexes lors du test n'est évaluée que par le temps nécessaire pour y répondre, peut-on affirmer qu'il y a une différence réelle entre la performance moyenne des candidats et celle des candidates ?
   Indication : utiliser la procédure `ttest` *($\alpha = 5\%$).*

Exercice 2 (SAS): Test d'indépendance de variables qualitatives

Sur 2000 personnes interrogées dans le Nord, 1040 disent acheter la marque de dentifrice X. Sur 1500 interrogées dans le reste de la France, 615 disent acheter la marque X.
Est-ce que ces résultats permettent de soutenir que les parts de marché de la marque X sont les mêmes dans le Nord que dans le reste de la France, au seuil de risque de 5%?
Indication : une solution peut être d'utiliser un test d'indépendance du χ^2 entre les deux variables région et achat. Ceci peut être réalisé à l'aide de la procédure `freq`.

Exercice 3 (SAS ou R): ANOVA

Le fichier `orge.dat` contient les valeurs de rendements de six engrais azotés pour 4 types de sols (dans l'ordre traitement, bloc, rendement). Les engrais sont les suivants :
1 : (NH4)2 SO2, 2 : NH4NO3, 3 : CO(NH2)2, 4 : CA(NO3)2, 5 : NaNO3, 6 : Rien.

1. L'engrais a-t-il une influence sur le rendement ?
2. Analyser ensuite les deux facteurs engrais et type de sols à l'aide d'une ANOVA à 2 facteurs.

Exercice 4 (SAS)

Récupérer le fichier `GermanCredit.data`.
En s'inspirant des méthodes statistiques vues en cours, répondre aux questions suivantes en justifiant et illustrant vos réponses :

1. Les clients de cette banque sont-ils jeunes (moins de 30 ans) ?
2. Le sexe a-t-il une influence sur le montant emprunté ? Si oui, les femmes empruntent-elles un montant plus important que les hommes ?

3. L'emploi et le sexe influent-ils sur la durée de l'emprunt ?
4. Le montant du crédit ainsi que la durée sont-elles des variables gaussiennes ?
5. Le montant du crédit est-il lié à la durée ?

Exercice 5 (R)

On s'intéresse au taux de fer présent dans le foie et le régime à suivre pour mieux contrôler ce taux. On souhaite comparer l'effet des 5 régimes. Il s'agit d'une étude sur des souris. Le plan d'expérience consiste à assigner de manière aléatoire 9 souris pour chaque régime (on considère que la durée du régime est suffisamment grande pour qu'elle *efface* les éventuelles différences entre les souris avant le régime). Les résultats obtenus sont :

A	B	C	D	E
2.23	5.59	4.50	1.35	1.40
1.14	0.96	3.92	1.06	1.51
2.63	6.96	10.33	0.74	2.49
1.00	1.23	8.23	0.96	1.74
1.35	1.61	2.07	1.16	1.59
2.01	2.94	4.90	2.08	1.36
1.64	1.96	6.84	0.69	3.00
1.13	3.68	6.42	0.68	4.81
1.01	1.54	3.72	0.84	5.21

Remarque : On organisera les données sous la forme d'un tableau à deux colonnes : X = tau de fer, Y = type de régime (variable qualitative = fonction R as. factor). Chaque ligne correspond donc à un individu.

1. Tracer sur un même graphique :
   - les 5 boîtes à moustaches correspondant aux 5 échantillons,
   - les 5 fonction de répartition empiriques correspondant aux 5 échantillons.
2. Est-ce qu'il y a une différence entre les régimes. On utilisera à la fois un test paramétrique (après avoir rappelé les hypothèses faites) et un test non paramétrique.

Exercice 6 (R)

Sur 10 patients choisis au hasard on observe l'évolution durant 5 jours du taux (en mg/litre sang) d'une certaine substance.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Jour 1	124	88	130	115	92	80	101	98	132	85
Jour 2	125	75	138	108	92	78	105	97	125	86
Jour 3	117	73	133	108	92	74	101	92	124	83
Jour 4	123	69	130	102	88	70	95	93	128	84
Jour 5	119	70	127	98	88	70	95	93	125	85

1. Tracer sur un même graphique les 5 fonctions de répartition empiriques ainsi que les 5 boîtes à moustaches correspondant aux 5 jours.
2. Les données observées permettent-elles de conclure à une variation significative dans le temps du taux mesuré.
3. Les données observées permettent-elles de conclure à une décroissance significative dans le temps du taux mesuré.

Exercice 7 (R)

(Re)faire les exercices 2 à 4 du TD 4.