1 Rappels de probabilité et de statistique

Exercice 1
Un examen est organisé conjointement pour 80 étudiants du collège A, 70 étudiants du collège B et 50 étudiants du collège C. Les taux de succès sont respectivement de 80%, 90% et 82% pour les collèges A, B et C.

(a) Quelle est la probabilité de succès pour un étudiant pris au hasard?

(b) Un étudiant pris au hasard a échoué. Quelle est la probabilité qu'il soit du collège A?

Exercice 2
Le poids des élèves d'un collège le premier jour de cours peut être approximé par une variable aléatoire normale avec écart-type de 2,8 kg. Un échantillon aléatoire simple de 8 élèves de ce collège donne les résultats suivants (en kg):

$$26, 27, 5, 31, 28, 25, 5, 30, 5, 32, 31, 5.$$

(a) Calculez un intervalle de confiance à 90% pour le poids moyen des élèves de ce collège.

(b) Calculez la taille minimale nécessaire pour qu'en valeur absolue, la différence entre la moyenne de l'échantillon et la moyenne de la population soit inférieure ou égale à 0,9 kg avec un niveau de signification de 0.97%

Exercice 3
Soient A et B deux événements aléatoires tels que $P(A \bigcap B) = 0,1$, $P(\overline{A} \bigcap \overline{B}) = 0,6$ et $P(A|B) = 0.5$. Calculez

(a) $P(B)$

(b) $P(A \bigcup B)$

(c) $P(A)$

(d) $P(\overline{B}|\overline{A})$

Indication: $\overline{S}$ est le complément de S. $P(S|T)$ est la probabilité de l'événement S, conditionnelle à l'événement T.

Exercice 4
On suppose que les dépenses de Noel d'une certaine population peuvent être approximées par une variable aléatoire normale avec moyenne et écart-type 45 EUR.

(a) On prend un échantillon simple et on obtient l'intervalle de confiance $[251, 6; 271, 2]$ pour un niveau de signification de 95%. Calculez la moyenne et la taille de l'échantillon que l'on a utilisé.

(b) On prend un échantillon aléatoire de taille 64 pour estimer la moyenne de la population, μ. Calculez l'erreur maximale que l'on commet ainsi pour un seuil de signification de $0, 90$.

Table 1: Distribution jointe de l'emploi et du diplôme universitaire de la population active des USA, en 2008

	Pas de diplôme universitaire (X=0)	Diplôme universitaire (X=1)	Total
Chômeur (Y=0)	0,037	0,622	0,659
Pas chômeur (Y=1)	0,009	0,332	0,341
Total	0,046	0,954	1,000

Exercice 5

La Table 1 donne la probabilité jointe de l'emploi et du diplôme universitaire dans la population active des USA en 2008, qui inclut donc à la fois ceux qui ont un emploi et ceux qui sont à la recherche d'un emploi.

(a) Calculez $E(Y)$

(b) Le taux de chômage est le pourcentage de chômeur dans la population active. Calculez ce taux.

(c) Calculez $E(Y|X=1)$ et $E(Y|X=0)$

(d) Calculez le taux de chômage des diplômés universitaires et des celle des non-diplômés.

(e) Un membre de cette population, pris au hasard, déclare être au chômage. Quelle est la probabilité qu'il soit diplômé? Et qu'il ne soit pas diplômé?

(f) Les variables X et Y sont-elles indépendantes?

Exercice 6

Calculez les probabilités suivantes:

(a) Si Y est distribué $N(50, 25)$, calculez $P(40 \leq Y \leq 52)$.

(b) Si Y est distribué $t_{[15]}$, calculez $P(Y \geq 1,75)$

(c) Si Y est distribué $t_{[90]}$, calculez $(1,99 \leq Y \leq 1,99)$

(d) Si Y est distribué $N(0, 1)$, calculez $P(1,99 \leq Y \leq 1,99)$

(e) Pourquoi les réponses de c) et d) sont-elles identiques?

(f) Si Y est distribué $\chi^2_{[4]}$, calculez $P(Y \leq 7,78)$

(g) Si Y est distribué $F_{[7;4]}$, calculez $P(Y > 4,12)$

(h) Si Y est distribué $\chi^2_{[10]}$, calculez $P(Y > 18,31)$

(i) Si Y est distribué $F_{[10;\infty]}$, calculez $P(Y > 1,83)$

(j) Pourquoi les réponses de h) et i) sont-elles identiques?

Exercice 7
Soient $\{Y_i\}_{i=1}^n$ des variables aléatoires Bernoulli, indépendantes et identiquement distribuées de moyenne $p = 0,4$.

(a) Utilisez le théorème central limite (CLT) pour calculer les approximations suivantes:

   (i) $P(\bar{Y} > 0,43)$ avec $n = 100$.

   (ii) $P(\bar{Y} \leq 0,37)$ avec $n = 400$.

(b) A combien doit-on augmenter la taille de l'échantillon pour s'assurer que $P(0,39 \leq \bar{Y} \leq 0,37) \geq 0,95$? (utilisez le théorème central limite (CLT) pour obtenir une réponse approximative à la question).

Exercice 8
Soient $\{Y_i\}_{i=1}^n$ des variables aléatoires avec moyenne identique μ_Y, variance commune σ_Y^2 et la même correlation ρ entre toutes les paires, c'est-à-dire, que la corrélation entre Y_i et Y_j est ρ pour tout $i \neq j$. Soit $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

(a) Démontrez que $Cov(Y_i, Y_j) = \rho\sigma_Y^2$ pour tout $i \neq j$.

(b) Supposez que $n = 2$, démontrez que $E[\bar{Y}] = \mu_Y$ et que $Var(\bar{Y}) = \sigma_y^2(1 + \rho)/2$.

(c) Pour n'importe quel $n \geq 2$, démontrez que $E[\bar{Y}] = \mu_Y$ et que $Var(\bar{Y}) = \sigma_y^2(1 + (n - 1)\rho)/n$.

(d) Pour n très grand, démontrez que $Var(\bar{Y}) = \rho\sigma_Y^2$.

Exercice 9
Une enquête est faite auprès de 1055 électeurs inscrits, et on leur demande s'ils voteront pour le candidat A ou le candidat B. Soient p et $\hat{p}$ les proportions d'électeurs qui préfèrent le candidat A dans la population et dans l'échantillon.

(a) On veut tester deux hypothèses rivales H_0: $p = 0,5$ et H_1: $p \neq 0,5$. Supposez que l'on décide de rejeter H_0 si $|\hat{p} - 0,5| > 0,02$.

   (i) Quel est le seuil de signification du test?

   (ii) Calculez la puissance du test quand $p = 0,53$.

(b) Dans l'échantillon, on trouve $\hat{p} = 0,54$.

   (i) Testez l'hypothèse H_0: $p = 0,5$ contre l'hypothèse alternative H_1: $p \neq 0,5$ en utilisant un seuil de confiance de 5%.

   (ii) Testez l'hypothèse H_0: $p = 0,5$ contre l'hypothèse alternative H_1: $p > 0,5$ en utilisant un seuil de confiance 5%.

   (iii) Construisez un intervalle de confiance à 95% pour p.

   (iv) Construisez un intervalle de confiance à 99% pour p.

   (v) Construisez un intervalle de confiance à 50% pour p.

(c) Supposez que l'enquête est réalisée 20 fois avec des électeurs choisis chaque fois de façon indépendante. Pour chacune de ces enquêtes, on réalise un intervalle de confiance à 95% pour p.

   (i) Quelle est la probabilité que la vraie valeur de p soit contenue dans la totalité des 20 intervalles de confiance?

   (ii) Quelle est le nombre espéré d'intervalles de confiance qui contiennent la vraie valeur de p?

(d) Dans le jargon des enquêtes, la "marge d'erreur" est définie comme $1,96$ x Ecart-type($\hat{p}$), c'est-à-dire que c'est la moitié de la longueur de l'intervalle de confiance à 95%. Supposons qu'on veuille faire un enquête qui ait une marge d'erreur maximale de 1%. C'est à dire que l'on aimerait avoir $P(|\hat{p} - p| > 0,01) \leq 0,05$. Quelle doit être la taille minimale, n, de l'échantillon, si l'enquête est faite avec un échantillon aléatoire simple?

Exercice 10

Pour voir si une entreprise pratique de la discrimination salariale entre les hommes et les femmes, on prend un échantillon aléatoire de 100 hommes et 64 femmes qui occupent des postes similaires. La table ci-dessous résume les salaires mensuels pour cet échantillon.

	Salaire moyen ($\bar{Y}$)	Ecart-type (s_Y)	n
Hommes	3.100	200	100
Femmes	2.900	320	64

(a) Que peut-on conclure à partir de ces données? Y a-t-il discrimination salariale statistiquement significative entre les hommes et des femmes? (Pour répondre à cette question, établissez tout d'abord les hypothèses nulle et alternative; ensuite, calculez la statistique t, calculez la p-valeur associée à ce test et finalement, utilisez cette p-valeur pour répondre à la question.)

(b) Ces données suggèrent-elles qu'il y a discrimination salariale dans cette entreprise? Justifiez votre réponse.

SOLUTIONS:
1. a) 84%; b) 50%.
2. a) (27.4,30.6); b) $n \geq 46$.
3. a) 0,2; b) 0,4; c) 0,3; d) 0,8571.
4. a) $\bar{X} = 261$; 4EUR, $n = 81$; b) $\varepsilon_{max} = 9,3$EUR.
5. a) 0,341; b) 0,659; c) $E(Y|X = 1) = 0,348$; $E(Y|X = 0) = 0,1567$; d) 0,652 et 0,8043; e) 0,9439 et 0,0561; f) Non, p. ex. $P(X = 0|Y = 0) = 0,0561 \neq 0,046 = P(X = 0)$
6. a) 0,6326; b) 0,05; c) 0,950; d) 0,950; e) $t_{[1]} = N(0,1)$ et 20 degrés de liberté est déjà très grand; f) 0,90; g) 0,10; h) 0,05; i) 0,05; j) $F_{[10;\infty]}d = \chi^2_{10}/10$.
7. a) i) 0,27; ii) 0,11; b) $n \leq 9220$.
9. a) i) 0,19; ii) 0,74; b) i) $t = 0,1$ et l'on rejette à 5%, ii) $P(t > 2,61) = 0,004$ et l'on rejette à 5%, iii) (0.51,0.57), iv) (0.5,0.58), v) (0.53,0.55); c) i) 0,36, ii) 95% des 20 intervalles de confiance; d) Si $n > 9604$, la marge d'erreur est moins que $0,01$ pour tout p.