Génie Biologique - Statistiques S2

Résumé de cours

Table des matières

— Interrogation de 1h en TD à mi-route
— Les séances de TP de PPP sont intégralement absorbées comme TPs de stats
— La dernière séance de TP servira pour l'interrogation (la note sera utilisée comme note de PPP)

Résumé des fonctions Excel / Calc

Fonctions de base

- max, min, moyenne, racine : rien à signaler
- var, ecartype : empirique ou estimé ?

Lois statistiques

- loi.normale(.standard) : cumulative = vrai pour avoir l'aire au lieu de la valeur de la fonction
- loi.normale(.standard).inverse
- loi.f :
- inverse.loi.f
- loi.student
- loi.student.inverse
- loi.khideux
- khideux.inverse

A chacune de ces fonctions se posent les questions fondamentales suivantes :

- Probas ou quantiles ?
- Probas d'être inférieur ou d'être supérieur au paramètre donné ?
- Quantiles unilatéraux ou bilatéraux ?

Fonctions techniques

- somme.ecarts.carres(plage)
- sommeprod(colonne1 ; colonne2) et sommeprod(colonne1 ; colonne1 ; colonne1) pour la moyenne et la variance avec effectifs
- sommeprod((cellule < plage)*1 ; (plage<=cellule)*1) pour les effectifs de classe (le *1 sert à convertir des "vrai / faux" en 0 / 1)

1 Statistiques descriptives (2h + 2TP)

— Représentation des données brutes
— Synthèse par indicateurs statistiques

1.1 Vocabulaire et présentation des données

— Concept de variable aléatoire
— Représentation graphique des données

1.1.1 Vocabulaire général

— Population : ensemble sur lequel porte l'étude. Ex. : les Français de 2 à à 35 ans, les canettes de coca ...
— Échantillon : (petite) partie de la population sur laquelle on a effectué des observations. Ex. : 189 personnes interrogées, Une palette de 200 canettes ...
— Individu : élément de base de l'échantillon. Ex. : une personne, une canette ...
— Taille de l'échantillon : nombre d'individus qui le composent. Généralement notée n
— Variable aléatoire : caractère étudié sur la population (et observé sur l'échantillon). Ex : âge, sexe, ville, volume, diamètre ...
— Modalités : valeurs que peut prendre la v.a. étudiée. Ex. : 20 ans, 21 ans, 22 ans ...35 ans, G,F, Paris, Strasbourg, Mulhouse, 15 ml, 33 ml, 50 ml, 4 cm, 5 cm ...

1.1.2 Variables aléatoires

— X : $\{\textit{événements}\} \mapsto$ modalités de X
— L'événement en question est l'observation d'un individu ; la modalité est la valeur relevée lors de cette observation
— X n'est *pas* un événement, $\mathbb{P}(X)$ n'a *pas* desens
— $\mathbb{P}$ (événements tels que X prenne certaines valeurs) est généralement abrégé en $\mathbb{P}(X = \text{modalités})$

V.A quantitative

Les modalités sont des quantités
— discrète : l'ensemble des modalités est dénombrable, les individus sont bien distincts
— continue : l'ensemble des modalités n'est pas dénombrable

Remarques

— En pratique, on dira que les unités physiques sont continues (la masse, le volume, le temps, l'âge ...) ; les comptages seront discrets (le nombre de canettes) ...)
— En pratique, on regroupe toujours les modalités continues en classes, c'est à dire des intervalles, et on travaille alors avec un ensemble discret de classes, par exemple : l'âge compté en années. Mais la variable aléatoire reste continue.

V.A. qualitative

Les modalités ne sont pas des quantités
— ordinale : il y a un ordre naturel pour les modalités. Ex. : pas d'accord, moyennement d'accord, tout à fait d'accord, froid, tiède, chaud, bouillant ...
— nominale : il n'y a pas d'ordre naturel. Ex. : G,F, Noir, Blanc, loup, renard, reloup ...

Remarque En pratique, il n'est pas rare de coder les modalités avec un nombre, par ex. 1 pour Homme et 2 pour Femme. La V.A. ne devient pas ordinale pour autant, et encore moins quantitative.

1.1.3 Présentation des données

L'ensemble des données brutes est généralement peu digeste. On commence alors par établir un tableau des effectifs n_i pour chacune des modalités x_i.
— Pour une V.A. quantitative continue on comptera toujours les effectifs par classe
— Pour les autres types de V.A. on peut aussi faire un regroupement par classes s'il y a trop de modalités.

Variable quantitative

Pour une V.A. quantitative, les valeurs doivent être reportées à l'échelle et les axes doivent être précisés.
— discrète : diagramme en bâtons de hauteur n_i et sans épaisseur, pour qu'on voie bien qu'on ne déborde pas de la modalité précise
— continue : histogramme.
  — Abscisses : bornes des classes à l'échelle
  — Ordonnées : densités de population $= \frac{\text{effectif de la classe}}{\text{amplitude de la classe}}$ ou l'amplitude d'une classe est la différence entre ses deux bornes.

  Remarques
  — Explication géométrique : pour chaque classe, l'aire du rectangle ainsi obtenu est alors égale à l'effectif de cette classe.
  — Explication visuelle : pour chaque classe, il faut placer n_i petits carrés individuels. On en place alors autant qu'on peut au sol, puis on fait des étages en les empilant.

Variable qualitative

— Les effectifs sont généralement représentés en barres épaisses bien séparées et équidistantes (puisqu'il n'y a pas d'échelle).
— La légende des modalités est indispensable

— Quand on parle de proportions, les barres peuvent être empilées en un diagramme uni-colonne
— Les camemberts, et plus généralement tout élan artistique, sont fortement déconseillés dans ce cadre précis, car nuisant à la comparaison des effectifs entre eux.

1.2 Indicateurs empiriques

Informations synthétiques pour des V.A. quantitatives : plus lisibles que l'ensemble des mesures, mais forcément moins complètes

1.2.1 Indicateurs de position empiriques (pour l'échantillon)

— Mode empirique : modalité du plus grand effectif (on peut avoir plusieurs modes ex-aequos)
— Médiane empirique : modalité qui partage l'échantillon ordonné en deux parties de même effectif. Si on range les modalités dans l'ordre croissant, ce sera la modalité de rang immédiatement supérieur à $n/2$ (n étant la taille de l'échantillon)
— Quartiles empiriques : modalités Q_1, Q_2 et Q_3 qui partagent l'échantillon ordonné en 4 parties de même effectif (Q_2 est donc la médiane). On appelle aussi parfois espace inter-quartile la grandeur $(Q_3 - Q_1)$.
— Déciles, Centiles ...quantiles empiriques : plus petites modalités telles que $\mathbb{P}(X \leqslant q_i) \geqslant \frac{1}{q+1}$, q étant le nombre de quantiles considéré.
Ex. : $\mathbb{P}(X \leqslant Q_1) \geqslant 25\%$, $\mathbb{P}(X \leqslant Q_2) \geqslant 50\%$ et $\mathbb{P}(X \leqslant Q_3) \geqslant 75\%$.
— Moyenne empirique : $\mu_E = \dfrac{1}{n} \displaystyle\sum_{i=1\ldots n} x_i$ où les x_i sont les valeurs de l'échantillon, comptées avec répétition.

Remarques

— Ne pas confondre moyenne et médiane. Les deux ont leur intérêt.
— Pour une V.A. continue dont on étudie les classes, on fait la moyenne sur les centres des classes

1.2.2 Indicateurs de dispersion

— Étendue empirique : étalement des données (max des modalités - min des modalités)
— Variance empirique : $\sigma_E^2 = \dfrac{1}{n} \displaystyle\sum_{i=1\ldots n} (x_i - \mu_E)^2 = \frac{1}{n} \left(\sum_{i=1\ldots n} (x_i)^2 \right) - \mu_E^2$
— Ecart-type empirique : écart moyen des valeurs par rapport à la moyenne $\sigma_E := \sqrt{\sigma_E^2}$
— Coefficient de variation empirique : écart-type relatif $\dfrac{\sigma_E}{\mu_E}$

Remarques

— la deuxième formule pour la variance est peut-être plus simple à manipuler, mais plus sensible aux erreurs d'arrondi à cause des carrés qui augmentent ces erreurs. Si on utilise la deuxième formule, il faut utiliser 4 chiffres après la virgule partout, alors que 2 peuvent suffire pour la première.
— Pour une V.A. continue dont on a regroupé les modalités par classes, on utilise les centres des classes.

— La variance est souvent plus simple pour les calculs, mais l'écart-type est généralement plus utile : il à la même unité que la moyenne, on peut ainsi mieux se représenter sa valeur en tant que distance moyenne des valeurs par rapport à la moyenne.
— Le coefficient de variation est un coefficent. En effet, on déduit de la remarque précédente qu'il est adimensionné. Ceci est utile pour comparer les variations d'échantillons qui ne sont pas à la même échelle.

Exemple Avec les valeurs $\{1, 2, 3\}$, la moyenne est trivialement égale à 2.
— On calcule très facilement de tête la variance avec la première formule : $\frac{2}{3}$ (rappelons que dans des cas moins basiques, on recommande l'utilisation de la seconde formule).
— Si les valeurs sont des mètres, la variance vaut 0.6m^2 ; une surface. Le lien avec les valeurs n'est pas évident. Alors que l'écart type vaut 0.82 mètres : en moyenne, les 3 points de l'exemple points sont à 0.82 mètres du point "central".
— L'écart-type est une valeur hors contexte : on ne sait pas si c'est beaucoup. Le coefficient de variation donne enfin l'écart-type dans son contexte : la distance moyenne entre les points et la valeur centrale vaut 41% de la valeur centrale, ce qui est très fort.

Statistiques descriptives - Exercices

Exercice 1 Pour la moitié des questions du questionnaire en annexe, nommez la variable, dites sa nature et le nombre de modalités (prenez par exemple les questions paires, ou les impaires ...)

Exercice 2 Voici un tableau de pourcentages obtenus pour la variable "Mode de logement" :

Modalité	%
Cité U	4.8
HLM	16.4
Résidence	38.6
Maison	28.6
Autre	11.6

Sachant que la taille de l'échantillon est $n = 189$, retrouver les effectifs pour chaque modalité.

Exercice 3 Pour la variable "Propriétaire/locataire" on recense 134 locataires et 55 propriétaires.

1. Dresser un tableau des fréquences
2. Représenter ces données sous forme de diagramme circulaire.

Exercice 4 Pour la variable "Lieu", représentez en diagramme "barres" les valeurs suivantes :

x_i	n_i
Centre	87
Banlieue	30
Village	32
Cité	30
Autre	10

Exercice 5 Pour la variable "Type de logement", on a relevé les données suivantes :

4 ,2 ,4 ,4 ,2 ,3 ,4 ,1 ,2 ,4 ,3 ,3 ,4 ,4 ,4 ,2 ,1 ,3 ,2 , 4,

où "1" = "chambre universitaire", "2" = "T1", "3" = "T2" et "4" = "T3 et +".

1. Dresser le tableau donnant les effectifs, fréquences et pourcentages .
2. Représentez cette synthèse sous forme de graphique uni-colonne.

Exercice 6 Calculer, quand cela est possible, le mode, la médiane, la moyenne, la variance, l'écart-type et le coefficient de variation sur toutes les variables étudiées ci-dessus.

Exercice 7 – A partir d'ici, on pourra (re)faire ces exercices en TP Les notes obtenues par une classe d'élèves de 5ème lors d'un devoir de Français fournissent le tableau suivant :

x_i	4	5	6	7	8	9	10	11	12	14
n_i	2	3	5	3	2	2	4	4	3	2

1. Préciser la variable étudiée ainsi que son type.
2. Compléter le tableau des effectifs cumulés
   — En TP : fonction SOMME en utilisant la poignée de recopie et les $ à bon escient : SOMME($A1 :F1), ou bien fonction somme en diagonale : l'effectif en cours + le cumulé précédent
3. Calcule mode, médiane, moyenne, variance, écart-type et coefficient de variation.
   — En TP : les fonctions excel du même nom ne fonctionnent pas avec des effectifs. On ne calculera alors pas le mode ni la médiane. Par contre, pour la moyenne on utilisera la fonction SOMMEPROD sur les colonnes valeurs et effectifs et pour la variance on utilisera la foncion SOMMEPROD sur les colonnes valeurs, valeurs et effectifs (pour calculer la moyenne des carrés).
4. Réaliser le diagramme en bâtons représentant la distribution de ces valeurs. Attention : il s'agit d'une V.A. quantitative discrète et la note 14 ne doit donc pas être accolée à la note 12...

Exercice 8 Pour la variable "Age" (regroupée en classes par année), on a extrait les observations suivantes :

21, 20, 19, 18, 20, 20, 19, 22, 18, 19, 20, 21, 16, 20, 25, 19, 22, 23, 20, 19, 20, 24, 16, 21, 18.

1. Que signifie par exemple l'observation 18 ?
2. Calculer la moyenne, la mediane, la variance et l'écart-type de cette série de données
   — En TP : fonctions MOYENNE, MEDIANE, VAR.P et ECARTYPE.P
3. Dresser le tableau des effectifs, fréquences et densités en utilisant comme modalités les classes suivantes : $[16; 19[$, $[19; 20[$, $[20; 21[$, $[21; 25]$.
   — En TP : Lister les bornes des classes, faire calculer les amplitudes et faire calculer les densités avec la fonction SOMMEPROD : SOMMEPROD((B2<$A$2 :$A$26)*1 ; ($A$2 :$A$26<B3)*1)
4. Tracer un histogramme de ces valeurs.

Exercice 9 On considère la variable X "Temps vécu dans le logement" pour laquelle on a obtenu le tableau d'effectifs suivants :

x_i	[0 ;1[	[1 ;2[	[2 ;3[	[3 ;5[	[5 ;11[	[11 ;16[	[16 ;21[	[21 ;26[
n_i	35	36	32	25	20	18	16	7

1. Préciser la variable étudiée ainsi que son type.
2. Calcule mode, médiane, moyenne, variance, écart-type et coefficient de variation.
3. Représenter l'histogramme.

Exercice 10 Statistique descriptive Un service de maternité s'interroge sur le poids de naissance des nouveaux-nés. Pour répondre à cette question, les poids en grammes de 100 nouveau-nés sont relevés de la manière suivante :

Classes de poids (en kg)	[2 ; 2.5[	[2.5 ;3[	[3 ;3.5[	[3.5 ;4[	[4 ;4.5]	[4.5 ;5]
Nombre d'enfants	6	22	33	31	7	1

1. Calculer les fréquences, les effectifs cumulés et les fréquences cumulées.
2. Donner une représentation graphique des fréquences cumulées.
3. Repérer dans quelles classes se trouvent le premier quartile, la médiane et le troisième quartile.
4. En utilisant le centre de chaque classe, calculer la moyenne, la variance et l'écart type du poids à la naissance pour l'échantillon.

Exercice 11 Exemple pratique Voici les notes d'une interrogation de statistiques :

5.38	7.27	6.36	7.27	14.23	13.64	6.82	dem	abs	13.08	6.36	8.18	1.15	8.64	14.5
1.92	14.55	9.55	10	13.08	dem	9.09	12.31	13.08	17.27	12.31	12.31	8.64	7.69	9.09
abs	11.54	13.08	8.08	14.55	6.36	12.73	6.92	17.27	dem	13.64	17.31	13.64	7.69	7.69
11.54	5	13.46	16.54	6.54	11.82	5	20	8.08						

Cette quantité de données étant fastidieuse à traiter, on ne le fera qu'avec Excel.

1. Avec les fonction nb, max, min, calculer l'effectif, le maximum, et le minimum de ces valeurs.
2. Avec les fonctions du même nom, calculer la moyenne et la médiane.
3. Avec la fonction $var.p$, calculer la variance, puis l'écart-type et enfin le coefficient de variation.

Réponse 1

1. Age, quantitative, continue, infinité de valeurs mais à priori regroupées par classes de 1 an, avec un nombre fini de classes (<150 par exemple)
2. Sexe, qualitative, nominale, 2
3. Origine du père/mère/soi-même, qualitative, nominale, 2
4. Lieu d'habitation, qualitative, nominale, 5 (si on regroupe tous les "autre" ensemble)
5. Type, qualitative, ordinale, 4
6. Mode, qualitative, nominale ("autre" est inclassable), 5
7. Durée, quantitative, continue, infinité de valeurs mais à priori regroupées en un nombre fini de classes
8. Rapport, qualitative, nominale, 2
9. Situation, qualitative, nominale, 5
10. Famille/amis, qualitative, nominale, 2 (ou quantitative, discrète, grand si on s'intéresse à la question "nombre de ...")
11. Fréquentation, qualitative, nominale, 2
    Relations, qualitative, ordinale, 4
12. Plusieurs modalités possibles en même temps : ce n'est pas une variable aléatoire
13. Plusieurs modalités possibles en même temps : ce n'est pas une variable aléatoire
14. Plusieurs modalités possibles en même temps : ce n'est pas une variable aléatoire
15. Plusieurs modalités possibles en même temps : ce n'est pas une variable aléatoire
16. xxx, qualitative, ordinale, 4 (pour chacune des lignes)
17. xxx, qualitative, ordinale, 5 (pour chacune des lignes)

Réponse 2 9, 31, 73, 54, 22

Réponse 5

x_i	n_i	f_i	%
Chambre U	2	0.1	10
T1	5	0.25	25
T2	4	0.2	20
T3 et +	9	0.45	45
TOTAL	N=20	1	100

Réponse 6

Exercice	mode	moyenne	médiane	variance	écart-type	coeff. variation
2	Résidence	-	-	-	-	-
3	Locataire	-	-	-	-	-
4	Centre	-	-	-	-	-
5	T3 et +	-	T2	-	-	-
6	$[21;25]$	20.5	20.5	4	2	0.09

Réponse 7

1. Variable quantitative discrète.
2. Effectifs cumulés : 2, 5, 10, 13, 15, 17, 21, 25, 28, 30.
3. Mode : 6, Médiane : 9, Moyenne : 8,533, Variance : 8,248, Ecart-type : 2,872, Coefficient de variation : 0,336.

Réponse 8 L'observation 18 signifie que l'individu est dans la classe d'âge $[18;19[$.

x_i	n_i	f_i	Densité
$[16;19]$	5	0.2	1.67
$[19;20]$	5	0.2	5
$[20;21]$	7	0.28	7
$[21;25]$	8	0.32	2
TOTAL	N=25	1	

Réponse 9

1. Variable quantitative continue regroupée par classes de mois.
2. Classe modale : $[1;2[$, Classe médiane : $[2;3[$, Moyenne : 5,899, Variance : 42,403, Ecart-type : 6,5, Coefficient de variation : 1,104.

Réponse 10

1.

Classes de poids (en kg)	[2 ; 2.5[	[2.5 ;3[	[3 ;3.5[	[3.5 ;4[	[4 ;4.5]	[4.5 ;5]
Nombre d'enfants	6	22	33	31	7	1
%	6	22	33	31	7	1
Effectifs cumulés	6 28	61	92	99	100	

2.

3. $Q1 \in [2,5; 3[$, $Q2 \in [3; 3,5[$, $Q3 \in [3,5; 4[$.
4. Moyenne : 3,32g, Variance : 0,28g^2, Ecart-type : 0,53g, Coefficient de variation : 0,1594

Réponse 11 Exemple pratique

— effectif = 49
— max = 20
— moyenne = 10.46
— mediane = 10
— min = 1.15
— var = 16.86
— ecart-type = 4.11
— coeff. variation = 39 % : les données sont assez largement étalées

Nous sommes étudiants en L2 de Psychologie à l'Université de Toulouse le Mirail.

Dans le cadre d'une UE de méthodologie nous faisons une étude sur l'attachement au quartier.

Ce questionnaire est anonyme.

Consigne : répondez à ce questionnaire en indiquant la réponse de votre choix ou en cochant dans la case appropriée.

1- Age : ……..

2- Sexe : ☐ Homme ☐ Femme

3- Quel est le pays d'origine de votre père : né en France ☐ né à l'étranger ☐

Quel est le pays d'origine de votre mère : née en France ☐ née à l'étranger ☐

Quel est votre pays de naissance: né en France ☐ né à l'étranger ☐

4- Quel est votre lieu d'habitation ?

Nom de votre ville : …….

Habitez vous : en centre ville ☐ en banlieue ☐

dans un village ☐ dans une cité ☐

Autres : ☐ ……

5- Votre type de logement actuel :

☐ chambre universitaire

☐ T1

☐ T2

☐ T3 et +

6-Votre mode de logement actuel :

☐ en cité Universitaire

☐ en HLM

☐ dans une résidence

☐ dans une maison

☐ autres :……

7- Depuis combien de temps vivez-vous dans ce logement ? ……

8. Vous êtes : Locataire ☐ propriétaire ☐

9- Dans ce logement vous vivez :

☐ seul

☐ en couple

☐ en famille (combien êtes vous ?.....)

☐ chez vos parents (combien êtes vous ?.....)

☐ en colocation (combien êtes vous ?.....)

10- Dans votre quartier :

- Avez-vous de la famille qui habite dans le quartier ?

Oui ☐ Non ☐

Si oui combien ? ……..

- Avez-vous des amis qui habitent dans le quartier ?

Oui ☐ Non ☐

Si oui combien ? ……..

11 Fréquentez vous des voisins dans votre quartier ?

Oui ☐ Non ☐ Si oui combien ? ……..

Quelle est la nature de vos relations avec vos voisins ? cochez la réponse de votre choix.

Très mauvaise	mauvaise	bonne	Très bonne

12- Dans votre quartier y a-t'il des commerces?
un marché ☐ une boucherie ☐
une boulangerie ☐ un supermarché ☐
une pharmacie ☐ une librairie ☐ autres ☐…….

13- Dans votre quartier y a-t'il des lieux culturels ?
théâtre ☐ cinéma ☐ associations ☐ bibliothèque ☐
musée ☐ autres ☐……

14- Dans votre quartier y a-t'il des lieux de sortie ?
fast-food ☐ restaurant ☐ boite de nuit ☐ bar ☐ autres ☐……

15- Dans votre quartier y a-t'il des lieux de formation ?
école☐ collège☐ lycée☐ université ☐ formation pour adulte ☐ autres ☐……

16- <u>Dans votre quartier,</u> dans quels lieux vous rendez-vous le plus souvent ?
(codez de 1 très souvent à 4 rarement)
_____ les commerces
_____ les lieux culturels
_____ les lieux de sortie
_____ les lieux de formation

17- Cochez la case qui correspond à votre réponse :

Que représente votre quartier pour vous ?	**Tout à fait en désaccord**				**Tout à fait d'accord**
1 Pour y vivre, c'est le quartier idéal.	**1**	**2**	**3**	**4**	**5**
2 Ce quartier fait partie de moi-même.	**1**	**2**	**3**	**4**	**5**
3 Je suis très attaché(e) à certains endroits de ce quartier.	**1**	**2**	**3**	**4**	**5**
4 Il me serait très difficile de quitter définitivement ce quartier.	**1**	**2**	**3**	**4**	**5**
5 Je pourrais facilement quitter ce quartier.	**1**	**2**	**3**	**4**	**5**
6 Je n'aimerais pas à avoir à quitter ce quartier pour un autre.	**1**	**2**	**3**	**4**	**5**

2 Lois de Variables aléatoires (3h + 1TP)

Modèles mathématiques basiques pour des comportements aléatoires.

2.1 Probabilités

— $\mathbb{P} : \{\textit{événements}\} \mapsto [0\,;\,1]$
— $\mathbb{P}\left(\bar{A}\right) = 1 - \mathbb{P}(A)$
— $\mathbb{P}(A \text{ ou } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ et } B)$
— $\mathbb{P}(A \text{ et } B) = \mathbb{P}(A|B)\,\mathbb{P}(B) = \begin{cases} 0 \text{ si } A \text{ et } B \text{ sont incompatbiles} \\ \mathbb{P}(A)\,\mathbb{P}(B) \text{ si } A \text{ et } B \text{ sont indépendants} \end{cases}$

2.2 Variables aléatoires (V.A.)

— $X : \{\textit{événements}\} \mapsto$ modalités de X (souvent notées x, ou même x_i si on peut les dénombrer)
— $\mathbb{P}$(événements tels que X prenne certaines valeurs) est généralement abrégé en $\mathbb{P}(X = \textit{valeurs})$
— Espérance : $\mu(X) = \int x * \textit{densite}(x)$ ou $\sum x * \mathbb{P}(X = x)$
— Variance : $\sigma^2(X) = \mu\left((X - \mu(X))^2\right) = \mu(X^2) - (\mu(X))^2$

Remarques

— Pour les variances et la covariance il faut calculer les moyennes avec 4 chiffres après la virgule
— X n'est *pas* un événement – $\mathbb{P}(X)$ n'a *pas* de sens
— X est une *fonction* qui transforme une expérience en modalité, qu'on appelle réalisation de X. Par exemple, on peut faire l'expérience de jeter une pièce en l'air : la variable aléatoire peut porter sur la prédiction du côté de la pièce qui finira sur le dessus ; mais une fois que la pièce a atterri on a une réalisation (c'est soit pile soit face, mais il n'y a plus rien à prédire)
— Espérance et variance d'une V.A. utilisent l'ensemble des modalités possibles pour X : ce sont la moyenne et la variance sur l'ensemble de la *population*. Par opposition, moyenne empirique et variance empirique sont calculées sur un échantillon : un ensemble de réalisations.

2.3 Lois de variables aléatoires discrètes

Une variable aléatoire est dite discrète si ses modalités sont dénombrables. La loi d'une V.A. discrète X est la donnée des valeurs $\mathbb{P}(X = x_i)$ pour toutes les modalités x_i de X.

2.3.1 Loi uniforme sur {1, . . ., n} : $X \sim \mathcal{U}(n)$

Cette loi modélise l'équiprobabilité sur n évènements.

- $\mathbb{P}(X = k) = \begin{cases} \frac{1}{n} \text{ si } 1 \leqslant k \leqslant n \\ 0 \text{ sinon} \end{cases}$
- $\mu(X) = \frac{n+1}{2}$
- $\sigma^2 = \frac{n^2-1}{12}$

2.3.2 Loi de Bernouilli : $X \sim \mathcal{B}(n; p)$

Cette loi modélise la somme des succès après n tentatives ayant chacune p chances de réussite.

- $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$
- $\mu(X) = np$
- $\sigma^2 = np(1-p)$

2.3.3 Loi de Poisson : $X \sim \mathcal{P}(\lambda)$

Cette loi modélise la somme des succès après un temps assez long en sachant qu'en moyenne on a λ succès sur cette durée.

- $\mathbb{P}(X = k) = \frac{\lambda^k}{e^\lambda k!}$
- $\mu(X) = \lambda$
- $\sigma^2 = \lambda$

2.4 Lois de variables aléatoires continues

Une variable aléatoire est dite continue si ses modalités sont des grandeurs physiques (le temps, le poids, ...). La loi d'une V.A. continue X est la donnée des valeurs $\mathbb{P}(X < x)$ pour toutes les modalités x de X.

2.4.1 Loi uniforme sur $[a; b]$: $X \sim \mathcal{U}[a; b]$

Cette loi modélise l'équiprobabilité sur un intervalle.

- $\mathbb{P}(X < x) = \begin{cases} 0 \text{ si } x \leqslant a \\ \frac{x-a}{b-a} \text{ si } a \leqslant k \leqslant b \\ 1 \text{ si } x \geqslant b \end{cases}$
- $\mu(X) = \frac{a+b}{2}$
- $\sigma^2 = \frac{(b-a)^2}{12}$

2.4.2 Loi exponentielle de paramètre λ

Cette loi modélise la durée de vie d'un phénomène sans mémoire : à un instant donné, la probabilité de sa durée de vie ne dépend pas de son age.

- $\mathbb{P}(X < x) = \begin{cases} 0 \text{ si } x \leqslant 0 \\ 1 - e^{-\lambda x} \text{ sinon} \end{cases}$
- $\mu(X) = \frac{1}{\lambda}$
- $\sigma^2 = \frac{1}{\lambda^2}$

2.4.3 Loi normale standard : $X \sim \mathcal{N}(0;1)$

Cette loi modélise quasiment tout le reste !

- $\mathbb{P}(X < x)$ sera donné par un tableau pré-calculé ou l'ordinateur
- $\mu = 0$
- $\sigma^2 = 1^2$
- Cette loi est symétrique : $\mathbb{P}(X < -x) = \mathbb{P}(X > x)$

2.4.4 Loi normale générale : $X \sim \mathcal{N}(\mu;\sigma)$

Si $X \sim \mathcal{N}(\mu;\sigma)$, alors on va utiliser une variable aléatoire qui servira d'intermédiaire de calcul : $Y = \left(\frac{X-\mu}{\sigma}\right)$. On a alors $Y \sim \mathcal{N}(0;1)$ et

$$\mathbb{P}(X < x) = \mathbb{P}\left(Y < \frac{x-\mu}{\sigma}\right)$$

Approximations classiques :

- $\mathbb{P}(\mu - \sigma < X < \mu + \sigma) \approx 68\%$
- $\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%$

Lois de Variables aléatoires - Exercices

Exercice 1 Loi binomiale

1. Une maladie congénitale a une prévalence de 1% chez les nouveaux-nés. On enregistre dans une clinique 10 naissances durant un week-end. Quelle est la probabilité pour que deux au moins de ces nouveaux-nés soient atteints de la maladie ?
2. Selon les statistiques américaines, le nombre de noyades accidentelles en un an est de 2 pour 100000 habitants. Quelle est la probabilité, pour une ville de 200000 habitants, de n'avoir aucune noyade durant une année ?

Exercice 2 Loi de poisson

1. Le nombre de pannes annuel d'une certaine machine suit une loi de poisson de paramètre $\lambda = 3$. Quelle est la probabilité pour que cette machine ait au moins 2 pannes dans l'année ?

Exercice 3 Loi exponentielle

1. Soit X une variable aléatoire qui suit la loi exponentielle de paramètre $\frac{1}{2}$. Calculer $\mathbb{P}(1 \leqslant X \leqslant 2)$.
2. Trouver λ, sachant que la variable X suit une loi exponentielle et que $\mathbb{P}(X \leqslant 50) = 1 - \frac{1}{e}$.
3. Avec ce paramètre, que vaut $\mathbb{P}(X \geqslant 25)$
4. De même, que vaut $\mathbb{P}(X \geqslant 50)$

Exercice 4 Loi normale

1. Soit X une variable aléatoire dont la loi est $\mathcal{N}(20, 5)$.
   a) Que vaut $\mathbb{P}(X \leqslant 20)$?
   b) Si $\mathbb{P}(X \leqslant 15) = 0.16$, que valent $\mathbb{P}(X \geqslant 15)$ et $\mathbb{P}(X \geqslant 25)$?
2. Soit X une variable aléatoire dont la loi est $\mathcal{N}(20, 5)$. Quelle est la valeur de a telle que :
   a) $\mathbb{P}(X \leqslant a) = 0.975$,
   b) $\mathbb{P}(X \geqslant a) = 0.005$,
   c) $\mathbb{P}(20 - a \leqslant X \leqslant 20 + a) = 0.95$.
3. Soit X une variable aléatoire dont la loi est $\mathcal{N}(176, 6)$. Déterminer a tel que $\mathbb{P}(176 - a \leqslant X \leqslant 176 + a) = 0.68$.
4. Soit X une variable aléatoire dont la loi est $\mathcal{N}(-5, 1)$. Déterminer a tel que $\mathbb{P}(-5 - a \leqslant X \leqslant -5 + a) = 0.95$.
5. On considère une V.A. X normalement distribuée, de moyenne 3 et d'écart-type inconnu σ. Déterminer σ, sachant que $\mathbb{P}(X > 0) = 0.9$.
6. (Plus difficile) Soit X une (autre) V.A. distribuée suivant une loi normale dont la moyenne et l'écart-type sont inconnus. Sachant que $\mathbb{P}(X > 0) = 0.3$ et $\mathbb{P}(X < -1) = 0.4$, déterminer les paramètres de la loi de X.

Exercice 5 Loi du χ^2 et loi T Soient X une V.A. suivant une loi du χ^2 à 24 degrés de liberté et S une V.A. suivant une loi de Student (T) à 10 degrés de liberté.

1. Que vaut $\mathbb{P}(X < 13.848)$? et $\mathbb{P}(X < 36.415)$? En déduire ce que calcule l'ordinateur quand on appelle la fonction "loi.khideux(36.415 ;24)" (cette question est fondamentale !).
2. Trouver a tel quel $\mathbb{P}(X < a) = 1\%$.
3. Trouver b tel que $\mathbb{P}(X > b) = 1\%$.
4. Que vaut $\mathbb{P}(S < 1.812)$? $\mathbb{P}(S > 1.812)$
5. Trouver a tel que $\mathbb{P}(-a < S < a) = 99\%$
6. Trouver b tel que $\mathbb{P}(S < b) = 2.5\%$

Réponse 1 Loi binomiale

1. 0.00427
2. 0.018315

Réponse 2 Loi de poisson

1. 0.8009

Réponse 3 Loi exponentielle

1. $-\frac{1}{e} + \frac{1}{\sqrt{e}}$.
2. $\lambda = \frac{1}{50}$.
3. $\frac{1}{\sqrt{e}}$.
4. $\frac{1}{e}$.

Réponse 4 Loi normale

1. a) 0.5
   b) 0.84 et 0.16
2. a) $a = 29.8$
   b) $a = 32.75$ (en réalité un peu plus, ceci est dû au manque de précision des tableaux)
   c) $a = 9.8$
3. $a = 6$ (en réalité un peu moins, ceci est une approximation rapide)
4. $a = 2$ (en réalité un peu moins, ceci est une approximation rapide)
5. $\sigma = 2.34$
6. $\sigma = 1.28$ et $\mu = -0.67$.

Réponse 5 Loi du χ^2 et loi T

1. 5% et 95%. De fait, la fonction "loi.khideux(36.415 ;24)" calcule la valeur $\mathbb{P}(X > 36.415)$ et il faut bien faire attention à la convention utilisée par le logiciel !
2. $a = 10,856$.
3. $b = 42,980$.
4. 95% et 5%
5. $a = 3.169$.
6. $b = -2.228$

3 Estimation et intervalles de confiance (3h)

Inférence des paramètres d'une population trop grande pour être recensée intégralement.

3.1 Estimation

Approximation sans garantie

3.1.1 Définitions

— L'estimation d'un paramètre d'une *population* est une valeur calculée sur un *échantillon* et censée être "proche" du paramètre estimé.
— L'estimateur est la fonction qui donne l'estimation à partir des valeurs de l'échantillon ; c'est une variable aléatoire.

Remarque "proche" a une définition rigoureuse dont voici l'idée :
— Si la taille de l'échantillon grandit, les estimations ainsi obtenues doivent se rapprocher du paramètre estimé.
— Si on prend de plus en plus d'échantillons différents, la moyenne des estimations ainsi obtenues doit se rapprocher du paramètre estimé.

3.1.2 Estimations principales

— Une "bonne" estimation de la proportion π de la *population* est $\pi_{est} = \pi_E$, où π_E est la proportion empirique.
— Une "bonne" estimation de l'espérance μ de la *population* est $\mu_{est} = \mu_E$, où μ_E est la moyenne empirique.
— Une "bonne" estimation de la variance σ^2 de la *population* est $\sigma^2_{est} = \dfrac{n}{n-1}\sigma^2_E$, où σ^2_E est la variance empirique.
— Une "bonne" estimation de l'écart-type σ de la *population* est $\sigma_{est} = \sqrt{\frac{n}{n-1}}\sigma_E$, où σ_E est l'écart-type empirique.

Remarques
— $\sigma^2(X) = \mu(X^2) - \big(\mu(X)\big)^2$
— $\sigma^2_E = \dfrac{1}{n}\sum(x_i - x_\cdot)^2$
— $\sigma^2_{est} = \dfrac{1}{n-1}\sum(x_i - x_\cdot)^2$

3.2 Intervalles de confiance (IDC) pour une V.A. $X \sim \mathcal{N}(\mu\ ;\ \sigma)$

Encadrement avec garantie

3.2.1 Définitions

- Un intervalle de confiance au risque α est un *intervalle* construit à partir des valeurs d'un échantillon tel que, sur l'ensemble des réalisations de la V.A. étudiée, exactement $\alpha\%$ des IDC qui découlent ne contiennent *pas* le paramètre étudié.
- $u(\alpha)$ est le quantile de la loi normale standard tel que $\mathbb{P}(U > u) = \alpha$, $U \sim \mathcal{N}(0\ ;\ 1)$
- $t(\alpha)$ est le quantile de la loi de Student à $(n-1)$ degrés de liberté tel que $\mathbb{P}(T > t) = \alpha$, $T \sim Student(n-1)$
- $k(\alpha)$ est le quantile de la loi du Khi-deux à $(n-1)$ degrés de liberté tel que $\mathbb{P}(K > k) = \alpha$, $K \sim \chi^2(n-1)$

3.2.2 Interprétation

- Le niveau de confiance $(1-\alpha)$ est le pourcentage d'intervalles qui contiendront le paramètre en prenant des échantillons différents ; et non pas la probabilité pour que le paramètre étudié soit dans l'intervalle calculé. C'est encore la différence entre la V.A. qui permet de calculer l'intervalle (la formule) et réalisation de cette V.A. (résultat de la formule avec l'échantillon particulier dont on dispose).
  Par exemple : le jeu des trois gobelets. On place un haricot sous un des gobelets et on les mélange. On a alors 33% de chances que le haricot soit sous le premier gobelet (niveau de confiance de la prédiction). Mais une fois le gobelet choisi, même si on ne regarde pas dessous, il n'y a plus de hasard : le haricot s'y trouve ou ne s'y trouve pas. De même, une fois les mesures faites et l'intervalle construit, il n'y a plus de hasard : la paramètre étudié s'y trouve ou ne s'y trouve pas, mais c'est définitif. Mais si on fait 100 fois la même expérience, 33% de nos choix devraient être bons.
- Un intervalle de confiance est un compromis entre faible risque d'erreur, pertinence de l'intervalle et petite taille de l'échantillon. On ne peut généralement pas avoir les trois en même temps.
- Les IDC qui vont suivre sont donnés pour quand on sait que X suit une loi normale. Ils sont aussi valables quand on ne le sait pas mais que n est "assez grand" : au delà de 30 c'est acceptable, au delà de 50 c'est correct et au delà de 100 c'est indiscernable.

3.2.3 IDC pour les valeurs de X

Aussi appelé intervalle de fluctuation.

- σ connu : $\left[\mu_{(est)} - u(\frac{\alpha}{2})\sigma\ ;\ \mu_{(est)} + u(\frac{\alpha}{2})\sigma\right]$
- σ estimé : $\left[\mu_{(est)} - t(\frac{\alpha}{2})\sigma_{est}\ ;\ \mu_{(est)} + t(\frac{\alpha}{2})\sigma_{est}\right]$

3.2.4 IDC pour la moyenne de X

- σ connu : $\left[\mu_{est} - u(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\ ;\ \mu_{est} + u(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right]$

— σ estimé : $\left[\mu_{est} - t(\frac{\alpha}{2})\frac{\sigma_{est}}{\sqrt{n}}\ ;\ \mu_{est} + t(\frac{\alpha}{2})\frac{\sigma_{est}}{\sqrt{n}}\right]$

Remarque On voit que la variance de la moyenne est n fois plus petite que la variance des valeurs

3.2.5 IDC pour la variance de X

— $\left[\frac{(n-1)\sigma_{est}^2}{k(\frac{\alpha}{2})}\ ;\ \frac{(n-1)\sigma_{est}^2}{k(1-\frac{\alpha}{2})}\right]$

Remarque On obtient celui pour l'écart-type en prenant la racine des bornes.

Estimation et intervalles de confiance - Exercices

Exercice 1 Estimation ponctuelle de la moyenne et de l'écart-type Lors d'un concours radiophonique, on note X le nombre de réponses reçues chaque jour. Durant les 10 premiers jours, on a obtenu :

$$\begin{array}{lllll} x_1 = 200 & x_2 = 240 & x_3 = 190 & x_4 = 150 & x_5 = 220 \\ x_6 = 180 & x_7 = 170 & x_8 = 230 & x_9 = 210 & x_{10} = 210 \end{array}$$

Déterminer une estimation ponctuelle de la moyenne et de l'écart-type.

Exercice 2 Estimation de variance Lors d'un contrôle d'une chaîne de médicaments, on s'intéresse au nombre de comprimés défectueux dans un lot. L'étude de 200 lots a donné les résultats suivants :

# de comprimés défectueux	0	1	2	3	4	5
Nombre de lots	75	53	39	23	9	1

1. Calculer la moyenne, le mode et les quartiles du nombre de comprimés défectueux pour cet échantillon de 200 lots.
2. Calculer la variance, l'écart-type et le coefficient de variation du nombre de comprimés défectueux pour cet échantillon de 200 lots.
3. Donner une estimation de la variance et de l'écart-type du nombre de comprimés défectueux sur l'ensemble de la production.

Exercice 3 Estimation par IDC de la moyenne Dans une station service, on suppose que le montant des chèques essence suit une loi normale de paramètres μ et σ. On considère un échantillon de taille $n = 51$ et on obtient une moyenne de 13€ et un écart-type (empirique) de 4€.

1. Donner une estimation de μ par un intervalle de confiance au niveau de confiance 95%.
2. Donner une estimation de σ par un intervalle de confiance au niveau de confiance 95%.

Exercice 4 Exemple pratique – suite Voici les indicateurs statistiques de base pour un contrôle de statistiques :

— effectif = 49
— max = 20
— moyenne = 10.4551
— mediane = 10
— min = 1.15
— var = 16.8648

— ecart-type = 4.1067
— coeff. variation = 39.28 %

1. Estimez la moyenne et la variance des notes en général.
2. Calculez l'intervalle de confiance à 95 % pour les valeurs (aussi appelé intervalle de fluctuation) des notes.
3. Calculez l'intervalle de confiance à 98 % pour les valeurs des notes (oui, on peut le faire avec seulement les tableaux).
4. Comparez avec le min et le max de l'échantillon.
5. [Avec Excel] Quel serait le pourcentage de risque exact pour que l'IDC aille précisément du min au max de l'échantillon ?
6. A l'inverse, quelle serait l'écart-type estimé pour la population si l'IDC à 95% pour les valeurs allait précisément du min au max de l'échantillon ?

Exercice 5 Estimation et IDC 1 On considère la variable X masse d'un ressort provenant d'une certaine fabrication. Cette variable suit une loi normale de moyenne μ et d'écart-type σ. On donne la répartition des masses de 219 ressorts :

X Masses (g)	[8,2 ;8,4[	[8,4 ;8,6[	[8,6 ;8,8[	[8,8 ;9,0[	[9,0 ;9,2[	[9,2 ;9,4[	[9,4 ;9,6[
Nb de ressorts	9	21	39	63	45	27	15

1. Donner une estimation ponctuelle de μ et σ.
2. Que dire de la qualité d'une estimation par Intervalle de Confiance dans ce cas ?

Exercice 6 Estimation et IDC 2 On veut estimer l'espérance mathématique μ d'une variable aléatoire Gaussienne X dont on connait l'écart-type $\sigma = 2,3$. Quelle est la taille minimale de l'échantillon de X qui est à prendre si l'on veut obtenir pour μ un intervalle de confiance de seuil 0,95 et dont la longueur ne dépasse pas 0,l.

Réponse 1 Estimation ponctuelle de la moyenne et de l'écart-type $\mu_{est} = 200$ et $\sigma_{est} = \sqrt{777.78} = 27.89$

Réponse 2 Estimation de variance

1. $\mu_E = 1.205$, mode = 0, effectifs cumulés = 75, 128, 167, 190, 199, 200 : $Q_1 = 0, Q_2 = 1$ et $Q_3 = 2$.
2. $\sigma_E = 1.214, CV = 1.007(> 100\%)$
3. $\sigma_{est} = \sqrt{1.4803} = 1.217$

Réponse 3 Estimation par IDC de la moyenne $\mu_{est} = 13, \sigma_{est} = 4.040$

1. $11.864 \leqslant \mu \leqslant 14.136$
2. $3.380 \leqslant \sigma \leqslant 5.022$

Réponse 4 Exemple pratique

1. $\mu = 10.4551, \sigma^2_{est} = 17.2159$.
2. Le quantile bilatéral de Student à 5% vaut 2.011 et l'IDC est alors $[2.1127; 18.7975]$.
3. Le quantile bilatéral de Student à 2% vaut 2.407 et l'IDC est alors [0.469620.4406].
4. On voit que l'IDC à 95% n'englobe pas toutes les valeurs de l'échantillon, mais que l'IDC à 98% les dépasse : un IDC à 95% est donc assez parlant pour décrire les valeurs communément mesurées, et ne nécessite que trois informations (moyenne, écart-type et effectif).
5. $max - min = 18.85 = 2t_{\alpha/2}\sigma_{est}$. En gardant $\sigma_{est} = 4.149$, on trouve $\alpha = 2.7638\%$.
6. En gardant $t_{\alpha/2} = 2.011$, la même formule donne $\sigma_{est} = 4.6877$.

Réponse 5 Estimation et IDC 1

1. $\mu_{est} = 8.9329, \sigma_{est} = 0.2986$
2. L'écart-type (numérateur) est petit et la taille de l'échantillon (dénominateur) est grande : la largeur de l'intervalle de confiance sera faible et l'encadrement sera donc précis.

Réponse 6 Estimation et IDC 2 Il faut considérer au moins $n = 8129$ individus.

4 Généralités sur les tests statistiques

— Ce sont des aides à la décision pour trancher systématiquement entre "oui" et "non" quand la réponse n'est pas évidente.
— La question porte sur la *population* ! On n'a pas besoin de faire de longues études pour savoir si 3.14 est plus grand que 3 ou pas.
— Comme tout résultat statistique, on n'obtiendra pas des vérités absolues, mais des éventualités assez probables en se basant sur les informations incomplètes disponibles

4.1 Préliminaires

— $\mathcal{H}_0$: hypothèse privilégiée — imposée par le test
— $\mathcal{H}_1$: hypothèse alternative — parfois orientable en fonction du résultat recherché

4.2 Estimation contextuelle

Sous l'hypothèse privilégiée $\mathcal{H}_0$, on aura alors une certaine V.A. qui suivra une certaine loi
— On peut ainsi calculer l'intervalle de confiance pour cette V.A., généralement avec un niveau de risque $\alpha = 5\%$: c'est la zone de validité pour $\mathcal{H}_0$
— En parallèle, on calcule une estimation de la V.A. à partir de l'échantillon disponible

4.3 Prise de décision

— Si l'estimation ne rentre pas dans l'intervalle de confiance, c'est donc que les observations ne sont pas compatibles avec la l'hypothèse privilégiée : on réfute $\mathcal{H}_0$ pour accepter $\mathcal{H}_1$
— Si l'estimation rentre dans l'intervalle de confiance, les observations sont donc compatibles avec $\mathcal{H}_0$: on accepte $\mathcal{H}_0$

4.4 Risques d'erreur

— On réfute $\mathcal{H}_0$ pour accepter $\mathcal{H}_1$ quand l'estimation est dans la zone de risque de l'IDC : α, appelé risque de première espèce, est donc précisément le risque d'erreur à réfuter $\mathcal{H}_0$ — On dit alors que le test est significatif au seuil α et idéalement on précise le risque le plus bas possible permettant de réfuter $\mathcal{H}_0$
— Quand l'estimation est dans l'IDC, on ne peut pas réfuter $\mathcal{H}_0$ et α n'est donc plus le risque d'erreur mais plutôt une sorte d'indicateur de qualité — On dit que le test est non significatif au seuil α et on précise idéalement le risque le plus haut possible ne réfutant pas $\mathcal{H}_0$

— Si on ne peut pas réfuter $\mathcal{H}_0$, le risque d'erreur, appelé risque de deuxième espèce, est noté β
— Il est généralement difficile à calculer mais c'est lui qui fait la différence d'efficacité (appelée puissance) entre deux tests différents pour la même question

Remarque Si on ne peut pas réfuter $\mathcal{H}_0$, cela ne veut pas dire que l'hypothèse privilégiée soit vraie. Cela veut juste dire qu'on n'a aucune preuve de contraire. Si on la pensait vraie pour d'autres raisons, alors on peut continuer d'y croire. Sinon on ne peut pas dire grand chose, d'où la non-significativité du test

5 Tests statistiques de base (5h + 2TP)

Ces tests sont valables quand la V.A. étudiée suit une loi normale, ou quand la taille de l'échantillon est supérieure à 30.

5.1 Comparaison d'un paramètre observé à un paramètre théorique

Un seul échantillon

5.1.1 Comparaison de variance

- $\mathcal{H}_0 : \Big(\sigma(X) = \sigma_0\Big)$
- $\mathcal{H}_1 : \Big(\sigma(X) \neq \sigma_0\Big)$ ou $\Big(\sigma(X) > \sigma_0\Big)$ ou $\Big(\sigma(X) < \sigma_0\Big)$
- Sous l'hypothèse $\mathcal{H}_0$, $K = \dfrac{(n-1)\sigma^2(X)}{\sigma_0^2} \sim \chi^2(n-1)$

5.1.2 Comparaison de moyenne

- $\mathcal{H}_0 : \Big(\mu(X) = \mu_0\Big)$
- $\mathcal{H}_1 : \Big(\mu(X) \neq \mu_0\Big)$ ou $\Big(\mu(X) > \mu_0\Big)$ ou $\Big(\mu(X) < \mu_0\Big)$
- Sous l'hypothèse $\mathcal{H}_0$:
  - Si $\sigma(X)$ est connu : $U = \dfrac{(\mu(X) - \mu_0)}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0;1)$
  - Si $\sigma(X)$ est estimé : $T = \dfrac{(\mu(X) - \mu_0)}{\sqrt{\frac{\sigma^2(X)}{n}}} \sim Student(n-1)$

5.2 Comparaison de deux paramètres observés - A

Deux échantillons *indépendants*

5.2.1 Comparaison de variances

- $\mathcal{H}_0 : \Big(\sigma(X_1) = \sigma(X_2)\Big)$

— $\mathcal{H}_1 : \Big(\sigma(X_1) \neq \sigma(X_2)\Big)$ ou $(\sigma(X_1) > \sigma(X_2))$
— Sous l'hypothèse $\mathcal{H}_0$, $F = \dfrac{\sigma^2(X_1)}{\sigma^2(X_2)} \sim Fisher(n_1 - 1\ ;\ n_2 - 1)$

5.2.2 Comparaison de moyennes

— $\mathcal{H}_0 : \Big(\mu(X_1) = \mu(X_2)\Big)$
— $\mathcal{H}_1 : \Big(\mu(X_1) \neq \mu(X_2)\Big)$ ou $\Big(\mu(X_1) > \mu(X_2)\Big)$ ou $\Big(\mu(X_1) < \mu(X_2)\Big)$
— Sous l'hypothèse $\mathcal{H}_0$:
  — Si $\sigma(X_1)$ et $\sigma(X_2)$ sont connus : $U = \dfrac{\Big(\mu(X_1) - \mu(X_2)\Big)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0;1)$
  — Si $\sigma(X_1)$ et $\sigma(X_2)$ sont estimés, il faut commencer par tester leur égalité.
    — Si on admet que $\sigma(X_1) = \sigma(X_2)$: $T = \dfrac{\Big(\mu(X_1) - \mu(X_2)\Big)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \sim \mathcal{S}tudent(n_1 + n_2 - 2)$, où
    $s^2 = \frac{(n_1-1)\sigma^2(X_1)+(n_2-1)\sigma^2(X_2)}{n_1+n_2-2}$ (la moyenne pondérée des variances)
    — Si on admet que $\sigma(X_1) \neq \sigma(X_2)$: $T = \dfrac{\Big(\mu(X_1) - \mu(X_2)\Big)}{\sqrt{\frac{\sigma^2(X_1)}{n_1} + \frac{\sigma^2(X_2)}{n_2}}} \sim \mathcal{S}tudent(k)$, où
    k = Partie entière de $\left(\dfrac{\left(\frac{\sigma_{1est}^2}{n_1} + \frac{\sigma_{2est}^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_{1est}^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\sigma_{2est}^2}{n_2}\right)^2}{n_2-1}} \right)$. Exemple : partie entière de $4.87 = 4$.
    Fonction Excel : ENT().
    Version simplifiée : k = minimum entre $n_1 - 1$ et $n_2 - 1$.

5.3 Comparaison de deux paramètres observés - B

Deux échantillons *appariés*

5.3.1 Comparaison de moyennes

On considère les valeurs de la V.A. $D = X_1 - X_2$: la différence entre les deux groupes.
— $\mathcal{H}_0 : (\mu(D) = 0)$
— $\mathcal{H}_1 : (\mu(D) \neq 0)$ ou $(\mu(D) > 0)$ ou $(\mu(D) < 0)$
— Sous l'hypothèse $\mathcal{H}_0$, $T = \dfrac{\mu(D)}{\sqrt{\frac{\sigma^2(D)}{n}}} \sim Student(n-1)$

Tests statistiques de base - Exercices

Exercice 1 La teneur en hémoglobine du sang des femmes non malades a pour valeur moyenne 14.5 g/100 mL et pour écart-type 1.1 g/100 mL qu'on supposera constant quelle que soit la population étudiée. Ce paramètre biologique suit une loi normale. Sur un échantillon de 20 femmes on trouve une teneur moyenne en hémoglobine de 13.8 g/100 mL. Au risque de 5%, peut-on conclure que la population de femmes dont est extrait cet échantillon présente une teneur en hémoglobine trop faible ?

Exercice 2 Dans un laboratoire pharmaceutique sont fabriquées des ampoules de soluté injectable dont le volume doit être de 1 mL. Afin de contrôler sa production, l'industriel effectue régulièrement des prélèvements de lots de 50 ampoules. Au cours d'un tel prélèvement, les résultats sont les suivants :

$$\sum_{i=1}^{50} x_i = 51.34 \text{ mL} \quad \text{et} \quad \sum_{i=1}^{50} x_i^2 = 53.64 \text{ mL}^2.$$

1. Calculez moyenne, variance et écart-type pour cet échantillon.
2. Déduisez-en une estimation ponctuelle de ces mêmes paramètres sur la population.
3. Au vu des ces résultats, peut-on conclure que le prélèvement est conforme à la valeur de référence ?

Exercice 3 Dans un laboratoire d'analyses médicales, on effectue le dosage de calcium sérique par une méthode colorimétrie dont l'écart-type sur la mesure est de 1.1 mg/L. Ce dosage suit une loi normale. Après une remise à niveau de l'appareil de mesure, le directeur de ce laboratoire veut vérifier que l'écart-type sur la mesure n'a pas augmenté. Il fait refaire par un même technicien, dans les mêmes conditions, 32 dosages d'un même prélèvement sérique et obtient les résultats suivants :

Teneur en Ca mg/L	[96.5 ;97.5[	[97.5 ;98.5[	[98.5 ;99.5[	[99.5 ;100.5[	[100.5 ;101.5[
Nombre	3	7	11	6	5

Exercice 4 Dans une usine sont fabriquées des micro-pipettes de laboratoire de volume calibré à 100μL. En cours de fabrication sur deux machines différentes, on prélève deux lots indépendants de 25 micro-pipettes et on obtient les valeurs suivantes pour les variances calculées sur les deux échantillons : $\sigma_{E1}^2 = 0.02\,\mu\text{L}^2$ et $\sigma_{E2}^2 = 0.015\,\mu\text{L}^2$. On suppose que la variable *volume d'une micro-pipette* suit une loi normale. Peut-on dire que les variances des volumes de micro-pipettes fabriquées par ces deux machines sont différentes (au risque de 5%) ?

Exercice 5 Dans une unité de fabrication, deux machines A et B produisent des ampoules pour soluté injectable. On sait que les variances du volume de ces ampoules dans toute la production de l'usine valent pour la machine A : 0.04 mL^2 et pour la machine B : 0.03 mL^2. Deux échantillons indépendants de 20 ampoules sont prélevés et on obtient les valeurs suivantes pour les volumes moyens : 4.9 mL pour l'échantillon issu de la production par la machine A et 5.3 mL pour l'échantillon issu de la production par la machine B.

1. Peut-on dire que la machine A produit des ampoules de volume moyen différent de celles produites par la machine B ?
2. On a donc conclu à une différence des volumes des machines à partir d'une différence des volumes des échantillons. A partir de quelle différence entre les volumes moyens des échantillons peut-on conclure, au risque de 5%, à une différence significative entre les volumes moyens de toutes les ampoules produites par les deux machines ?

Exercice 6 Le dosage des transaminases (exprimés en UI/L) informe sur le fonctionnement hépatique. Lors de la mise au point d'un traitement qui peut perturber la fonction hépatique, deux variantes du protocole thérapeutique sont appliquées a deux groupes indépendants de malades :

— la variante 1 à un groupe de 12 malades,
— la variante 2 à un groupe de 15 malades.

On suit la sensibilité au traitement par dosage des transaminases dont les valeurs sont reportées dans le tableau suivant :

Variante 1	22	18	25	14	16	22	17	19	30	17	23	17			
Variante 2	31	35	32	30	28	26	27	19	25	20	18	31	27	29	24

La variable *dosage des transaminases* est supposée suivre une loi normale. Peut-on dire, au risque de 5%, que les deux variantes du protocole modifient de façons différentes la fonction hépatique ?

Exercice 7 Dans un laboratoire pharmaceutique, deux machines assurent la fabrication des comprimés, l'une est rotative, l'autre est verticale. La production est contrôlée par pesée des comprimés et ce poids est supposé suivre une loi normale. On extrait de la production deux échantillons indépendants, de même taille $n = 25$, sur lesquels les comprimés sont pesés. On obtient les résultats suivants (en g et en g^2) :

— Pour la machine rotative, $\sum_{i=1}^{25} x_{i,1} = 2541$ et $\sum_{i=1}^{25} x_{i,1}^2 = 258\,505$,
— Pour la machine verticale, $\sum_{i=1}^{25} x_{i,2} = 2555$ et $\sum_{i=1}^{25} x_{i,2}^2 = 262\,547$.

Peut-on dire que la machine rotative produit des comprimés de poids inférieur à ceux produits par la machine verticale ?

Exercice 8 On étudie l'effet d'une nouvelle forme médicamenteuse sur la tension artérielle. On constitue un groupe de 12 personnes et on mesure leur tension artérielle avant et après ingestion de la substance. Les résultats sont les mesures de la tension systolique exprimées en mm de mercure.

Avant	120	130	132	125	140	145	135	125	133	140	138	137
Après	122	125	118	135	142	138	125	115	123	135	122	126

Peut-on dire que cette forme médicamenteuse a fait baisser la tension artérielle ?

Réponse 1 Comparaison à une moyenne théorique d'un petit échantillon suivant une loi normale d'écart type connu : avec $U_{est} = -2.85, U_{5\%} = -1.6449$ (test unilatéral), au risque de 5% on peut conclure que la population dont est extrait l'échantillon est constituée de femmes dont la teneur en hémoglobine est trop faible.

Réponse 2

1. $\mu_E = 1.027, \sigma_E^2 = 0.018482, \sigma_E = 0.13595$.
2. $\mu_{est} = 1.027, \sigma_{est}^2 = 0.018859, \sigma_{est} = 0.13733$.
3. Comparaison à une moyenne théorique avec un échantillon suffisamment grand pour être supposé suivre une loi normale d'écart type inconnu : avec $T_{est} = 1.39$ et $T_{2.5\%} = \pm 2.009$ (test bilatéral), au risque de 5% on peut conclure que le volume moyen semble conforme à la valeur de référence.

Réponse 3 Ecart type estimé $\sigma_{est} = 1.2011$mg/L. Avec $K_{est} = 36.96$ et $\chi^2_{5\%} = 44.9$ (test unilatéral), au risque de 5% on ne peut pas conclure que l'écart type après intervention soit supérieur à l'écart type initial. Remarque : Sur les tableaux, on n'aura que la ligne 30 DDL, au lieu de 31, et on prendra donc la limite $\chi^2_{5\%} = 43.7$.

Réponse 4 $\sigma_{est.1}^2 = 0.02083, \sigma_{est.2}^2 = 0.015625$. Avec $F_{est} = 1.33$ et $F_{2,5\%} = 2.27$, au risque de 5% (test bilatéral – avec les tableaux, comme on n'a pas (24 ;24), on prendra la valeur 2.3) on ne peut pas conclure que les variances associées à chacune des deux machines soient différentes.

Réponse 5 Comme on a un échantillon petit, il faut ajouter l'hypothèse de normalité pour la V.A. étudiée (le volume des ampoules) – les conditions énoncées ne permettant pas de la garantir.

1. Avec $U_{est} = -6.76$ et $U_{5\%} = -1.96$ (test bilatéral), au risque de 5% on peut dire que les ampoules A sont d'un volume moyen significativement inférieur aux B (conclusion valable même avec un risque d'erreur de 0.1%).
2. On conclut à une différence, au risque de 5%, entre les productions dès que $|U_{est}| > 1.96$, ce qui donne $|\mu_A - \mu_B| > 0.116$ mL.

Réponse 6 Comparaison de moyennes observées sur des échantillons indépendants de variances inconnues.

1. Comparaison des variances : avec des variances estimées $\sigma_{est.1}^2 = 20.54545, \sigma_{est.2}^2 = 24.45714$, une variable de décision $F_{est} = 1.19039$ et une valeur seuil $F_{2.5\%} = 3.36$, on peut conclure au risque de 5% (test bilatéral) que les variances sont égales pour les deux protocoles. On conserve alors comme variance commune $s^2 = 22.73$.
2. Comparaison des moyennes : avec les moyennes $\mu_{E1} = 20, \mu_{E2} = 26.8$, la variable de décision $T_{est} = -3.68$ et la valeur seuil $T_{2.5\%} = 2.06$ (test bilatéral), au risque de 5% on peut conclure que les moyennes sont significativement différentes selon le protocole utilisé. Cette conclusion est même valable avec un risque d'erreur de 1%.

Réponse 7

1. Comparaison des variances : avec des variances estimées $\sigma_{est.1}^2 = 9.91, \sigma_{est.2}^2 = 59.42$, une variable de décision $F_{est} = 5.998$ et une valeur seuil $F_{2.5\%} = 2.27$ (test bilatéral – avec les tableaux, comme on n'a pas (24 ;24), on prendra la valeur 2.3), on peut conclure au risque $\alpha = 5\%$ que les variances sont différentes. On ne peut donc pas calculer de variance commune.
2. Comparaison des moyennes : avec les moyennes $\mu_{E1} = 101.64, \mu_{E2} = 102.2$, la variable de décision $T_{est} = -0.336$ et la valeur seuil $T_{2.5\%} = \pm 2.04$ (test unilatéral avec 31 DDL – valeur calculée avec la formule adaptée au cas de variances non connues mais supposés différentes), au risque de $\alpha = 5\%$ on ne peut pas conclure que les poids moyens diffèrent.

Réponse 8 $\mu_{est}(D) = 6.16667, \sigma^2_{est}(D) = 56.69697, T_{est} = 2.837, T_{5\%} = 1.796$ (test unilatéral) : on peut conclure que la tension artérielle diminue après le traitement, au risque de 5% et même au risque de 1%.

6 Corrélation linéaire et régression linéaire (3h + 1TP)

— Une population, plusieurs variables aléatoires : un échantillon et plusieurs mesures par individu
— Soit on a une V.A. X contrôlée par l'expérimentateur (conditions de test), appelée indépendante/explicative, et une V.A. Y appelée dépendante/expliquée
— Soit on a deux V.A. différentes et on ne contrôle ni l'une ni l'autre
— Y-a-t-il une relation assez simple mais assez satisfaisante entre deux jeux de données ? Si oui, laquelle ?
— Il existe de nombreuses réponses, en fonction de ce qu'on considère comme simple, et surtout en fonction de ce qu'on considère satisfaisant.

6.1 Relation entre deux variables

Y'a-t-il une relation entre X et Y ?

6.1.1 Covariance

Variation entre les deux variables - indicateur de variance commune

— $\text{cov}(X,Y) = \dfrac{\sigma^2(X+Y) - \sigma^2(X) - \sigma^2(Y)}{2}$
$\Rightarrow \sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2\text{cov}(X,Y)$
— $\text{cov}(X,Y) = \mu(XY) - \mu(X)\mu(Y)$
$\Rightarrow \mu(XY) = \mu(X)\mu(Y) + \text{cov}(X,Y)$

Remarques

— De même que pour la variance, il faut utiliser 4 chiffres après la virgule pour les calculs de la covariance
— De même que pour la variance, une bonne estimation de $\text{cov}(X,Y)$ est $\text{cov}_{est} = \frac{n}{n-1}\text{cov}_E$
— Attention : X et Y indépendantes implique toujours $\text{cov}(X,Y) = 0$, mais la réciproque n'est vraie que quand X et Y suivent une loi normale

6.1.2 Coefficient de corrélation linéaire

Indicateur normalisé de la pertinence du modèle affine (souvent improprement appelé linéaire)

$$Y = aX + b$$

— $R(X,Y) := \dfrac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$
— $-1 \leqslant R(X,Y) \leqslant +1$
— Plus $|R|$ est proche de 1, plus le modèle *linéaire* est pertinent et le signe de R donne le signe de la pente
— Plus $|R|$ est proche de 1, plus X et Y sont dépendantes

6.1.3 Coefficient de détermination

— R^2 donne le pourcentage d'explication de Y par $(aX + b)$

6.1.4 Prise de décision par rapport au coefficient de corrélation linéaire

Comme pour tous les tests statistiques de ce résumé de cours, il faut que X et Y suivent une loi normale
— $\mathcal{H}_0 : R^2(X,Y) = 0$ (pas de relation linéaire entre X et Y)
— $\mathcal{H}_1 : R^2(X,Y) > 0$ (relation linéaire)
— Sous l'hypothèse $\mathcal{H}_0$, $F = \dfrac{R^2}{1-R^2}(n-2) \sim Fisher(1, n-2)$

Remarque Dans certains livres, on verra la statistique de test $T = \sqrt{F} \sim Student(n-2)$. C'est parfaitement équivalent, au détail près que le test avec *Student* est bilatéral ! (on y teste $R = 0$ et $R \neq 0$)

6.2 Droite de régression linéaire

Qu'on ait une relation linéaire entre X et Y ou pas, tant que les mesures ne sont pas toutes faites pour la même valeur de x, il est possible d'estimer les paramètres du modèle linéaire $Y = aX + b$.
— Soit on a montré que ce modèle est pertinent, et on pourra alors prédire (une estimation) des valeurs de Y
— Soit on a montré que ce modèle n'est pas pertinent, et on appellera alors cette droite la droite de tendance

6.2.1 Droite des moindres carrés

— $a = \dfrac{\text{cov}(X,Y)}{\sigma^2(X)} = R(X,Y)\dfrac{\sigma(Y)}{\sigma(X)}$
— $b = \mu(Y) - a\mu(X)$

Remarques
— La droite $Y = aX + b$ passe par le point $\big(\mu(X); \mu(Y)\big)$
— Comme a est proportionnel à R, les hypothèses du test de la section 6.1.4 peuvent aussi se lire $a = 0$ ou $a \neq 0$: on a bien une équation de droite, ou pas.

6.3 Modèles de régression non linéaire

Si le modèle linéaire n'est pas pertinent ($|R|$ proche de 0) on peut quand même utiliser les outils précédents dans certains cas. Par exemple :

- Si $\tilde{b} > 0$, le modèle $Y = \tilde{b}e^{aX}$ est équivalent à $\ln(Y) = aX + b$ (et $\tilde{b} = e^b$) : on analyse la pertinence de ce modèle en faisant les calculs avec $\ln(Y)$ et X
- Si $\tilde{b} > 0$, le modèle $Y = \tilde{b}X^a$ est équivalent à $\ln(Y) = a\ln(X) + b$ (et $\tilde{b} = e^b$) : on analyse la pertinence de ce modèle en faisant les calculs avec $\ln(Y)$ et $\ln(X)$

Statistiques bivariées et régression - Exercices

Exercice 1 On a injecté à 15 souris, prises au hasard, des doses (en mg/L) d'une drogue et observé les survies (en jours). On admet que le lien entre dosage et survie suit un modèle linéaire.

Dose	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Survie	8	8	8	8	10	10	10	10	10	12	10	12	12	12	12

1. Calculer et interpréter coefficient de détermination entre dosage et survie (penser à regrouper les valeurs pour aller plus vite à la calculatrice).
2. Pour une dose de 3.6mg/L, combien de jours de survie peut-on prévoir ?

Exercice 2 Tel père, tel fils Dans l'optique d'étudier la liaison entre le poids d'un père et de son fils aîné, on a relevé le poids de 12 pères et de leur fils aîné respectif. Les résultats (en kg) sont dans le tableau suivant :

Père	67	63	70	63	67	63	70	67	67	67	70	70
Fils	68	66	68	66	68	66	68	66	71	66	68	71

(Pensez à regrouper les valeurs pour aller plus vite avec la calculatrice)

1. Peut-on dire que les poids des pères et fils sont liés linéairement ? (on supposera que les hypothèses de normalité sont vérifiées)
2. Un père pèse 61kg, quel poids peut-on prévoir pour son fils ?

Exercice 3 Radioacivité Un élément de la classification périodique est bombardé dans un accélérateur de particules. Il donne naissance à de nouveaux éléments dont l'un est radioactif. On étudie sa décroissance en effectuant des comptages à différents moments pendant une durée de 31 heures. Les résultats obtenus sont :

t	1	19	31
N	1240	1160	1065

Sachant que la décroissance radioactive obéit à une relation du type

$$N = N_0 e^{-\lambda t},$$

avec $\lambda = \frac{\ln 2}{T}$, où T est la demi-vie du corps, le but est ici de donner la radioactivité du corps au temps initial.

1. Vérifier *à la louche* la validité du modèle proposé avec le coefficient de corrélation
2. Calculer les paramètres N_0 et λ par régression des données
3. Enfin extrapoler la valeur en $t = 0$ grâce à ce modèle

Exercice 4 Deux méthodes sont proposées pour doser un antibiotique dans un liquide biologique : une méthode radio-immunologique R et une méthode immuno-enzymatique I. On veut déterminer si ces deux méthodes fournissent des résultats comparables. On fait donc un essai sur 14 prélèvements, chacun d'eux étant séparé en deux aliquote sur lesquels l'antibiotique est dosé par l'une ou l'autre méthode. Les résultats obtenus (en mg) sont :

R	1.42	3.21	3.21	1.42	3.21	2.19	2.19	1.42	2.19	1.42	1.42	1.42	3.21	1.42
I	3.36	3.36	2.22	1.18	3.36	2.22	2.22	1.18	2.22	1.18	2.22	1.18	3.36	1.18

Peut-on dire que les résultats obtenus par les deux méthodes sont liés linéairement ? (pensez à regrouper les valeurs pour aller plus vite avec la calculatrice)

Exercice 5 On aimerait savoir s'il existe une corrélation linéaire entre le temps de mise en solution d'un certain type de streptomycine en poudre et la densité de la solution de streptomycine avant séchage. Pour cela on a obtenu les résultats suivants :

Densité	1140	1092	1127	1175	1162	1105	1160	1143	1170	1105	1150	1145	1120
Temps	95	35	15	110	105	20	70	90	100	45	45	55	45

Réponse 1 Variable contrôlée : D = doses de drogue, variable aléatoire : S = survie.

1. $R_{est} = 0.916$. Le coefficient de détermination vaut donc $R^2_{est} = 0.839$. On peut alors dire que le modèle linéaire est un très bon modèle : 83.9% de la variation totale est expliquée par la droite.
2. La droite de régression est estimée par $S = D + 7.13$. Pour une dose de 3.6mg/L, l'estimation ponctuelle est ainsi de 10.73 jours.

Réponse 2 Tel père, tel fils

1. X = poids du père, Y = poids du fils. $R^2_{est} = 35.36\%$.
   — Hypothèse privilégiée : $R^2 = 0$
   — Hypothèse alternative : $R^2 > 0$
   — Variable de décision $F_{est} = \frac{0.3536}{1-0.3536} 10 = 5.47$
   — Intervalle de validité : $[4.9646, +\infty]$
   — On rejette donc l'hypothèse privilégiée au risque $\alpha = 5\%$
   Conclusion, au risque $\alpha = 5\%$, on peut dire que le poids du père est corrélé linéairement avec celui de son fils aîné, malgré un R^2 très faible. C'est le nombre de mesures qui aide.
2. Le modèle linéaire $Y = aX + b$ semble donc pertinent. On estime les paramètres de cette droite par $a_{est} = 0.393$ et $b_{est} = 41.345$. Avec ce modèle et $x = 61$, on trouve $y = 65.3kg$.

Réponse 3 Radioactivité Variable contrôlée : temps t, variable aléatoire : nombre de particules N, nombre de mesures : n=3. Le modèle proposé est de la forme $\ln N = \ln N_0 - \lambda t$. On fait alors une régression linéaire pour estimer les paramètres $\ln N_0$ et λ : $\lambda_{est} = 4.96 \cdot 10^{-3}$ et $\ln N_{0\ est} = 7.134$. On trouve un très bon $R^2_{est} = 0.965$ (mais il faudra utiliser la valeur ajustée F pour tirer des conclusions précises). Avec $t = 0$, on trouve alors $N_{0\ est} = 1254.26$.

Réponse 4 $R_{est} = 49.60\%$
- Hypothèse privilégiée : $R = 0$
- Hypothèse alternative : $R \neq 0$
- Variable de décision $F_{est} = \frac{0.4960}{1-0.4960}12 = 11.81$
- Intervalle de validité à 5% : $[4.747; +\infty]$
- On rejette donc l'hypothèse privilégie au risque $\alpha = 1\%$

Conclusion, au risque $\alpha = 5\%$, on peut dire que les résultats donnés par les deux méthodes sont corrélés linéairement.

Réponse 5 $R^2_{est} = 60.45\%$.
- Hypothèse privilégiée : $R^2 = 0$
- Hypothèse alternative : $R > 0$
- Variable de décision $F_{est} = \frac{0.6045}{1-0.6045}11 = 16.81$
- Intervalle de validité à 5% $[4.8443; +\infty]$

Conclusion, au risque $\alpha = 5\%$, on peut dire que le le temps de mise en solution est corrélé linéairement à la densité de la solution de streptomycine avant séchage.