

Probabilités : données, lois et estimation

OSSIK • COURS 13-01 • LE CORRIGÉ • RÉVISION 0.5

PASS - majeure santé - classement d'après la maquette publique de Paris Cité, sans lien avec l'université. Cours original rédigé avec une IA. Le cours de ta faculté fait foi.

Sommaire du cours

Partie 1 • Description des données et variabilité	1
Partie 2 • Dénombrement, lois discrètes et calcul probabiliste	6
Partie 3 • Conditionnement, arbres et théorème de Bayes	12
Partie 4 • Variables aléatoires et lois usuelles	18
Partie 5 • Échantillonnage et estimation par intervalle	23

Partie 1 • Description des données et variabilité

Prérequis et objectifs

Savoir additionner, calculer une proportion et mettre une quantité au carré. Identifier l'unité statistique avant de compter les observations.

À l'issue de ce cours, vous devez pouvoir :

- Choisir un résumé adapté à la variable.
- Calculer moyenne, variance et dispersion.
- Repérer les effets d'un codage ou de données manquantes.

Fondements du thème

Population, échantillon et variable

Une population est l'ensemble sur lequel porte la question scientifique ; un échantillon est le sous-ensemble effectivement observé. Le paramètre décrit la population, tandis qu'une statistique est calculée dans l'échantillon. Une moyenne calculée sur des étudiants volontaires peut être parfaitement exacte pour ces volontaires sans estimer correctement celle de tous les étudiants. La représentativité dépend du recrutement, pas seulement de la taille.

Une variable qualitative nominale classe sans ordre naturel, comme un groupe sanguin. Une variable ordinale possède un ordre, comme une catégorie de sévérité, mais l'écart entre catégories n'est pas nécessairement constant. Une variable quantitative discrète compte des valeurs séparées ; une variable continue représente une mesure pouvant théoriquement prendre toute valeur dans un intervalle. Coder des catégories par 1, 2 et 3 ne crée pas automatiquement une échelle quantitative.

Position et dispersion

La moyenne $\bar{x} = \text{somme}(x_i)/n$ utilise toutes les valeurs et est sensible aux extrêmes. La médiane partage la série ordonnée en deux moitiés et résiste mieux à quelques valeurs très éloignées. Les quartiles décrivent les positions à 25 %, 50 % et 75 %, selon une convention de calcul précisée pour les petits échantillons. L'étendue interquartile est $Q3 - Q1$; elle résume la dispersion centrale.

La variance d'échantillon usuelle $s^2 = \text{somme}[(x_i - \bar{x})^2]/(n - 1)$ estime sans biais la variance d'une population sous échantillonnage indépendant identiquement distribué. L'écart-type s retrouve l'unité des données. La variance descriptive d'une série considérée comme exhaustive peut utiliser n au dénominateur : les objectifs diffèrent. L'erreur standard $s/\sqrt{n}$ décrit la variabilité de la moyenne estimée, et non la dispersion des individus.

Graphiques et transformations

Un histogramme représente une variable quantitative découpée en classes ; si leurs largeurs diffèrent, la surface des rectangles doit être proportionnelle à la fréquence. Un diagramme en barres convient à des catégories. Un nuage de points montre la relation entre deux variables quantitatives. Une échelle tronquée ou une sélection de classes peut exagérer une différence sans changer les nombres.

La moyenne géométrique peut être utile pour des grandeurs strictement positives variant de façon multiplicative : elle vaut l'exponentielle de la moyenne des logarithmes. Une transformation logarithmique modifie l'interprétation des différences en rapports. Elle ne doit pas être appliquée automatiquement à des zéros ou à des valeurs négatives. Les données manquantes doivent être décrites ; remplacer toutes les valeurs absentes par la moyenne réduit artificiellement la dispersion.

Développements et raisonnements

L'unité statistique précède le tableau

Un laboratoire peut mesurer dix cellules chez chacune de vingt personnes. Son fichier contient deux cents lignes, mais les cellules d'une personne partagent un contexte biologique. Il faut distinguer le nombre de mesures et le nombre d'unités recrutées de façon indépendante. Résumer toutes les cellules peut répondre à une question descriptive ; faire comme si elles provenaient de deux cents personnes change l'inférence.

La population cible doit aussi être distinguée de la population accessible. Des étudiants présents lors d'un cours ne représentent pas nécessairement tous les inscrits. L'absence peut être liée à la santé, au travail ou à d'autres facteurs étudiés. Une taille importante ne remplace donc pas une description du mécanisme de recrutement.

Décomposer la dispersion

La somme des écarts à la moyenne est nulle par construction. Pour mesurer une dispersion, on utilise donc leurs carrés ou leurs valeurs absolues. La variance met davantage de poids sur les écarts élevés. L'écart-type revient à l'unité de départ, mais ne doit pas être assimilé à une distance maximale : certaines observations peuvent être à plusieurs écarts-types de la moyenne.

Pour une transformation $y = ax + b$, la moyenne devient $a \text{ moyenne}(x) + b$, tandis que l'écart-type devient valeur absolue de a fois s_x . Un changement d'origine ne modifie pas la dispersion ; un changement d'échelle la modifie. Passer de degrés Celsius à Fahrenheit modifie donc à la fois

moyenne et écart-type, mais le terme ajouté à la température n'entre pas dans ce dernier.

Moyennes pondérées et données groupées

Si deux groupes ont des effectifs différents, leur moyenne commune est $(n_1m_1 + n_2m_2)/(n_1 + n_2)$. La moyenne simple de leurs moyennes ne convient que si les effectifs sont égaux. La variance globale dépend de la dispersion à l'intérieur des groupes et de l'écart entre leurs moyennes. On ne peut pas la retrouver en faisant simplement la moyenne des écarts-types.

Variable ou forme	Résumé souvent utile	Ce qu'il faut encore montrer
Qualitative nominale	Effectifs et proportions	Dénominateur et manquants
Ordinale	Distribution ordonnée, médiane selon cas	Sens des catégories
Quantitative symétrique	Moyenne et écart-type	Valeurs extrêmes
Quantitative asymétrique	Médiane et quantiles	Forme de la distribution

La moyenne géométrique décrit un niveau multiplicatif pour des valeurs positives. Elle n'est pas une solution automatique aux données asymétriques : il faut que le logarithme corresponde à une interprétation scientifique utile. Une moyenne calculée après exclusion des zéros répond à une population différente si ces zéros sont réels.

Atelier de calcul et analyse de sensibilité

Considérons 4, 5, 5, 6 et 10 unités. La moyenne vaut 6 et la médiane 5. Les carrés des écarts à 6 sont 4, 1, 1, 0 et 16, soit 22. La variance descriptive exhaustive est $22/5 = 4,4$; la variance d'échantillon corrigée est $22/4 = 5,5$. Les écarts-types correspondants valent 2,098 et 2,345. Les deux calculs sont arithmétiquement justes mais répondent à des conventions différentes.

Un groupe de dix personnes possède une moyenne de 12 unités, un autre de trente personnes une moyenne de 20. La moyenne commune est $(10 \times 12 + 30 \times 20)/40 = 18$. La moyenne non pondérée des moyennes, 16, donnerait trop de poids au petit groupe. Sans les dispersions internes, on ne peut pas déterminer l'écart-type global à partir de ces seuls nombres.

Pour les valeurs positives 1, 10 et 100, la moyenne arithmétique vaut 37, la médiane 10 et la moyenne géométrique $(1 \times 10 \times 100)^{1/3} = 10$. La dernière reflète l'espacement régulier sur une échelle logarithmique. Si la valeur 100 est une erreur de saisie, aucun choix de moyenne ne dispense de revenir à la donnée d'origine ; si elle est réelle, la supprimer pour rapprocher les résumés fausserait la description.

Approfondissement et contrôle du modèle

Effet d'une transformation sur les résumés

Une transformation monotone croissante conserve l'ordre des observations et transforme donc la médiane selon la convention retenue dans de nombreux cas usuels. Elle ne transforme pas la moyenne par la même simple règle : la moyenne de $\ln X$ n'est pas $\ln$ de la moyenne de X . La courbure de la fonction explique cette différence.

Une transformation logarithmique peut rendre les variations multiplicatives plus faciles à décrire. Après retour à l'échelle initiale, une différence de moyennes logarithmiques devient un rapport de moyennes géométriques, pas une différence de concentrations. Les unités et le sens de l'effet doivent être réexprimés dans la conclusion. Une présentation claire peut donner simultanément le modèle transformé et un effet interprétable sur l'échelle originale, en précisant ce qui a été estimé.

Mise à l'épreuve : Même moyenne et même dispersion, données différentes

Problème. Comparer $A = (-1 ; -1 ; 1 ; 1)$ et $B = (-\sqrt{2} ; 0 ; 0 ; \sqrt{2})$. Calculer la moyenne et la variance avec le diviseur n , puis décrire ce que ces résumés ne montrent pas.

Résolution. Dans les deux cas, la somme est nulle et la somme des carrés vaut 4. La moyenne est donc 0 et la variance descriptive 1. Avec le diviseur $n - 1$, les deux variances estimées seraient également identiques : $4/3$. Pourtant, A ne contient aucune valeur centrale, tandis que B contient deux zéros et des valeurs extrêmes plus éloignées. Un graphique des valeurs individuelles révèle cette différence. L'égalité de deux premiers moments ne prouve ni l'identité des distributions ni leur normalité. Avec seulement quatre observations, on évite aussi de conclure sur une forme populationnelle stable.

Schéma de raisonnement

Décrire des données

1. Unité statistique • 2. Type de variable • 3. Distribution • 4. Résumés adaptés • 5. Limites de généralisation

Un résumé condense l'information. Le schéma ne transforme pas un échantillon de convenance en échantillon représentatif.

Exercices progressifs

1. Pour 2, 3, 3, 4 et 18, calcule moyenne et médiane.
2. Un écart-type vaut 12 avec $n = 36$. Quelle erreur standard de la moyenne ?
3. Un code de douleur « faible = 1, modérée = 2, forte = 3 » autorise-t-il à dire que forte vaut trois fois faible ?
4. Une série a moyenne 8 et écart-type 2. Pour $Y = 3X + 4$, donner moyenne et écart-type.
5. Deux groupes de 20 et 80 personnes ont des moyennes de 10 et 15. Moyenne globale ?

6. Pourquoi remplacer dix valeurs manquantes par la moyenne des valeurs observées réduit-il artificiellement la dispersion ?

Corrections argumentées

1. Somme = 30, moyenne = 6, médiane = 3. La valeur 18 attire fortement la moyenne ; la médiane reste une observation centrale.
2. $SE = 12/\sqrt{36} = 2$, dans l'unité de la variable. Les individus restent dispersés avec un écart-type de 12.
3. Non. Les nombres codent un ordre sans garantir une proportionnalité ni des écarts égaux.
4. Moyenne 28, écart-type 6. Le décalage de 4 modifie le centre mais pas les écarts autour de ce centre.
5. $(20 \times 10 + 80 \times 15)/100 = 14$. Les effectifs sont les poids de la moyenne.
6. Les valeurs imputées ont un écart nul à cette moyenne et ne reproduisent pas la variabilité inconnue. Leur ajout augmente le dénominateur sans apporter la dispersion qui aurait accompagné de vraies observations.

Références et portée

- Penn State : cours de statistique, référence 1

Sources consultées le 1er octobre 2026 pour les notions signalées. Développements, situations et calculs pédagogiques OSSIK ; données fictives. La concordance complète avec les supports universitaires de l'année et la relecture scientifique humaine restent à réaliser. Les modèles biomédicaux ne fournissent aucune décision thérapeutique individuelle.

Partie 2 • Dénombrement, lois discrètes et calcul probabiliste

1. Définir l'expérience avant de compter

Une probabilité ne se calcule pas en comptant des mots dans un énoncé. Il faut définir les issues possibles, le mécanisme de sélection et l'événement recherché. Le quotient « cas favorables/cas possibles » n'est valide que lorsque les issues comptées sont équiprobables. Regrouper des issues en catégories ne rend pas les catégories équiprobables.

Si deux tirages indépendants donnent chacun A ou B avec probabilité $1/2$, les quatre séquences AA, AB, BA et BB sont équiprobables. Les catégories « deux A », « un A » et « aucun A » ne le sont pas : leurs probabilités sont $1/4$, $1/2$ et $1/4$. Cette différence prépare le coefficient binomial. Une liste courte et exhaustive est souvent un meilleur départ qu'une formule mémorisée sans comprendre ses objets.

Les probabilités appartiennent à $[0 ; 1]$. Une somme sur une partition complète doit valoir 1. Une probabilité conditionnelle nécessite un événement conditionnant de probabilité strictement positive. Ces contrôles sont nécessaires mais pas suffisants : une mauvaise formule peut fournir un nombre parfaitement compris entre zéro et un.

2. Ordre, répétition et remise

Pour choisir successivement k objets distincts parmi n sans remise, en tenant compte de l'ordre, on obtient $n(n-1)\dots(n-k+1)$ séquences. Avec remise et n choix possibles à chaque étape, on obtient n^k séquences. Sans tenir compte de l'ordre ni autoriser de répétition, le nombre de sous-ensembles de taille k est $C(n,k) = n!/[k!(n-k)!]$.

La division par $k!$ retire les $k!$ façons d'ordonner le même groupe. Choisir deux personnes parmi cinq donne dix paires ; désigner parmi elles un premier et un second rôle donne vingt possibilités. Le contexte détermine si l'ordre compte. Une numérotation technique des observations ne transforme pas nécessairement le résultat scientifique en séquence ordonnée.

Une sélection sans remise crée une dépendance. Dans une boîte de cinq tubes dont deux sont marqués, la probabilité de tirer un marqué au premier tirage vaut $2/5$. Après un marqué, elle devient $1/4$; après un non marqué, elle devient $2/4$. Le mécanisme modifie donc la probabilité conditionnelle au fil des tirages.

3. Reconstruire la loi binomiale

Considérons n essais indépendants, avec une probabilité de succès p identique à chaque essai. Pour une séquence précise contenant k succès, la probabilité vaut $p^k(1-p)^{(n-k)}$. Comme $C(n,k)$ séquences conduisent au même compte k , $P(X=k) = C(n,k)p^k(1-p)^{(n-k)}$.

Le mot « succès » désigne seulement l'événement compté ; il peut s'agir d'une défaillance ou d'un événement indésirable dans un exercice. Une série de résultats corrélés ne devient pas binomiale parce que les observations sont binaires. Des probabilités individuelles différentes rendent également la loi binomiale usuelle inadaptée, même si les essais restent indépendants.

Pour $n = 8$ et $p = 0,25$, $E(X) = np = 2$ et $\text{Var}(X) = np(1-p) = 1,5$. L'espérance n'est pas une prédiction certaine de deux succès. La loi répartit la probabilité entre neuf valeurs possibles de zéro à huit. $P(X=0)$

$= 0,75^8 = 0,1001$ environ ; $P(X \geq 1) = 0,8999$. Le complément simplifie le calcul « au moins un » sans additionner huit termes.

4. Espérance, indicateurs et dépendance

Écrire $X = I_1 + \dots + I_n$, où I_j vaut un en cas de succès et zéro sinon, permet de retrouver $E(X)$ = somme des p_j . Cette additivité des espérances ne demande pas l'indépendance. En revanche, la variance de la somme contient les covariances : $\text{Var}(X)$ = somme des variances + 2 fois la somme des covariances par paires.

Deux indicateurs ayant $p = 0,5$ et toujours égaux donnent une somme égale à zéro ou deux avec probabilités égales. Son espérance reste un, mais sa variance vaut un. S'ils étaient indépendants, cette variance serait 0,5. La même moyenne peut donc masquer des fluctuations très différentes selon la dépendance.

Pour une transformation affine $Y = aX + b$, $E(Y) = aE(X) + b$ et $\text{Var}(Y) = a^2 \text{Var}(X)$. Ajouter une constante décale la distribution sans modifier sa dispersion. Une transformation non linéaire ne permet généralement pas de remplacer $E[f(X)]$ par $f[E(X)]$. Cette distinction rejoint les biais d'approximation du cours 22-01, partie 3.

5. Tirage sans remise : loi hypergéométrique

Une population finie contient N éléments dont K marqués. On prélève n éléments sans remise, tous les sous-ensembles de taille n étant équiprobables. Pour avoir exactement k marqués, choisir k parmi K puis $n-k$ parmi $N-K$ donne $C(K, k)C(N-K, n-k)$ possibilités favorables. Le dénominateur vaut $C(N, n)$.

Les valeurs admissibles vérifient $\max(0, n-N+K) \leq k \leq \min(n, K)$. Cette borne évite de calculer une combinaison impossible. L'espérance vaut nK/N . La variance est $np(1-p)(N-n)/(N-1)$, avec $p=K/N$ et $N>1$. Le facteur de correction de population finie réduit la variance par rapport au modèle avec remise.

Si tous les éléments sont tirés, $n=N$ et la variance devient nulle : le compte final est nécessairement K . Une approximation binomiale peut convenir quand le prélèvement représente une faible fraction de la population, mais le critère dépend de la précision demandée et de la probabilité étudiée.

6. Poisson et fenêtre d'observation

Une loi de Poisson de paramètre λ décrit un compte entier avec $P(X=k) = e^{-\lambda} \lambda^k / k!$. Son espérance et sa variance sont λ . Dans un processus homogène idéal de taux r , le paramètre d'une fenêtre de durée t est $\lambda = rt$. Le taux porte une unité inverse du temps, tandis que λ est un nombre moyen d'événements et donc sans dimension.

Si $r=0,4$ événement par minute, une fenêtre de cinq minutes a $\lambda=2$. La probabilité de zéro événement vaut e^{-2} , et non $e^{-0,4}$. Aggrandir la fenêtre change la loi du compte. Le modèle suppose des propriétés de stationnarité et d'indépendance appropriées ; des regroupements, une évolution du taux ou une période réfractaire peuvent le contredire.

Lorsque n est grand et p faible, une binomiale peut être approchée par une Poisson de $\lambda=np$. C'est une approximation, pas une identité. Pour une probabilité de queue importante dans une décision, vérifier l'erreur plutôt que d'invoquer une règle numérique universelle.



7. Temps d'attente et propriété sans mémoire

Dans un processus de Poisson homogène, le temps T avant le premier événement satisfait $P(T > t) = P(\text{aucun événement jusqu'à } t) = e^{-(rt)}$. Il suit une loi exponentielle de moyenne $1/r$. Le compte sur une fenêtre et le temps d'attente sont donc deux variables différentes reliées par le même modèle de processus.

Le rapport $P(T > s+t)/P(T > s)$ vaut $e^{-(rt)}$: conditionnellement à l'absence d'événement pendant s , le temps supplémentaire suit la même loi. Cette propriété sans mémoire est une hypothèse restrictive. L'usure d'un composant, l'âge biologique ou l'évolution d'un état clinique peuvent rendre un taux constant inadéquat ; aucune loi exponentielle ne s'impose par le seul mot « survie ».

Dans une suite de Bernoulli indépendantes, le rang G du premier succès a $P(G=k) = (1-p)^{(k-1)}p$ pour $k \geq 1$. Sa moyenne vaut $1/p$. Certains ouvrages comptent le nombre d'échecs avant le succès et commencent à zéro ; il faut vérifier la convention avant de comparer les formules.

8. Approximation normale et correction de continuité

Une binomiale suffisamment peu asymétrique peut être approchée par une normale de moyenne np et variance $np(1-p)$. La normale est continue alors que X est entier. Pour approcher $P(X \leq k)$, on peut utiliser la borne $k+0,5$; pour $P(X \geq k)$, la borne correspondante est $k-0,5$. Cela améliore souvent l'approximation, mais ne garantit pas sa qualité dans les queues ou près des bornes zéro et n .

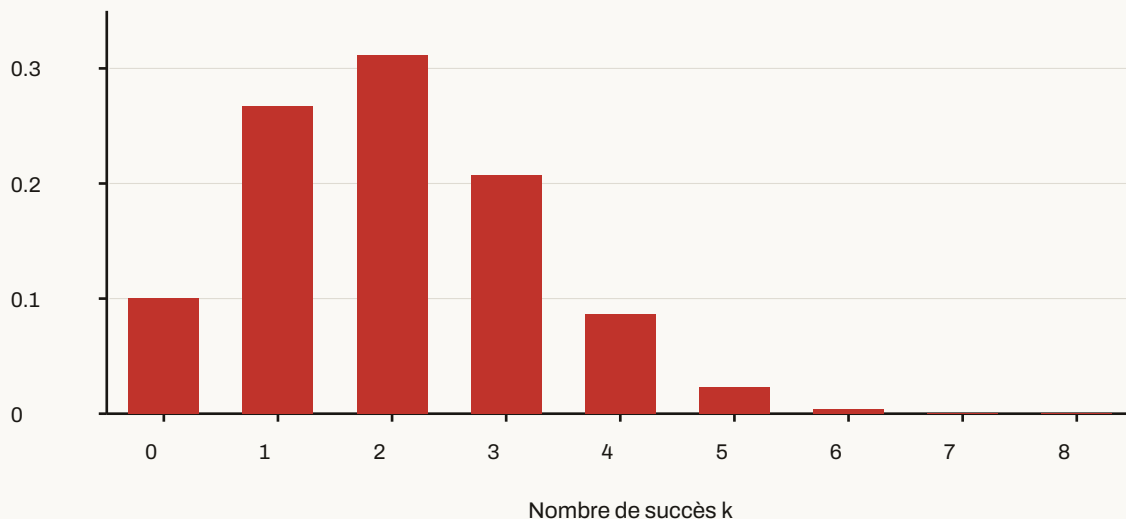
Standardiser consiste à calculer $z = (\text{borne} - \text{moyenne}) / \text{écart-type}$. La variance ne doit pas remplacer l'écart-type au dénominateur. Une probabilité de variable continue en un point exact est nulle ; une densité n'est pas une probabilité ponctuelle. Le cours 13-01, partie 4 donne les bases des lois ; ce chapitre explicite le choix de modèle et les calculs associés.



Lecture visuelle : Loi binomiale : lire des masses de probabilité

Loi binomiale : huit épreuves, probabilité 0,25

Épreuves indépendantes, même p ; $E(X) = 2$; variance = 1,5.
Probabilité $P(X = k)$



Les hauteurs des neuf barres somment à 1 : il ne s'agit pas d'une densité.

Huit épreuves de Bernoulli indépendantes, même probabilité de succès $p = 0,25$. Chaque barre donne $P(X = k) = C(8, k) p^k (1-p)^{8-k}$. Les neuf hauteurs somment à 1. L'espérance vaut 2 et la variance 1,5 ; ces paramètres ne rendent pas la distribution continue.

Question de lecture

Quelle est la probabilité d'obtenir exactement deux succès ?

Correction de lecture

$C(8, 2) \times 0,25^2 \times 0,75^6 = 0,31146$ environ. Cette probabilité se lit comme une hauteur de barre, et non comme une aire sous une densité continue.

Choisir entre les modèles : protocole avant calculatrice

Trois laboratoires observent des résultats binaires. Le premier répète des essais indépendants avec une probabilité stable. Le deuxième tire sans remise dans une collection finie. Le troisième réalise plusieurs lectures du même échantillon, avec un effet commun de préparation. Les trois tableaux peuvent contenir le même nombre de lignes et la même proportion observée, mais ils n'ont pas la même distribution du nombre de résultats positifs.

La loi binomiale convient au premier modèle si les hypothèses sont justifiées. La loi hypergéométrique décrit le deuxième lorsque la composition de la collection est fixée et le tirage équiprobable. Le troisième nécessite de représenter la dépendance ; compter chaque lecture comme une nouvelle unité indépendante sous-estime souvent l'incertitude. Le nombre de lignes n'est pas le nombre d'informations indépendantes.

Pour choisir, reconstruire une réalisation possible de l'expérience. Après un résultat, la probabilité du suivant change-t-elle ? Tous les éléments ont-ils la même chance d'être sélectionnés ? Le paramètre est-il stable pendant la fenêtre ? Des éléments partagent-ils une préparation, un sujet ou un lieu ? Les réponses déterminent la structure avant la formule.

Question supplémentaire. Une série de comptes a une variance empirique supérieure à sa moyenne. Faut-il conclure immédiatement que la loi de Poisson est impossible ? **Réponse.** Pas sur un seul petit échantillon : moyenne et variance empiriques fluctuent. Cependant, une surdispersion persistante peut signaler une hétérogénéité de taux, des dépendances ou des regroupements. Il faut examiner ces mécanismes et la précision de l'écart, plutôt que modifier arbitrairement le paramètre λ .

Un mélange de populations de taux différents peut présenter plus de dispersion qu'une Poisson de taux moyen, car la variabilité entre taux s'ajoute à la variabilité conditionnelle des comptes. Ce raisonnement explique pourquoi une moyenne correctement estimée ne suffit pas à prévoir les queues de distribution. Une vérification de normalisation est utile pour un calcul, mais seule l'analyse du protocole rend le modèle pertinent.

9. Exercices progressifs

1. Parmi six personnes, combien de groupes de trois sans rôle distinct ? Combien de triples ordonnés ?
2. Pour X binomiale de $n=8$, $p=0,25$, calculer $P(X=2)$ et $P(X \geq 1)$.
3. Dans dix tubes dont trois marqués, tirer deux sans remise. Calculer la probabilité que les deux soient marqués, de deux façons.
4. Pour ce tirage, calculer espérance et variance du nombre de marqués, puis comparer à un tirage avec remise.
5. Un compte de Poisson a taux $0,4/\text{min}$. Quelle est la probabilité d'exactly trois événements en cinq minutes ?
6. Dans le même processus, calculer la probabilité d'attendre plus de cinq minutes et le temps moyen d'attente.
7. Deux indicateurs identiques valent un avec probabilité $0,5$. Calculer la variance de leur somme et expliquer pourquoi la formule binomiale échoue.
8. Un étudiant approxime $P(X \geq 15)$ par $P(Y \geq 15,5)$ pour Y normale. Repérer l'erreur de correction de continuité.

10. Corrections en étapes

1. $C(6,3)=6 \times 5 \times 4 / (3 \times 2 \times 1)=20$ groupes. Pour les rôles ordonnés : $6 \times 5 \times 4=120$. Chaque groupe correspond à six permutations.
2. $P(X=2)=28 \times 0,25^2 \times 0,75^6=0,31146$ environ. Pour au moins un, utiliser $1-0,75^8=0,89989$. Ces deux événements ne sont pas complémentaires.
3. Par tirages successifs : $(3/10)(2/9)=1/15$. Par groupes : $C(3,2)/C(10,2)=3/45=1/15$. La seconde méthode ignore l'ordre au numérateur et au dénominateur de façon cohérente.
4. $E=2 \times 0,3=0,6$. $\text{Var}=2 \times 0,3 \times 0,7 \times 8/9=0,37333$. Avec remise, la variance serait $0,42$. L'espérance reste identique, la dépendance négative réduit la variance.



5. $\lambda=0,4 \times 5=2$. $P(X=3)=e^{-2} \times 2^3/6=0,18045$. Le taux initial n'est pas le paramètre de la fenêtre de cinq minutes.
6. $P(T>5)=e^{-2}=0,13534$ et $E(T)=1/0,4=2,5$ min. Une moyenne de 2,5 min n'interdit pas une attente supérieure à cinq minutes.
7. La somme vaut zéro ou deux ; sa moyenne est un et sa variance $0,5(0-1)^2+0,5(2-1)^2=1$. Les deux essais sont parfaitement dépendants et ne définissent pas une binomiale à deux essais indépendants.
8. Le seuil correct est 14,5 : la barre correspondant à l'entier 15 commence à 14,5 dans cette représentation. Il faut ensuite standardiser et discuter si l'approximation est suffisamment précise.

Références et portée

NIST, loi binomiale ; NIST, loi de Poisson ; NIST, loi exponentielle. Références vérifiées le 2 octobre 2026. Modèles et jeux numériques pédagogiques ; la sélection d'un modèle pour une donnée réelle demande l'étude du protocole et de ses dépendances.

Partie 3 • Conditionnement, arbres et théorème de Bayes

Prérequis et objectifs

Connaître proportions, ensembles et lecture d'un tableau à double entrée. Le conditionnement change le groupe de référence.

À l'issue de ce cours, vous devez pouvoir :

- Définir indépendance et incompatibilité.
- Construire un arbre ou un tableau de probabilités.
- Calculer une probabilité après observation.

Fondements du thème

Événements et règles de calcul

Une probabilité attribuée à un événement a une valeur entre 0 et 1. Le complément A^c vérifie $P(A^c) = 1 - P(A)$. L'union correspond à « A ou B », avec inclusion éventuelle des deux ; $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Deux événements incompatibles ne peuvent se réaliser ensemble ; leur intersection est vide.

Deux événements indépendants vérifient $P(A \cap B) = P(A)P(B)$. Indépendance et incompatibilité sont très différentes : deux événements incompatibles de probabilités strictement positives ne sont pas indépendants. Le fait qu'une personne ait une caractéristique peut changer la probabilité d'une autre sans que l'une cause nécessairement l'autre.

Changer de dénominateur

La probabilité conditionnelle $P(A|B) = P(A \cap B)/P(B)$, si $P(B) > 0$, restreint l'univers aux personnes ou situations où B est réalisé. $P(\text{maladie}|\text{test positif})$ et $P(\text{test positif}|\text{maladie})$ n'ont pas le même dénominateur. Les confondre est une erreur fréquente de raisonnement diagnostique.

Le théorème de Bayes donne $P(A|B) = P(B|A)P(A)/P(B)$. Si A et A^c forment une partition, $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$. Un arbre probabiliste permet de lire les chemins : on multiplie le long d'une branche et on additionne les branches incompatibles menant au même résultat. Des effectifs fictifs peuvent rendre le calcul plus intuitif.

Exemple de test

Dans une population de 10 000 personnes, supposons une prévalence de 1 %, une sensibilité de 90 % et une spécificité de 95 %. Il y a 100 malades : 90 tests positifs vrais et 10 négatifs faux. Parmi les 9900 non-malades, 495 tests sont faussement positifs. La valeur prédictive positive vaut $90/(90 + 495) = 15,4$ % environ.

Le test peut donc être assez sensible et spécifique sans qu'un résultat positif suffise à confirmer une maladie rare. Cette conclusion dépend des hypothèses de transportabilité des performances du test vers la population étudiée. Une confirmation diagnostique peut être nécessaire selon le contexte ; le calcul seul ne définit pas une conduite clinique.

Développements et raisonnements

Une probabilité conditionnelle est un changement d'univers

Si 30 personnes sur 100 possèdent B et que 12 possèdent A et B, $P(A|B)$ vaut 12/30. Le dénominateur n'est plus l'ensemble initial de 100 personnes. Le même tableau peut fournir $P(B|A)$, mais il faut connaître le total des personnes possédant A. La notation avec une barre verticale doit être lue comme « parmi celles qui vérifient ».

La formule du produit $P(A \text{ et } B) = P(A|B)P(B)$ ne demande pas l'indépendance. Celle-ci est une condition supplémentaire permettant de remplacer $P(A|B)$ par $P(A)$. Deux événements incompatibles de probabilités non nulles sont dépendants : savoir que l'un s'est produit rend l'autre impossible. Confondre ces deux mots conduit à attribuer une probabilité positive à une intersection vide.

Retrouver Bayes par deux écritures du même événement

L'événement « malade et positif » peut se compter parmi les malades ou parmi les positifs. Son expression est à la fois $P(+|M)P(M)$ et $P(M|+)P(+)$. Les égaux puis diviser par $P(+)$ donne Bayes. Le dénominateur $P(+)$ comprend les vrais positifs et les faux positifs : $Se \times \text{prévalence} + (1 - Sp) \times (1 - \text{prévalence})$.

Le passage aux cotes fournit une autre écriture. La cote avant test est $p/(1 - p)$. Après un résultat positif, elle est multipliée par le rapport de vraisemblance $Se/(1 - Sp)$. La probabilité finale vaut cote/(1 + cote). Une cote de 4 signifie une probabilité de 80 %, pas de 400 %.

Deux tests ne sont pas automatiquement deux informations indépendantes

Multiplier deux rapports de vraisemblance suppose une structure de dépendance adaptée, souvent l'indépendance conditionnelle des tests chez les malades et chez les non-malades. Deux tests mesurant presque le même mécanisme peuvent être très corrélés. Les traiter comme deux preuves indépendantes surestime alors le gain d'information. La répétition exacte d'une même lecture ne crée pas une nouvelle observation biologique.

Quantité	Condition donnée	Population du dénominateur
Sensibilité	Maladie présente	Malades
Spécificité	Maladie absente	Non-malades
VPP	Test positif	Positifs
VPN	Test négatif	Négatifs

Les performances observées dans une étude peuvent changer avec le spectre de sévérité, la méthode de référence et la réalisation du test. La formule de Bayes reste vraie ; ce sont les valeurs conditionnelles utilisées qui doivent correspondre au contexte.

Atelier de calcul et analyse de sensibilité

Dans une population fictive de 5000 personnes, $p = 4\%$, $Se = 80\%$ et $Sp = 95\%$. On attend 200 malades, dont 160 positifs, et 4800 non-malades, dont 240 positifs. La VPP vaut $160/(160 + 240) = 40\%$. Le rapport de vraisemblance positif est $0,80/0,05 = 16$. La cote initiale $0,04/0,96 = 1/24$ devient $16/24 = 2/3$; la probabilité finale est $(2/3)/(5/3) = 0,40$. Les deux méthodes se contrôlent mutuellement.

Si p passe à 20% avec les mêmes performances, les vrais positifs deviennent 800 et les faux positifs 200, donc $VPP = 80\%$. La sensibilité n'a pas augmenté. La composition du groupe des positifs a changé parce que davantage de personnes étaient malades avant le test.

Pour un test négatif dans le premier contexte, le rapport de vraisemblance négatif vaut $0,20/0,95 = 0,210526$. La cote devient $1/114$ et la probabilité de maladie $1/115 = 0,8696\%$. La VPN est donc $99,1304\%$. Il faut préciser si la question demande une probabilité de maladie après un négatif ou une probabilité de non-maladie : ces deux valeurs sont complémentaires.

Approfondissement et contrôle du modèle

Conditionnement sur une information imparfaite

Bayes peut être appliqué à toute observation définie, pas seulement à un test médical. La probabilité après observation dépend de sa fréquence chez les personnes possédant ou non la caractéristique recherchée. Si ces fréquences sont elles-mêmes estimées, le résultat final possède une incertitude ; remplacer les paramètres par des valeurs ponctuelles ne la supprime pas.

Un résultat indéterminé doit être traité comme une catégorie observable avec ses probabilités conditionnelles, ou faire l'objet d'une règle explicite. Le supprimer revient à conditionner sur le fait d'avoir un résultat interprétable, ce qui peut changer la population. Deux groupes ayant des fréquences d'indétermination différentes ne se comparent pas correctement si cette étape disparaît du tableau.

Enfin, une probabilité prétest peut provenir d'une fréquence populationnelle ou d'informations propres au contexte étudié. Utiliser une prévalence générale après avoir déjà sélectionné un sous-groupe à risque peut être incohérent. Le calcul doit commencer dans la population effectivement concernée par l'observation. Le meilleur moyen de vérifier ce point est de nommer en mots le dénominateur de chaque probabilité avant d'appliquer la formule.

Contrôle de cohérence des probabilités

Une probabilité calculée doit rester entre zéro et un et les événements d'une partition doivent totaliser un. Ce contrôle détecte une erreur de dénominateur ou de complément, mais ne valide pas les hypothèses du modèle. Il faut aussi vérifier que les groupes considérés recouvrent bien la population annoncée et que les valeurs conditionnelles correspondent au même contexte. Une substitution correcte de nombres incompatibles entre études peut produire une réponse arithmétiquement plausible mais scientifiquement inadaptée.



Lecture visuelle : Test : les dénominateurs changent le sens du pourcentage

Test : les dénominateurs changent le sens du pourcentage

Population fictive : 1 000 ; prévalence 10 %

	Maladie + (100)	Maladie - (900)
Test +	90 vrais +	90 faux +
Test -	10 faux -	810 vrais -

Sensibilité = $90 / 100 = 90 \%$

VPP = $90 / (90 + 90) = 50 \%$

Spécificité = $810 / 900 = 90 \%$; aucun seuil clinique déduit.

Les effectifs concernent 1 000 personnes fictives : 100 avec maladie et 900 sans maladie. Sensibilité et spécificité valent chacune 90 %. La valeur prédictive positive est pourtant 50 %, car seuls 90 des 180 résultats positifs correspondent à la maladie.

Question de lecture. Quel changement de dénominateur sépare sensibilité et VPP ?

Correction. Sensibilité : tous les malades (100). VPP : tous les résultats positifs (180). Remplacer l'un par l'autre ignore la prévalence et la population de référence.

Mise à l'épreuve : Tester une dépendance plutôt que la supposer

Problème. Dans une population fictive, 10 % ont la condition recherchée. Un test a une sensibilité de 80 % et une spécificité de 90 %. Comparer la VPP après deux résultats positifs si les erreurs sont conditionnellement indépendantes, puis si le second résultat est une copie parfaite du premier.

Résolution. Sous indépendance conditionnelle dans chaque groupe, les probabilités de double positivité sont $0,8^2$ et $0,1^2$. La VPP vaut $0,10 \times 0,64 / [0,10 \times 0,64 + 0,90 \times 0,01] = 87,7 \%$. Avec une copie parfaite, aucune information n'est ajoutée : la VPP reste $0,08 / (0,08 + 0,09) = 47,1 \%$. Ces deux scénarios illustratifs n'estiment pas un protocole réel et ne sont pas des bornes générales pour toutes les dépendances possibles. Elles montrent pourquoi multiplier les performances exige une hypothèse sur la dépendance conditionnelle, et pas simplement deux dates ou deux appareils.

Schéma de raisonnement

Bayes en effectifs

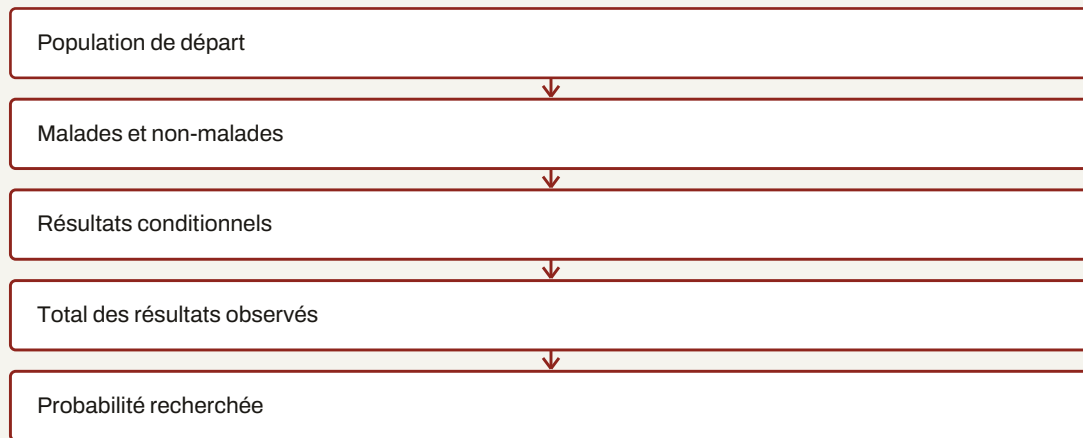


Schéma de principe original • lire les conditions et limites dans le texte

Le tableau fictif utilise des effectifs attendus. Une étude réelle possède aussi une incertitude sur prévalence, sensibilité et spécificité.

Exercices progressifs

1. $P(A) = 0,4$, $P(B) = 0,3$ et $P(A \cap B) = 0,1$. Calcule $P(A \cup B)$.
2. Deux événements indépendants ont des probabilités 0,2 et 0,5. Probabilité de leur intersection ?
3. Dans l'exemple du test, calcule la valeur prédictive négative.
4. $P(A) = 0,5$ et $P(B|A) = 0,3$. Donner $P(A \text{ et } B)$.
5. Une cote après test vaut 0,25. Probabilité correspondante ?
6. Pourquoi deux tests répétés très corrélés ne doivent-ils pas être supposés indépendants sans justification ?

Corrections argumentées

1. $0,4 + 0,3 - 0,1 = 0,6$. L'intersection était comptée deux fois.
2. $0,2 \times 0,5 = 0,10$. Cette multiplication exige l'indépendance annoncée.
3. Vrais négatifs = 9405 ; négatifs totaux = 9415 ; $VPN = 9405/9415 \approx 99,89 \%$. Cette forte valeur est en partie liée à la rareté de la maladie.
4. 0,15. La formule utilise une probabilité conditionnelle et n'exige donc pas que A et B soient indépendants.
5. $0,25/(1 + 0,25) = 0,20$, soit 20 %. La cote compare survenue et non-survenue.
6. Une information commune peut être comptée deux fois. Il faut connaître les performances conjointes ou formuler une hypothèse d'indépendance conditionnelle appropriée.



Références et portée

- Penn State : cours de statistique, référence 1
- Penn State : cours de statistique, référence 2

Sources consultées le 1er octobre 2026 pour les notions signalées. Développements, situations et calculs pédagogiques OSSIK ; données fictives. La concordance complète avec les supports universitaires de l'année et la relecture scientifique humaine restent à réaliser. Les modèles biomédicaux ne fournissent aucune décision thérapeutique individuelle.

Partie 4 • Variables aléatoires et lois usuelles

Le cours 13-01, partie 2 construit les coefficients de dénombrement et les lois binomiale, hypergéométrique et de Poisson à partir du protocole ; il complète ces définitions par des calculs et cas de dépendance.

Prérequis et objectifs

Savoir calculer factorielle, exponentielle et moyenne pondérée. Identifier si l'observation est un compte, une catégorie ou une mesure continue.

À l'issue de ce cours, vous devez pouvoir :

- Définir espérance et variance.
- Choisir entre binomiale, Poisson et normale.
- Vérifier les hypothèses d'une approximation.

Fondements du thème

Espérance et variance

Une variable aléatoire associe une valeur numérique à chaque issue d'une expérience. Pour une variable discrète, l'espérance est la somme des valeurs pondérées par leurs probabilités. Elle représente une moyenne théorique sur des répétitions, pas nécessairement une valeur observable. La variance vaut $E[(X - E(X))^2] = E(X^2) - E(X)^2$.

Pour $Y = aX + b$, $E(Y) = aE(X) + b$ et $\text{Var}(Y) = a^2\text{Var}(X)$. Ajouter une constante décale la distribution sans changer sa dispersion. Pour une somme, les espérances s'additionnent même sans indépendance ; les variances s'additionnent si les covariances sont nulles, notamment sous indépendance. Oublier les covariances sous-estime souvent l'incertitude de mesures liées.

Bernoulli et binomiale

Une variable de Bernoulli prend 1 avec probabilité p et 0 avec probabilité $1 - p$. Son espérance vaut p et sa variance $p(1 - p)$. Une loi binomiale $B(n, p)$ compte les succès de n essais indépendants ayant la même probabilité p . Sa probabilité en k est $C(n, k)p^k(1 - p)^{n - k}$, avec $C(n, k)$ le nombre de façons de choisir les k succès.

La moyenne vaut np et la variance $np(1 - p)$. Un échantillonnage sans remise dans une petite population finie n'est pas strictement binomial : les tirages sont dépendants. Une approximation peut être pertinente si la fraction prélevée est faible, mais doit être annoncée.

Poisson et normale

La loi de Poisson décrit un nombre d'événements dans une unité de temps ou d'espace sous un modèle de taux homogène et d'occurrences indépendantes. $P(X = k) = \frac{\exp(-\lambda)\lambda^k}{k!}$, avec espérance et variance égales à λ . Le paramètre λ représente le nombre moyen attendu dans la fenêtre considérée ; doubler cette fenêtre double λ si le taux est constant. Des événements regroupés ou un taux variable peuvent rendre le modèle inadéquat.

La loi normale $N(\mu, \sigma^2)$ est continue, symétrique autour de μ . La variable standardisée $Z = (X - \mu)/\sigma$ suit $N(0,1)$. Environ 95 % des observations se trouvent entre $\mu - 1,96\sigma$ et $\mu + 1,96\sigma$. Pour une variable continue, la probabilité d'une valeur exacte isolée est nulle ; les probabilités portent sur des intervalles. Une densité peut dépasser 1 sans que la probabilité totale dépasse 1, puisque c'est l'aire qui compte.

Développements et raisonnements

Une loi est un modèle de production des données

Une variable de Bernoulli décrit une seule issue binaire. Une binomiale est la somme de n Bernoulli indépendantes de même probabilité. Cette construction explique son espérance np et sa variance $np(1 - p)$. Si les probabilités diffèrent entre personnes, l'espérance reste la somme des probabilités, mais la loi n'est plus la binomiale ordinaire. Si les issues sont corrélées, des covariances interviennent dans la variance.

Le coefficient $C(n,k)$ compte les positions possibles des k succès parmi n essais. Le produit $p^k(1 - p)^{n - k}$ décrit une séquence particulière ; multiplier par le coefficient additionne les séquences incompatibles ayant le même total. Oublier ce facteur sous-estime fortement la probabilité des valeurs centrales.

Poisson : un compte dépend de sa fenêtre

Le paramètre λ est le nombre moyen d'événements dans la fenêtre définie. Si le taux est r par heure et que l'on observe t heures, $\lambda = rt$. La probabilité de zéro vaut $\exp(-rt)$, celle d'au moins un vaut $1 - \exp(-rt)$. Deux heures ne donnent pas une probabilité exactement double de celle d'une heure, même si le nombre moyen double.

Une variance empirique bien supérieure à la moyenne peut signaler une hétérogénéité de taux ou une dépendance entre événements. Cela ne prouve pas un mécanisme précis, mais invite à discuter le modèle de Poisson simple. Un appareil saturant peut aussi produire des comptes déformés par son temps mort.

Normale, standardisation et densité

Standardiser une mesure consiste à retirer son centre puis à diviser par son écart-type. Le résultat Z est sans unité. Si X suit une normale, on utilise la normale centrée réduite pour calculer des aires. Une densité au point x n'est pas la probabilité de $X = x$: la probabilité d'un intervalle est l'aire correspondante.

Loi	Valeurs possibles	Paramètres et contrôle
Bernoulli	0 ou 1	Une probabilité p
Binomiale	Entiers de 0 à n	n fixé, même p , indépendance
Poisson	Entiers positifs ou nuls	Taux et fenêtre définis
Normale	Réels	Moyenne et variance positives

Une approximation normale d'un compte discret demande des effectifs suffisants dans les deux queues pertinentes. Une correction de continuité peut améliorer certaines probabilités. Le simple fait que n soit grand ne suffit pas si np reste très petit.

Atelier de calcul et analyse de sensibilité

Pour X binomiale avec $n = 5$ et $p = 0,20$, $P(X = 2) = 10 \times 0,20^2 \times 0,80^3 = 0,2048$. L'espérance vaut 1, mais 1 n'est pas une prédiction certaine du nombre observé. La probabilité d'au moins un succès se calcule plus vite par le complément : $1 - 0,80^5 = 0,67232$.

Un dispositif reçoit en moyenne 2 événements par minute suivant un modèle de Poisson homogène. Sur 30 secondes, $\lambda = 1$ et $P(0) = e^{-1} = 0,367879$. Sur deux minutes, $\lambda = 4$ et $P(0) = e^{-4} = 0,018316$. Le nombre moyen et la probabilité de ne rien voir ne suivent donc pas la même relation linéaire avec le temps.

Une mesure suit une normale de moyenne 50 et d'écart-type 4. L'intervalle $[42 ; 58]$ correspond à Z entre -2 et $+2$ et couvre environ 95,45 % des observations du modèle. Si l'on moyenne 16 mesures indépendantes, l'écart-type de la moyenne devient 1 ; environ 95,45 % des moyennes sont alors entre 48 et 52. La distribution des individus n'a pas été resserrée : c'est une autre variable aléatoire que l'on étudie.

Approfondissement et contrôle du modèle

Sommes de variables et indépendance

L'espérance d'une somme égale la somme des espérances même lorsque les variables sont dépendantes. La variance demande en revanche les covariances. Deux mesures très corrélées n'apportent donc pas la même réduction de variabilité de la moyenne que deux mesures indépendantes. Cette propriété relie directement un modèle probabiliste au plan de recueil.

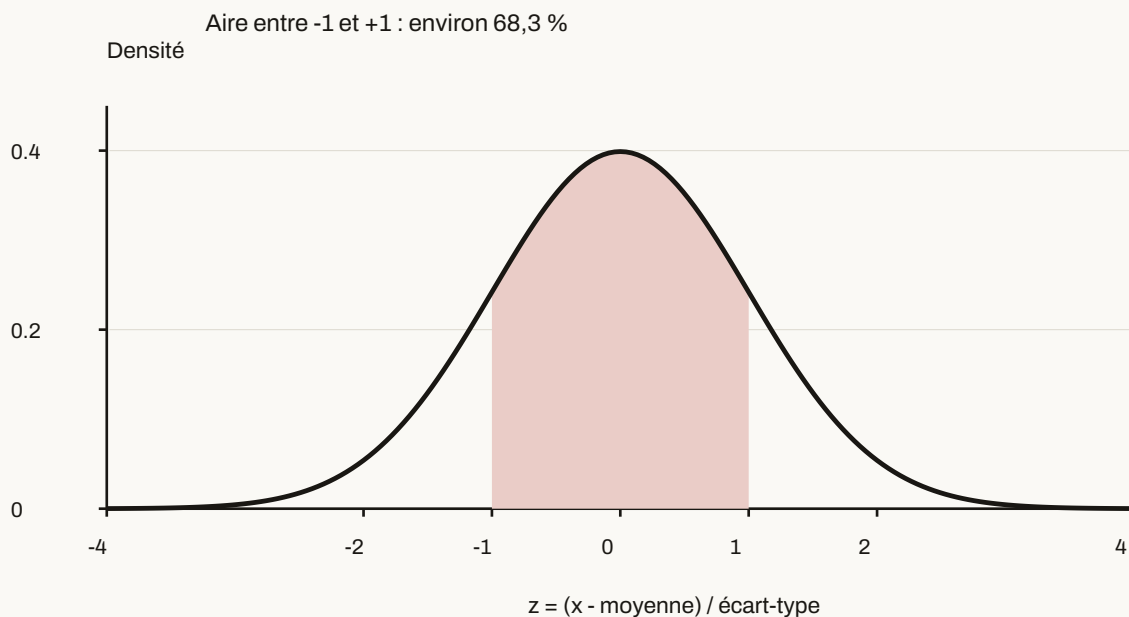
La somme de variables de Poisson indépendantes possède une loi de Poisson dont le paramètre est la somme des paramètres. Cette propriété peut justifier l'agrégation de fenêtres indépendantes dans un processus homogène ou de sources indépendantes. Si des événements surviennent en grappes, le même total peut avoir une dispersion plus grande que sa moyenne et ne plus suivre ce modèle simple.

Une variable normale transformée de façon affine reste normale, avec moyenne et variance adaptées. Une transformation non linéaire ne conserve pas en général la normalité. Par exemple, l'exponentielle d'une normale est positive et asymétrique. Reconnaître ces transformations aide à choisir une loi à partir du mécanisme plutôt qu'à partir d'une ressemblance visuelle approximative de l'histogramme.



Lecture visuelle : Loi normale centrée réduite : densité et aire

Loi normale centrée réduite : densité et aire



La zone colorée est l'aire entre -1 et +1 sous la densité normale centrée réduite, environ 68,3 %. La hauteur n'est pas une probabilité ponctuelle. Les queues continuent au-delà du domaine représenté $[-4, 4]$.

Question de lecture. Standardiser une distribution asymétrique la rend-il normale ?

Correction. Non : centrage et réduction changent position et échelle, pas sa forme. La référence normale exige une hypothèse ou une approximation justifiée.

Mise à l'épreuve : Une variable de Poisson n'est pas son indicateur

Problème. Le nombre X d'événements dans une fenêtre suit une loi de Poisson de paramètre 2. On code $Y = 1$ si au moins un événement survient et $Y = 0$ sinon. Donner $E(X)$, $E(Y)$ et $\text{Var}(Y)$.

Résolution. $E(X) = 2$: un compte peut avoir une moyenne supérieure à un. En revanche, Y suit une Bernoulli de paramètre $p = 1 - \exp(-2) = 0,8647$. $E(Y) = 0,8647$ et $\text{Var}(Y) = p(1 - p) = 0,1170$. Assimiler ces deux variables produirait une « probabilité de 200 % ». Le codage binaire perd l'information sur les répétitions. Si la durée de la fenêtre double à taux constant, le paramètre du compte double, mais la probabilité d'au moins un événement ne double pas : elle reste inférieure à un.

Schéma de raisonnement

Choisir une loi

1. Nature de la variable • 2. Mécanisme de répétition • 3. Paramètres • 4. Probabilité calculée • 5. Concordance aux données



Une loi ne doit pas être choisie uniquement parce qu'une formule est familière. Les dépendances et l'hétérogénéité peuvent invalider les calculs usuels.

Exercices progressifs

1. Dix essais indépendants ont chacun $p = 0,2$. Moyenne et variance du nombre de succès ?
2. Pour X de Poisson de paramètre 3, calcule $P(X = 0)$.
3. X suit $N(100, 15^2)$. Quelle valeur de Z correspond à $X = 130$?
4. Pour X binomiale $B(20 ; 0,1)$, calculer moyenne et écart-type.
5. Un taux Poisson vaut 3/h. Probabilité d'au moins un événement en 20 minutes ?
6. Pour $Y = 2X - 3$, avec $E(X) = 5$ et $\text{Var}(X) = 4$, donner $E(Y)$ et $\text{Var}(Y)$.

Corrections argumentées

1. Espérance = $10 \times 0,2 = 2$; variance = $10 \times 0,2 \times 0,8 = 1,6$. L'écart-type est $\sqrt{1,6}$, pas 1,6.
2. $P(0) = \exp(-3) \approx 0,0498$, car $3^0 = 1$ et $0! = 1$.
3. $Z = (130 - 100)/15 = 2$. La valeur est située deux écarts-types au-dessus de la moyenne.
4. Moyenne 2 ; variance $20 \times 0,1 \times 0,9 = 1,8$; écart-type racine(1,8) = 1,342.
5. La fenêtre vaut un tiers d'heure, donc $\lambda = 1$. La probabilité est $1 - e^{-1} = 0,632121$.
6. $E(Y) = 7$ et $\text{Var}(Y) = 2^2 \times 4 = 16$. La constante -3 ne change pas la variance.

Références et portée

- Penn State : cours de statistique, référence 1

Sources consultées le 1er octobre 2026 pour les notions signalées. Développements, situations et calculs pédagogiques OSSIK ; données fictives. La concordance complète avec les supports universitaires de l'année et la relecture scientifique humaine restent à réaliser. Les modèles biomédicaux ne fournissent aucune décision thérapeutique individuelle.

Partie 5 • Échantillonnage et estimation par intervalle

Prérequis et objectifs

Connaître écart-type, indépendance, quantiles et distribution normale. Distinguer population fixe et échantillon aléatoire.

À l'issue de ce cours, vous devez pouvoir :

- Construire un intervalle de moyenne.
- Évaluer une précision et un effectif.
- Interpréter la couverture sans la confondre avec une probabilité individuelle.

Fondements du thème

D'une série à une estimation

La loi des grands nombres explique la stabilisation de certaines moyennes d'échantillons indépendants autour de leur espérance. Elle ne supprime pas un biais de sélection : répéter massivement une mauvaise méthode peut produire une estimation très précise d'une mauvaise cible. Le théorème central limite donne, sous des conditions adaptées, une distribution approximativement normale de la moyenne standardisée quand n augmente. Il ne transforme pas la distribution individuelle en loi normale.

Un estimateur possède un biais, une variance et éventuellement d'autres propriétés. L'intervalle de confiance décrit l'incertitude d'estimation selon une procédure répétable. Dans l'interprétation fréquentiste, un intervalle à 95 % est issu d'une méthode qui couvre le paramètre fixe dans 95 % des échantillons répétés sous les hypothèses du modèle. Il ne contient pas nécessairement 95 % des observations individuelles. Cette distinction et les formules de moyenne sont vérifiées dans le **NIST**, *Confidence Limits for the Mean*.

Moyenne et proportion

Pour des observations indépendantes normales avec écart-type de population inconnu, l'intervalle de la moyenne est $\bar{x} \pm t_{(n-1;0,975)} s/\sqrt{n}$. Le quantile de Student dépasse 1,96 pour les petits échantillons ; il s'en rapproche quand n grandit. Une approximation normale peut être employée dans d'autres conditions, en précisant ses limites, notamment asymétrie marquée et valeurs extrêmes.

Pour une proportion $\hat{p}$ et un grand effectif, l'intervalle de Wald $\hat{p} \pm 1,96\sqrt{\hat{p}(1-\hat{p})/n}$ est simple mais parfois médiocre. Il peut sortir de $[0,1]$ et sous-couvrir près des limites. Des méthodes comme Wilson ou des méthodes binomiales exactes sont souvent préférables. Le **NIST** rappelle les limites des approximations symétriques lorsque les effectifs sont faibles.

Précision et effectif

À dispersion et niveau de confiance constants, la largeur de l'intervalle décroît comme $1/\sqrt{n}$. Diviser la marge d'erreur par deux demande environ quatre fois plus d'observations. Passer de 95 % à 99 % élargit l'intervalle pour le même échantillon. Ces règles supposent un plan d'échantillonnage comparable ; des observations groupées par centre ou famille peuvent apporter moins d'information que le même nombre d'observations indépendantes.

Développements et raisonnements

L'objet aléatoire est la procédure d'estimation

Avant l'échantillonnage, la moyenne, l'écart-type et les bornes de l'intervalle sont variables. Après observation, l'intervalle est fixé, tandis que le paramètre demeure inconnu. La propriété de couverture à 95 % concerne la répétition de la procédure sous ses hypothèses. Elle ne dit pas que 95 % des individus sont entre les bornes, ni que 95 % des futures moyennes seront dans cet intervalle particulier.

Un intervalle de prédiction pour une observation future comprend la variabilité individuelle en plus de l'incertitude sur la moyenne. Il est donc généralement plus large. Un intervalle étroit peut coexister avec une forte diversité entre personnes lorsque l'échantillon est grand.

Dériver l'effectif à partir de la marge recherchée

Dans une approximation normale avec écart-type anticipé σ , la demi-largeur vaut $e = z\sigma/\text{racine}(n)$. Isoler n donne $n = (z\sigma/e)^2$. On arrondit vers le haut, car un effectif inférieur ne respecte pas la précision prévue. Ce calcul suppose que l'estimation de σ et le plan d'échantillonnage soient adaptés. Des pertes de suivi attendues demandent une inflation séparée.

Pour une proportion, $n \approx z^2 p(1-p)/e^2$. Le produit $p(1-p)$ est maximal à $p = 0,5$, ce qui fournit une option prudente lorsque p est inconnu. Cette formule ne remplace pas une méthode adaptée aux très faibles proportions ou aux petits effectifs. La méthode de Wald peut donner une largeur nulle si aucun événement n'est observé, ce qui contredit l'incertitude réelle.

Effet de plan et non-réponse

Des personnes recrutées dans les mêmes classes ou centres peuvent être corrélées. Avec des grappes de taille m et une corrélation intraclasse ρ dans un modèle simple, un effet de plan approximatif est $1 + (m-1)\rho$. L'effectif informatif peut être inférieur au nombre total de participants. Cette expression est un repère de modèle, pas une correction universelle de tous les plans.

Une non-réponse sélective peut produire un biais qui n'apparaît pas dans l'erreur standard ordinaire. Les intervalles usuels quantifient l'erreur d'échantillonnage sous le modèle ; ils ne couvrent pas automatiquement les erreurs de mesure, de sélection et de spécification. Présenter un intervalle très précis sans ces limites peut donner une impression trompeuse de certitude.

Atelier de calcul et analyse de sensibilité

Un échantillon de 16 observations indépendantes, approximativement normales, donne moyenne 40 et $s = 8$. L'erreur standard est 2. Avec le quantile de Student à 15 degrés de liberté, environ 2,131, l'intervalle à 95 % vaut $40 \pm 4,262$, soit [35,738 ; 44,262]. Utiliser 1,96 donnerait un intervalle un peu plus étroit, sans respecter ici l'hypothèse de variance inconnue en petit échantillon.

Pour viser une marge de 2 unités avec σ anticipé à 10 et $z = 1,96$, $n = (1,96/2)^2 = 96,04$, soit 97 personnes observées. Avec 20 % de pertes prévues indépendantes du résultat, un calcul de recrutement simple donne $97/0,80 = 121,25$, soit 122 personnes. Cette inflation protège le nombre final, mais ne corrige pas un biais si les pertes sont informatives.

Pour une proportion anticipée de 0,20, une marge de 0,04 donne $n = 1,96^2 \times 0,20 \times 0,80/0,04^2 = 384,16$, soit 385. Avec p inconnu et $p = 0,5$ retenu par prudence, l'effectif monte à 601. Le choix d'une précision

absolue doit être distingué d'une précision relative : 4 points de pourcentage n'ont pas le même poids relatif autour de 2 % et de 50 %.

Approfondissement et contrôle du modèle

Bootstrap : reproduire une procédure, pas inventer une population

Le bootstrap rééchantillonne avec remise les observations disponibles pour examiner la variabilité d'une statistique. Il peut aider lorsque sa distribution analytique est compliquée. Mais il hérite de la qualité et de la structure des données initiales : un échantillon biaisé reste biaisé et des observations groupées demandent un rééchantillonnage respectant les groupes.

Rééchantillonner des cellules une à une quand elles proviennent de quelques personnes peut simuler une indépendance inexistante. Le niveau de tirage doit correspondre à l'unité pertinente. La méthode ne corrige pas non plus automatiquement une petite taille ne contenant aucune observation d'une région importante de la distribution.

Un intervalle bootstrap peut prendre plusieurs formes, avec des propriétés différentes. Il faut indiquer la méthode, le nombre de réplifications et les hypothèses plutôt que présenter « bootstrap » comme une garantie universelle. Une simulation supplémentaire réduit surtout l'erreur numérique de la procédure ; elle ne crée pas de nouveaux participants et ne supprime pas l'incertitude scientifique liée au recrutement. Cette distinction explique pourquoi mille réplifications bootstrap ne correspondent pas à mille nouvelles observations indépendantes.

Mise à l'épreuve : Zéro observation : une borne, pas zéro risque

Problème. Aucun événement n'est observé chez 20 personnes indépendantes, toutes suivies sur une même durée. Calculer la borne supérieure unilatérale exacte à 95 % du risque binomial.

Résolution. Pour une probabilité p , $P(\text{zéro}) = (1 - p)^{20}$. La limite retenue résout $(1 - p)^{20} = 0,05$, donc $p = 1 - 0,05^{1/20} = 0,1391$, soit 13,9 %. La règle approximative $3/n$ donnerait 15 %. Un intervalle bilatéral exact à 95 % utiliserait 0,025 dans cette équation et aurait une borne supérieure différente, environ 16,8 %. « À 95 % » ne suffit donc pas à spécifier la procédure. Ces bornes supposent l'indépendance, un risque commun et un suivi complet ; elles ne traitent pas les perdus de vue.

Estimer une proportion près des bornes

Pourquoi une erreur standard peut devenir trompeuse

L'intervalle de Wald utilise $p_{\text{chapeau}} \pm z\sqrt{[p_{\text{chapeau}}(1-p_{\text{chapeau)})/n]}$. Quand aucun événement n'est observé, $p_{\text{chapeau}}=0$ et cette expression donne une largeur nulle. Cela contredit l'incertitude réelle sur une probabilité inconnue : un échantillon sans événement n'exclut pas une petite probabilité positive.

L'intervalle de Wilson vient de l'inversion d'une inégalité de score utilisant la variance $p(1-p)/n$ sous chaque valeur candidate p , plutôt que de figer cette variance à l'estimation observée. Son centre est $[p_{\text{chapeau}}+z^2/(2n)]/[1+z^2/n]$, et sa demi-largeur $z\sqrt{[p_{\text{chapeau}}(1-p_{\text{chapeau)})/n+z^2/(4n^2)]/[1+z^2/n]}$. Les bornes restent dans $[0 ; 1]$. Ce n'est pas une garantie de couverture exactement nominale pour toute binomiale, mais une réponse mieux adaptée à de nombreux petits échantillons.



Exercice guidé

Sur 20 observations indépendantes identiquement distribuées, un événement survient une fois. Avec $z=1,96$, calculer l'intervalle de Wald puis celui de Wilson à 95 %.

Correction. $p_{\text{chapeau}}=0,05$. Wald donne $0,05 \pm 0,09552$, soit $[-0,04552 ; 0,14552]$: la borne négative est physiquement impossible. Wilson donne un centre 0,12251 et une demi-largeur 0,11363, soit environ $[0,00888 ; 0,23614]$. Remplacer seulement la borne négative de Wald par zéro ne reconstitue pas la procédure de Wilson et ne répare pas sa couverture.

Exercice d'interprétation

Le résultat $[0,00888 ; 0,23614]$ signifie-t-il que 95 % des individus ont un risque dans cet intervalle ?

Non. L'intervalle concerne un paramètre de proportion dans le modèle et la procédure d'échantillonnage, pas la distribution de risques individuels. Sa lecture fréquentiste porte sur la couverture de la procédure lors de répétitions hypothétiques.

Dépendance entre observations, sélection et mauvaise définition de l'événement ne sont pas corrigées par le choix de Wilson. Un intervalle mathématiquement mieux construit ne répare pas un protocole biaisé. Référence de méthode : **NIST, intervalles de confiance d'une proportion.**

Schéma de raisonnement

Dimensionner une estimation

1. Paramètre cible • 2. Dispersion anticipée • 3. Niveau de confiance • 4. Marge et effectif • 5. Plan et pertes

L'effectif calculé dépend d'hypothèses préalables. Il ne garantit pas la validité externe ni l'absence de biais.

Exercices progressifs

1. $\bar{x} = 50$, $s = 10$, $n = 100$. Donne un intervalle normal approximatif à 95 %.
2. Combien faut-il multiplier n pour diviser la marge d'erreur par trois ?
3. Un intervalle de confiance à 95 % contient-il 95 % des patients ?
4. Pour $s = 12$ et $n = 144$, quelle erreur standard de la moyenne ?
5. Combien de fois faut-il augmenter n pour diviser la marge par quatre ?
6. Zéro événement parmi 20 personnes prouve-t-il que la probabilité est nulle ?

Corrections argumentées

1. $SE = 10/10 = 1$; $IC \approx [48,04 ; 51,96]$. Si les conditions appellent Student, le quantile correspondant remplace 1,96.
2. Par neuf, puisque la marge est proportionnelle à $1/\sqrt{n}$.
3. Non. Il concerne un paramètre, par exemple une moyenne. Un intervalle de prédiction pour une observation future répond à une autre question et est généralement plus large.
4. $SE = 12/12 = 1$ unité sous indépendance. L'écart-type individuel reste 12.
5. Par seize, car la marge dépend de $1/\text{racine}(n)$.
6. Non. Une issue rare peut ne pas apparaître dans un petit échantillon. Une borne binomiale adaptée reste strictement positive ; le Wald dégénéré à zéro est trompeur.

Références et portée

- Penn State : cours de statistique, référence 1
- NIST : manuel de statistique, référence 2

Sources consultées le 1er octobre 2026 pour les notions signalées. Développements, situations et calculs pédagogiques OSSIK ; données fictives. La concordance complète avec les supports universitaires de l'année et la relecture scientifique humaine restent à réaliser. Les modèles biomédicaux ne fournissent aucune décision thérapeutique individuelle.