François Dress

Les PROBABILITÉS et la STATISTIQUE

de A à Z

500 DÉFINITIONS, FORMULES ET TESTS D'HYPOTHÈSE

DUNOD

AVERTISSEMENT

Ce dictionnaire est un dictionnaire généraliste principalement destiné aux étudiants des premières années d'université et aux utilisateurs professionnels non mathématiciens.

Une attention toute particulière a été portée aux étudiants et aux utilisateurs dans les domaines des sciences expérimentales et des sciences économiques et sociales. Dans le souci de leur fournir un outil de travail adapté, les articles ont été rédigés en restant au niveau mathématique le plus élémentaire possible. En outre, chaque fois que cela apparaît nécessaire, la définition est précédée par une courte introduction en langage « courant ».

Certains termes et concepts mathématiques de base, en nombre très restreint et dont la vertu est de clarifier les définitions et d'éviter les ambiguïtés, sont néanmoins utilisés. En particulier, on se réfère souvent à l'espace probabilisé $(\Omega, \mathcal{A}, \mathrm{P})$, notion qui se comprend facilement : Ω est l'ensemble de tous les résultats possibles, $\mathcal{A}$ est l'ensemble de tous les événements auxquels on pourra attribuer une probabilité, et P est cette mesure de probabilité, dont les valeurs sont comprises entre 0 et 1… le niveau d'abstraction n'est pas insupportable !

Toujours dans le souci de la meilleure efficacité, ce dictionnaire a un parti pris de redondance : redondance de contenu entre les articles, et souvent redondance du vocabulaire à l'intérieur de chaque article.

Enfin, ce dictionnaire inclut un certain nombre d'articles, ou de parties d'articles, dont on pourrait qualifier le niveau mathématique d'intermédiaire, qui lui permettront d'être également utile aux étudiants de mathématiques des premières années.

Nous avons choisi de garder pour certains concepts le vocabulaire usuel même lorsqu'il n'est pas très bien choisi et représente un héritage malheureux des siècles précédents (la « variable » aléatoire !). En outre la terminologie n'est pas complètement fixée. Nous avons systématiquement indiqué les variantes les plus courantes. En tout état de cause, si l'utilisateur trouve insolite l'utilisation d'un mot dans tel ou tel manuel, il ne doit pas se laisser perturber, mais chercher simplement quel sens précis a ce mot dans ce manuel. La même remarque peut être faite pour les notations. Seule exception à notre ouverture aux notations variées, nous noterons P(A) la probabilité de l'événement A, et n'utiliserons jamais Pr(A) ou Prob(A).

Un détail enfin : en règle générale, toutes les valeurs numériques qui sont données dans les exemples sont des valeurs approchées (par arrondi à la décimale la plus proche) sans que cela soit marqué par la (trop lourde) présence de points de suspension. Bien entendu, une valeur approchée peut aussi être exacte, et il est laissé au lecteur le soin de le détecter.

Quelques conseils matériels d'utilisation

- Lorsqu'une expression contient plusieurs mots, l'article correspondant est répertorié au mot le plus significatif (mais de nombreux renvois facilitent la recherche).
- Les crochets dans une entrée d'article ou un synonyme encadrent une partie optionnelle de l'expression.
- Les mots du texte qui renvoient à une autre entrée sont mis en italiques.
- Enfin, quelques rares expressions n'ont aucun équivalent dans l'usage anglo-saxon (par exemple « épreuves répétées »), elles ne sont donc pas traduites.

François Dress
dress@math.u-bordeaux.fr

absolument continue (variable aléatoire) *(absolutely continuous random variable)*

Voir *variable aléatoire (typologie).*

acceptation (région d') *(acceptance region)*

Voir *région d'acceptation*.

additivité *(additivity)*

Soit un espace probabilisable $(\Omega, \mathcal{A})$. On s'intéresse aux applications définies sur l'ensemble $\mathcal{A}$ des évènements et à valeurs réelles positives, et on veut caractériser la propriété qui fera d'elles des mesures (en langage mathématique général) ou des mesures de probabilité alias probabilités (en langage spécifique au calcul des probabilités).

À un niveau élémentaire, on se limite à la réunion finie, avec deux formules qui sont d'un usage extrêmement fréquent (P est une mesure de probabilité) :

si A et B sont disjoints : $P(A \cup B) = P(A) + P(B)$

de façon générale : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Exemple On considère une population humaine $\mathcal{P}$ et le système ABO de groupes sanguins. Les probabilités qu'un individu tiré au hasard dans la population $\mathcal{P}$ présente l'un des 6 génotypes possibles dans ce système sont données par le tableau ci-dessous :

OO	OA	AA	OB	BB	AB
0,435	0,364	0,076	0,086	0,004	0,035

Si l'on examine maintenant les phénotypes, l'évènement « phénotype A » est constitué par la réunion des évènements « OA » et « AA », qui sont disjoints, et sa probabilité est donc P(OA) + P(AA) = 0,440.

À un niveau plus approfondi, il faut envisager la réunion infinie dénombrable, pour laquelle l'axiome de σ-additivité énoncé ci-dessous doit être vérifié.

On dit qu'une application $\mu : \mathcal{A} \to [0, +\infty[$ est σ-additive si, pour toute suite $(A_n)(n = 1, 2, ...)$ d'éléments de $\mathcal{A}$, deux à deux disjoints, on a :

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

On notera une conséquence importante : si les A_n forment une suite croissante (au sens de la théorie des ensembles, *i.e.* si $A_n \subset A_{n+1}$), alors $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} \mu(A_n)$.

Voir *espace probabilisé, mesure de probabilité, Poincaré (formule de).*

affine (fonction) *(affine function)*

Fonction de la forme $y = a + bx$ que l'on rencontre notamment en statistique comme courbe de régression. Une telle fonction est souvent appelée « linéaire » par abus de langage.

ajustement linéaire, polynomial, exponentiel, logarithmique *(fitting)*

On considère deux grandeurs numériques x et y intervenant dans l'étude d'un phénomène. On suppose qu'il existe une liaison entre ces deux grandeurs, et que cette liaison est aléatoire (qu'il s'agisse d'un phénomène déterministe pertubé par exemple par des erreurs de mesure, ou d'un phénomène intrinsèquement aléatoire). Le plus souvent, cette liaison sera représentée – ou sera présumée pouvoir être représentée – par une relation mathématique simple, par exemple : $y = bx$ (liaison linéaire), $y = a + bx$ (liaison affine, le plus souvent qualifiée aussi de linéaire), $y = bx^P$ (liaison puissance), $y = a_0 + a_1x + ... + a_nx^n$ (liaison polynomiale), $y = ae^{kx}$ (liaison exponentielle), $y = a + b \ln x$ (liaison logarithmique).

Le problème se pose alors, connaissant un ensemble de points expérimentaux (un « nuage », dans le vocabulaire statistique), de trouver la forme mathématique de la liaison et les « bons » coefficients (ou paramètres) numériques. Lorsque le phénomène sera déterministe pertubé aléatoirement, le nuage de points sera très voisin de la courbe $y = f(x)$ que l'on cherche à déterminer. Lorsque le phénomène sera intrinsèquement aléatoire, le nuage sera plus dispersé et la courbe $y = f(x)$ sera la courbe qui passera au mieux au « milieu » du nuage des points. La courbe ainsi déterminée, dans un cas comme dans l'autre, sera appelée courbe d'ajustement ou, dans le cas particulier affine, *droite d'ajustement*.

Si $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ est le nuage de points, et si $f = f_{a, b, ...}$ est la « forme » connue ou présumée de la relation mathématique, dépendant des paramètres a, b,…, la méthode standard (due à Gauss et Legendre) s'appelle la méthode des moindres carrés : elle consiste à déterminer les valeurs des paramètres a, b,… comme celles qui minimisent la somme des carrés des écarts $\sum_{n=1}^{\infty} (y_i - f(x_i))^2$. Dans le cas particulier affine et dans une situation de type déterministe pertubé aléatoirement, on obtient ainsi la *droite des moindres carrés*.

Dans une situation de type intrinsèquement aléatoire, la somme des carrés des écarts n'est pas exactement une variance, mais elle est directement liée aux variances des variables aléatoires que l'on doit introduire pour modéliser la situation. L'ensemble des conditions à satisfaire inclut une minimisation de la variance et on aboutit exactement aux mêmes formules et aux mêmes valeurs des paramètres. Le contexte probabiliste–statistique induit un changement de vocabulaire : plutôt que de droite d'ajustement ou de droite des moindres carrés, on parle de *droite de régression* ; plutôt que de courbe d'ajustement, on parle de *courbe de régression*.

Signalons pour terminer que les liaisons linéaire, affine et polynomiale sont directement traitées par les techniques (équivalentes) de moindres carrés ou de régression linéaire, tandis que les liaisons puissance et exponentielle sont le plus souvent ramenées aux cas précédents par passage aux logarithmes : $y = bx^P$ devient $\ln y = \ln b + p \ln x$, $y = ae^{kx}$ devient $\ln y = \ln a + kx$.

ajustement

Voir *droite d'ajustement* et *khi–deux d'ajustement (test du)*.

aléatoire *(random)*

Cet adjectif qualifie dans la langue courante des phénomènes dont les « résultats » sont variables, non ou mal maîtrisables, imprévisibles, …, et s'oppose à « déterministe ». Il est égale-

ment employé, avec un sens mathématique codifié, dans des expressions qui désignent certains concepts du calcul des probabilités (expérience aléatoire, variable aléatoire, processus aléatoire, ...).

Le calcul des probabilités a pour but la « modélisation » des phénomènes aléatoires, dès lors qu'ils sont reproductibles dans des conditions identiques. Selon les cas, cette reproduction pourra être provoquée par l'homme (expérimentation), ou sera « naturelle » (observation). La reproductibilité est essentielle notamment pour fonder l'estimation des probabilités, ce qui explique l'extrême difficulté – conceptuelle et pratique – d'évaluer la probabilité d'un évènement très rare.

On notera que la variabilité des phénomènes qualifiés d'aléatoires peut être soit intrinsèque (variabilité d'un caractère biologique, imprévisibilité d'une désintégration radioactive), soit liée à l'imprécision ou à l'erreur de la mesure de la grandeur qui peut être déterministe.

Voir *hasard, modélisation*, *variable aléatoire*.

algèbre d'évènements *(field of events)*

Famille (ensemble) $\mathcal{A}$ d'évènements qui vérifient les propriétés qui caractérisent une « algèbre de Boole » de parties d'un ensemble E : appartenance de E à $\mathcal{A}$, « stabilité » par réunion finie et par passage au complémentaire.

Si on ajoute la stabilité par réunion dénombrable, on obtient une σ-algèbre ou *tribu*.

σ–algèbre

Voir *tribu*.

alphabet grec

Voir *grec (alphabet)*.

amplitude *(amplitude)*

Terme souvent utilisé pour désigner la largeur d'une classe (bornée) : l'amplitude de la classe $]a_j, a_{j+1}]$ est $|a_{j+1} - a_j|$.

analyse combinatoire *(combinatorial analysis, combinatorics)*

Branche des mathématiques qui étudie les « configurations », formées à partir d'« objets » pris dans un ensemble fini donné et disposés en respectant certaines contraintes ou certaines structures fixées. Les deux problèmes principaux sont l'énumération des configurations, et leur dénombrement.

Les dénombrements (arrangements, combinaisons, permutations) jouent un rôle important en *probabilités combinatoires*, où l'hypothèse d'équiprobabilité ramène la détermination des probabilités à des comptes d'évènements élémentaires.

analyse de la variance (test d') *(variance analysis test)*

Test paramétrique qui permet de comparer globalement plusieurs espérances mathématiques entre elles. La dénomination du test s'explique par son fonctionnement, qui décompose la variance de l'ensemble des observations en deux variances partielles, la première fournissant une estimation « inter-classes » de la variance commune, et la seconde une estimation « intra-classes » (ou « résiduelle »). On teste ensuite le quotient : si la première estimation est sensiblement plus grande que la deuxième, on rejette alors l'hypothèse H_0 d'égalité de toutes les espérances.

test de comparaison de q espérances mathématiques μ_1 , μ_2, ..., μ_q

• Données. q classes ou groupes, soit pour chaque k ($1 \leq k \leq q$) : un échantillon $(x_{k1}, x_{k2}, ..., x_{kn_k})$ de n_k valeurs observées d'une variable aléatoire numérique X_k d'espérance mathématique μ_k. On note N le nombre total $n_1 + n_2 + ... + n_q$ de valeurs observées.

• Hypothèse testée. H_0 = « $\mu_1 = \mu_2 = ... = \mu_q$ » contre H_1 = « il existe au moins deux espérances différentes ».

• Déroulement technique du test

1a. Pour chaque k, on calcule la moyenne de l'échantillon n° k :

$$m_k = \frac{x_{k1} + x_{k2} + \ldots + x_{kn_k}}{n_k}.$$

1b. On calcule la moyenne générale :

$$M = \frac{n_1 m_1 + n_2 m_2 + \ldots + n_q m_q}{N}.$$

2. On calcule les 2 sommes de carrés qui « analysent » la variance complète :

$$Q_1 = \sum_{k=1}^{q} n_k (m_k - M)^2,$$

$$Q_2 = \sum_{k=1}^{q} \left(\sum_{i=1}^{n_k} (x_{ki} - m_k)^2 \right).$$

(*nota* : la somme $Q_1 + Q_2$ est la somme complète $\displaystyle\sum_{k=1}^{q} \left(\sum_{i=1}^{n_k} (x_{ki} - M)^2 \right)$).

On calcule ensuite $s_1^2 = \dfrac{Q_1}{q-1}$, estimation « inter-classes » de la variance commune, et

$s_2^2 = \dfrac{Q_2}{N-q}$, estimation « intra-classes » (« résiduelle »).

3. On calcule enfin la valeur observée de la variable de test :

$$F_{q-1,\,N-q} = \frac{s_1^2}{s_2^2}.$$

Les valeurs de référence de la variable de test sont à lire dans les tables de la loi de Fisher-Snedecor. Elles dépendent des deux degrés de liberté $q-1$ et $N-q$, et du risque α. On notera que les tables de la loi de Fisher-Snedecor sont unilatérales. En effet, si H_0 est fausse, alors l'estimation inter-classes de la variance ne peut être qu'augmentée.

• Conditions et précautions

– En théorie les X_k doivent être des v.a. normales ; en outre elles doivent avoir même variance, exigence difficile à apprécier et à contrôler... S'il semblait que cela ne soit pas le cas, on peut appliquer le (*test de*) *Bartlett*, ou bien rechercher dans un ouvrage spécialisé d'autres procédures applicables ;

– lorsque les X_k ne sont pas normales, le test est robuste et reste applicable si les effectifs des groupes sont « assez grands ».

analyse de la variance à un facteur *(one-way ANOVA)*

Méthode d'analyse d'une situation où l'on expérimente l'effet d'un facteur sur une variable, facteur supposé agir exclusivement sur l'espérance mathématique. Dans un premier temps de l'analyse, on teste l'existence d'un effet du facteur par le (*test d'*) *analyse de la variance* : la variance inter-classes s'interprète alors comme la part de variance due au facteur considéré, et le terme de variance résiduelle se justifie alors pleinement.

Si l'hypothèse H_0 est rejetée, cela signifie que le facteur a un effet, et dans un deuxième temps de l'analyse, on estime les espérances dont l'ensemble décrit l'effet du facteur.

La présentation globale et les notations données ci-dessous sont très légèrement différentes de celles du test d'analyse de la variance, de façon à pouvoir les généraliser naturellement au cas de deux facteurs.

analyse de la variance à un facteur A

• Données. p classes ou groupes, avec pour chaque niveau i ($1 \leq i \leq p$) du facteur : un échantillon $(x_{i1}, x_{i2}, ..., x_{in_i})$ de n_i valeurs observées d'une variable aléatoire numérique X_i d'espérance mathématique μ_i.

On note N le nombre total $n_1 + n_2 + ... + n_p$ de valeurs observées.

• Modèle étudié

Les observations individuelles sont de la forme $X_{ij} = \mu_i + E_{ij}$, où E_{ij} est un écart de loi normale centrée et d'écart-type σ (indépendant de l'effet du facteur A).

On pose $\mu_i = \mu + \alpha_i$, où μ représente la moyenne globale et α_i l'effet du facteur A (en convenant que la moyenne pondérée des effets est nulle).

➤ Premier temps

• Hypothèse sur l'effet du facteur. H_0 = « le facteur ne possède aucun effet » = « $\mu_1 = \mu_2 = ... = \mu_p$ » contre H_1 = « il existe au moins deux espérances différentes ».

• Déroulement technique du test

– On calcule la moyenne observée m_i de l'échantillon n° i ;
– on calcule la moyenne générale observée M (moyenne pondérée des m_i) ;
– on calcule les sommes de carrés relatives aux parts de variance observées et on regroupe les résultats comme indiqué dans le tableau suivant :

Origine de la dispersion	Somme de carrés	ddl	Estimation de la variance	Test F
totale	$Q = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - M)^2$	$N - 1$	$\frac{Q}{N-1}$	
facteur A	$Q_A = \sum_{i=1}^{p} (m_i - M)^2$	$p - 1$	$s_A^2 = \frac{Q_A}{p-1}$	$F_{p-1,\,N-p} = \frac{s_A^2}{s_R^2}$
résiduelle	$Q_R = Q - Q_A$	$N - p$	$s_R^2 = \frac{Q_R}{N-p}$	

Les valeurs de référence de la variable de test sont à lire dans les tables de la loi de Fisher-Snedecor.

• Conditions et précautions. Voir *analyse de la variance* (test d').

➤ Deuxième temps (en cas de rejet de H_0)

• estimations des effets

– μ est estimée par M,

– les α_i sont estimés par $m_i - M$.

analyse de la variance à deux facteurs *(two-way ANOVA)*

Méthode d'analyse d'une situation où l'on expérimente l'effet de deux facteurs sur une variable, facteurs supposés agir exclusivement sur l'espérance mathématique. Il faut distinguer les effets séparés des deux facteurs et envisager *a priori* une interaction possible.

Dans un premier temps de l'analyse, on teste l'existence de l'effet des facteurs et l'existence d'une interaction par plusieurs tests de type *analyse de la variance*.

Si l'un ou l'autre de ces tests rejette l'hypothèse de non-existence d'un effet, il faut alors, dans un deuxième temps de l'analyse, estimer les espérances dont l'ensemble décrit les effets des deux facteurs ainsi que leur interaction.

Il existe plusieurs « plans d'expérience » possibles pour préciser le détail de l'expérimentation. Les formules qui suivent concernent les deux plans d'expérience les plus courants, qui sont des plans « complets » et « équilibrés » : toutes les associations entre un « niveau » du premier facteur et un niveau du second sont observées, et toutes les classes présentent le même nombre r d'observations. Le premier plan est « à répétitions » ($r \geq 2$), le deuxième sans répétition ($r = 1$), auquel cas l'on ne peut pas évaluer les interactions et l'on estime seulement l'effet des deux facteurs.

1. analyse de la variance à deux facteurs A et B
plan d'expérience complet équilibré avec répétitions

• Données. pq classes ou groupes, avec pour chaque couple (i, j) $(1 \leq i \leq p,\ 1 \leq j \leq q)$ de niveaux des facteurs : un échantillon $(x_{ij1}, x_{ij2}, ..., x_{ijr})$ de r valeurs observées d'une variable aléatoire numérique X_{ij} d'espérance mathématique μ_{ij}.

Le nombre total N de valeurs observées est égal à pqr.

• Modèle étudié. Les observations individuelles sont de la forme $X_{ijk} = \mu_{ij} + E_{ijk}$, où E_{ijk} est un écart de loi normale centrée et d'écart-type σ (indépendant de l'effet des facteurs A et B et de leur interaction).

On pose $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, où μ représente la moyenne globale et α_i l'effet du facteur A, β_j l'effet du facteur B, γ_{ij} l'interaction entre A et B (en convenant que la moyenne pondérée des effets et de l'interaction est nulle).

➤ Premier temps

• Hypothèse globale sur l'existence d'effet(s). H_0 = « il n'y a aucun effet » = « toutes les μ_{ij} sont égales » contre H_1 = « il existe au moins deux espérances différentes ».

• Déroulement technique du test

– On calcule la moyenne observée m_{ij} de chaque classe (i, j) ;

– on calcule les moyennes marginales observées $m_{i.}$ et $m_{.j}$;

– on calcule la moyenne générale observée M ;

– on calcule les sommes de carrés relatives aux parts de variance observées et on regroupe les résultats comme indiqué dans le tableau ci-contre.

– on effectue trois tests séparés, les valeurs de référence de la variable de test sont à lire dans les tables de la loi de Fisher-Snedecor.

Origine de la dispersion	Somme de carrés	ddl	Estimation de la variance	Test F
totale	$Q = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(x_{ijk} - M)^2$	$pqr - 1$	$\frac{Q}{pqr-1}$	
facteur A	$Q_A = qr\sum_{i=1}^{p}(m_{i.} - M)^2$	$p-1$	$s_A^2 = \frac{Q_A}{p-1}$	$F_{p-1,\,pq(r-1)} = \frac{s_A^2}{s_R^2}$
facteur B	$Q_B = pr\sum_{j=1}^{q}(m_{.j} - M)^2$	$q-1$	$s_B^2 = \frac{Q_B}{q-1}$	$F_{q-1,\,pq(r-1)} = \frac{s_B^2}{s_R^2}$
interaction AB	$Q_{AB} = r\sum_{i=1}^{p}\sum_{j=1}^{q}(m_{ij} - m_{.j} - m_{i.} + M)^2$	$(p-1)(q-1)$	$s_{AB}^2 = \frac{Q_{AB}}{(p-1)(q-1)}$	$F_{(p-1)(q-1),\,pq(r-1)} = \frac{s_{AB}^2}{s_R^2}$
résiduelle	$Q_R = Q - Q_A - Q_B - Q_{AB}$	$pq(r-1)$	$s_R^2 = \frac{Q_R}{pq(r-1)}$	

• Conditions et précautions. Voir *analyse de la variance* (test d').

➤ Deuxième temps (en cas de rejet de H_0)

• estimations des effets
– μ est estimée par M ;
– les α_i sont estimés par $m_{i.} - M$, les β_j sont estimés par $m_{.j} - M$;
– les γ_{ij} sont estimés par $m_{ij} - m_{.j} - m_{i.} + M$.

2. analyse de la variance à deux facteurs A et B plan d'expérience complet équilibré sans répétition

• Données. pq classes ou groupes, avec pour chaque couple (i, j) $(1 \leq i \leq p,\ 1 \leq j \leq q)$ de niveaux des facteurs : une valeur unique x_{ij} observée d'une variable aléatoire numérique X_{ij} d'espérance mathématique μ_{ij}.
On note N le nombre total pq de valeurs observées.

• Modèle étudié. Les observations individuelles sont de la forme $X_{ij} = m_{ij} + E_{ij}$, où E_{ij} est un écart de loi normale centrée et d'écart-type σ (indépendant de l'effet des facteurs A et B).
On pose $\mu_{ij} = \mu + \alpha_i + \beta_j$, où μ représente la moyenne globale et α_i l'effet du facteur A, β_j l'effet du facteur B (en convenant que la moyenne pondérée des effets est nulle).

➤ Premier temps

• Hypothèse globale sur l'existence d'effet(s). H_0 = « il n'y a aucun effet » = « toutes les μ_{ij} sont égales » contre H_1 = « il existe au moins deux espérances différentes ».

• Déroulement technique du test
– On calcule les moyennes marginales observées $m_{i.}$ et $m_{.j}$;
– on calcule la moyenne générale observée M ;