Université Lumière Lyon2

ISPEF

Département de Sciences de l'éducation

Licence SHS Sciences de l'éducation
Master1 Sciences de l'éducation

Statistique et traitement des données

Vocabulaire élémentaire de statistique

Cours de Jean-Claude Régnier
Diane Diaz
Fadhila Fahrane Horrigue

Amplitude d'une classe (intervalle d'une variable continu)
C'est la longueur d'un intervalle. L'amplitude de la classe [20 ;55[est de 55-20=30

Caractère qualitatif nominal
Voir variable qualitative nominale

Caractère qualitatif ordinal
Voir variable qualitative ordinale

Caractère quantitatif continu
Voir variable quantitative continue

Caractère quantitatif discret
Voir variable quantitative discrète

Caractère statistique
Voir variable statistique
C'est ce qui est observé sur les individus d'une population ou d'un échantillon

Centile
Les centiles C1, C2, ….C99 partagent une série statistique en 100 parties d'effectifs égaux.

Khi-deux
La variable du Khi-deux est utilisée par l'intermédiaire de tables statistiques pour le test du Khi-deux (*voir* test du Khi-deux)

Classe (intervalle) modale
C'est la classe ou intervalle d'une variable continue d'effectif ou de fréquence maximum. Elle correspond sur l'histogramme de la distribution d'un caractère quantitatif continu au « rectangle de surface maximale ».

Classes ou intervalles d'une variable continue
Pour un caractère (une variable) quantitatif continu, il convient de découper l'ensemble des valeurs possibles en intervalles ou classes d'amplitudes égales ou inégales. Par exemple si la distribution des âges de la population étudiée et comprise entre 0 et moins de 100 ans, on pourra construire des classes : [0 ;15[, [15 ;25[,[25 ;35[,[35 ;45[etc….

Coefficient de corrélation linéaire de Bravais-Pearson
Le coefficient de corrélation entre deux variables quantitatives X et Y est le nombre noté r tel que : $r = \frac{\operatorname{cov}(X,Y))}{\sigma(X).\sigma(Y)}$.

On peut montrer que : $-1 \leq r \leq +1$

Interprétation : Le coefficient de corrélation linéaire est un nombre sans unité ; Si r est nul ou proche de zéro, les deux variables étudiées sont non-corrélées linéairement. Le nuage de points n'a pas de direction privilégiée. Si r est proche de 1 ou -1, les deux variables X et Y sont bien corrélées linéairement, c'est-à-dire qu'il existe entre elles une liaison de type linéaire. Le nuage de points est presque aligné le long d'une droite (ascendante si r proche de 1, et descendante si r proche de -1)

Coefficient de variation

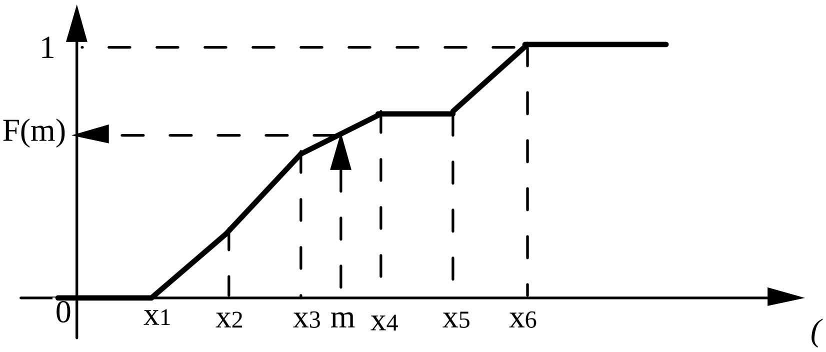
Ce coefficient est obtenu en calculant le rapport de l'écart-type à la moyenne.

$$CV = \frac{\text{écart – type}}{\text{moyenne}} = \frac{\sigma}{\mu}$$

Il est parfois exprimé en % et il permet de s'affranchir des unités de mesure et des ordres de grandeur de la variable. Il permet ainsi de comparer plusieurs distributions de variables.

Courbe cumulative croissante

La courbe cumulative croissante est du type :



Elle représente la fréquence cumulée associée à une valeur x, c'est à dire le % d'observation dont la valeur est strictement inférieure à x.

Cette courbe permet de déterminer graphiquement les fractiles. La médiane correspond à l'antécédent de la valeur 0,5 (50%) sur la courbe des fréquences cumulées. La médiane est le fractile d'ordre 0,5

Covariance entre deux variables quantitatives X et Y

La covariance est un indicateur de co-variation entre deux séries numériques X et Y. Rappelons-nous que la variance d'une variable statistique X peut s'exprimer comme la moyenne des carrés des écarts à la moyenne. La covariance est définie comme la moyenne des produits des écarts à la moyenne. On peut alors calculer cette covariance de la manière suivante :

$$\operatorname{cov}(X,Y) = \frac{\sum_i (xi - \overline{x})(yi - \overline{y})}{n}$$

On peut vérifier que cette formule est équivalente à la formule suivante :

$$\operatorname{cov}(X,Y) = \frac{\sum_i xi.yi}{n} - \overline{x}.\overline{y}$$

Cette expression s'interprète comme étant : la moyenne des produits moins le produit des moyennes

Déciles

Les déciles D1, D2, …, D9 partagent une série statistique en 10 parties d'effectifs égaux.

Densité de fréquence

La densité de fréquence intervient pour exprimer la distribution des fréquences d'une variable continue. Dans le cas des variables statistiques que nous étudions empiriquement, la densité de fréquence s'obtient en divisant la fréquence relative à l'intervalle par l'amplitude cet intervalle. Il s'agit alors de la fréquence par unité d'intervalle et la densité d'effectif s'obtient en divisant l'effectif par l'amplitude.

Diagramme circulaire

Cette représentation graphique fournit une visualisation de la distribution des fréquences ou de celle des effectifs. On obtient une telle représentation graphique en établissant une correspondance entre la mesure en degré des angles et la valeur des effectifs ou des fréquences. Ainsi, pour la modalité x_k de fréquence $f_k\%$ exprimée en pourcentage :

mesure en degrés de l'angle de la portion représentant la modalité x_k vaut $\dfrac{360 \times f\%_k}{100}$

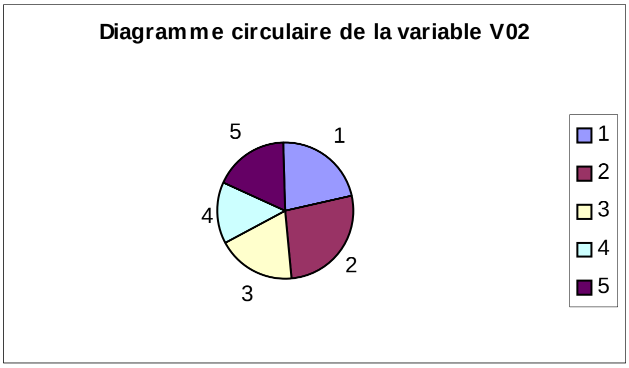
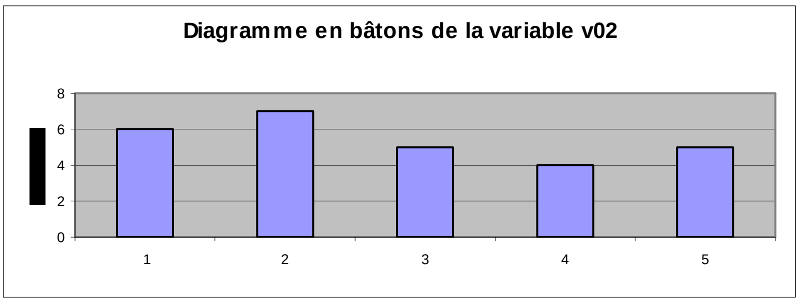


Diagramme en bâtons

Diagramme représentant une variable qualitative ou quantitative discrète. Cette représentation graphique fournit une visualisation de la distribution des fréquences ou de celle des effectifs. Chaque segment nommé « bâton », a une longueur proportionnelle à la fréquence ou à l'effectif correspondant à la modalité représentée. La largeur de ce « bâton » n'a aucun sens statistique.



Ecart-type d'une variable quantitative

L'écart-type est la racine carrée de la variance. Pour un variable quantitative discrète :

sur la population d'effectif N:

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{k=p} n_k\,(x_k - \mu)^2}$$

sur l'échantillon d'effectif n :

$$\sigma_{ech} = \sqrt{\frac{1}{n}\sum_{k=1}^{k=p} n_k (x_k - m)^2}$$

Pour une variable quantitative continue, les formules sont identiques en prenant pour valeur x_k, les centres des intervalles c_k.

Échantillon

L'échantillon est un sous-ensemble de la population.

Effectif

L'effectif est la quantité d'individus relative à un résultat (valeur ou modalité). Nous noterons $n_1,n_2,\ldots,n_p$, les effectifs correspondant respectivement aux résultats $x_1,x_2,\ldots,x_p$. La distribution des effectifs de la variable X n'est autre que la suite des (x_k, n_k) avec k = 1,..., p.

Effectif total

L'effectif total, noté N ou n, est la somme des effectifs relatifs à chaque résultat distinct. Nous avons avec $N = \sum_{k=1}^{k=p} n_k$ si nous travaillons sur la population et $n = \sum_{k=1}^{k=p} n_k$ si l'étude est restreinte à l'échantillon.

Effectifs cumulés

Voir courbe cumulée croissante

Estimation ponctuelle ou par intervalle de confiance d'une moyenne d'une variable sur une population

Estimation ponctuelle

Pour estimer la moyenne μ, nous mettons en œuvre la statistique $\overline{X}$ nommée moyenne empirique ou échantillonnale (le nom de la variable, ici X, surmonté d'une barre, on lit « X barre ») suivante qui consiste à calculer la moyenne des n résultats réalisés sur l'échantillon. L'opération est représentée symboliquement par la formule suivante:

$$m = \overline{X} = \frac{1}{n}\sum_{i=1}^{i=n} X_i$$

Estimation par intervalle de confiance au niveau 1-α

La fourchette d'estimation de la moyenne s'obtient avec la formule suivante appliquée aux résultats obtenus sur un échantillon aléatoire avec remise de taille n :

$$m - \kappa\frac{s}{\sqrt{n}} < \mu < m + \kappa\frac{s}{\sqrt{n}}$$

avec

$m = \overline{X} = \frac{1}{n}\sum_{i=1}^{i=n} X_i$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - m)^2$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - m)^2}$

Ainsi construit, cet intervalle [m- $\kappa \frac{s}{\sqrt{n}}$; m + $\kappa \frac{s}{\sqrt{n}}$] a une probabilité de $1-\alpha$ de contenir la vraie valeur μ qui nous est inconnue. La valeur critique κ est à chercher dans la table de Student ddl=n-1 ou dans la table de Laplace-Gauss, si n est supérieur à 50

Estimation ponctuelle et par intervalle de confiance d'une proportion π sur une population

Estimation ponctuelle

Elle est le résultat d'une réalisation de la fréquence empirique sur l'échantillon choisi. Elle est tout simplement obtenue à partir de la proportion calculée avec les valeurs observées sur l'échantillon : $\pi= f_n$ où f_n est la valeur de F_n calculée à partir de l'échantillon obtenu à l'issue de n tirages. Notons que $\frac{\pi(1-\pi)}{n}$ la variance de F_n est estimée ponctuellement soit par $\frac{f_n(1-f_n)}{n-1}$, soit par sa valeur maximale $\frac{1}{4(n-1)}$

Estimation par intervalle de confiance

Les deux situations usuelles sont:

- Intervalle de confiance à 95% : k=1,9600
- Intervalle de confiance à 99% : k=2,5758

Ensuite on utilise les résultats obtenus précédemment dans le cadre de l'estimation ponctuelle à savoir f_n. La fourchette d'estimation de la proportion s'obtient avec la formule suivante :

$$f_n - \kappa\sqrt{\frac{f_n(1-f_n)}{n-1}} < \pi < f_n + \kappa\sqrt{\frac{f_n(1-f_n)}{n-1}}$$

Ainsi construit, cet intervalle $\left[f_n - \kappa\sqrt{\frac{f_n(1-f_n)}{n-1}} ; f_n + \kappa\sqrt{\frac{f_n(1-f_n)}{n-1}} \right]$ a une probabilité de (1-α de contenir la vraie valeur π qui nous est inconnue. La valeur critique κ est à chercher dans la table de Laplace-Gauss, si n est supérieur à 100

Estimation ponctuelle de l'écart-type σ d'une variable sur une population

Elle est tout simplement obtenue à partir de la racine carrée de la variance calculée avec les valeurs observées sur l'échantillon à laquelle on applique le coefficient multiplicateur $\frac{n}{n-1}$ corrigeant le biais

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - m)^2} = \sqrt{\frac{n}{n-1}}\,\sigma_{\text{échantillon}}$$

Estimer

Estimer, c'est attribuer une valeur (ou une modalité) à un paramètre, ou une caractéristique inconnue tel qu'en particulier une moyenne, une variance, un écart type, une proportion, une médiane, un mode, un effectif. L'estimation peut être ponctuelle ou par intervalle de confiance, dite fourchette d'estimation.

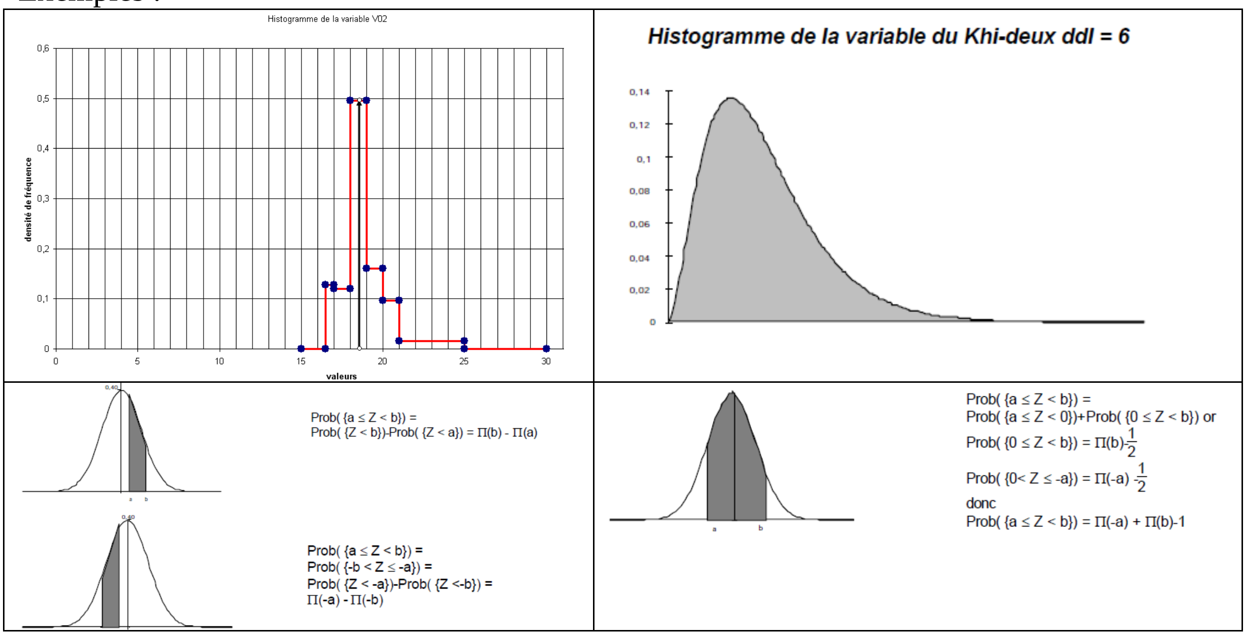
Étendue
L'étendue est l'intervalle entre la borne inférieure de rang 1 prise par X , et la borne supérieure de rang p prise par X c'est à dire [x_1 ;x_p]

Fractiles d'ordre α d'une variable quantitative
Le nombre f_α tel que la fréquence cumulée associée à f_α est égale à a. La médiane est le fractile d'ordre 0,5. Les quartiles sont les fractiles d'ordre 0,25 ; 0,5 ; 0,75.

Fréquence
La fréquence est définie par $f_k = \dfrac{n_k}{effectiftotal}$ où l'effectif total est N si le référent est la population et n si le référent est l'échantillon. (x_k, f_k) avec k = 1,..., p la distribution des fréquences de la variable X. Cette fréquence est souvent ramenée en pourcentage en multipliant f_k par 100.

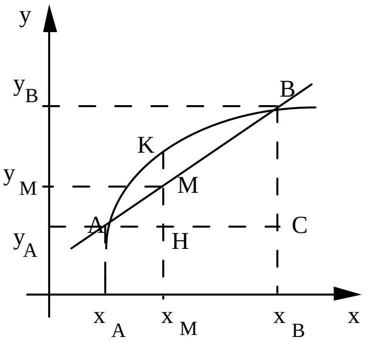
Fréquences cumulées
Voir courbe cumulée croissante.

Histogramme d'une variable continue
Représentation graphique délimitée par la courbe de la densité de fréquence, dont la surface représente la fréquence.
Exemples :



Hypothèses statistiques
Elles correspondent à un couple d'énoncés conjecturaux par rapport auxquels le chercheur prend une décision de choix au vu d'une observation sur un échantillon. Le premier énoncé est désigné par Ho (hypothèse nulle) et le second, désigné par H1 (hypothèse alternative) est obtenu par la négation logique de l'hypothèse nulle. Les tests statistiques ont pour objet d'aider le chercheur dans cette prise de décision en contrôlant les risques d'erreur encourus : [erreur de première espèce = rejet à tort de Ho ; erreur de seconde espèce = conservation à tort de l'hypothèse Ho]

Indépendance de deux événements



Supposons connues x_M, l'abscisse du point M et x_K, l'abscisse du point K , avec $x_M = x_K$, nous pouvons calculer y_M, l'ordonnée de M, qui fournit une valeur approchée de y_K , par ordonnée de K, grâce à la relation qui se fonde sur l'alignement des points A, M et B:

$$\frac{x_M - x_A}{x_B - x_A} = \frac{y_M - y_A}{y_B - y_A}$$

ou encore par la relation

$$\boxed{\frac{y_M - y_A}{x_M - x_A} = \frac{y_B - y_A}{x_B - x_A}}$$

Si c'est y_K qui est connue, nous pouvons alors calculer x_M en tant que valeur approchée de x_K.

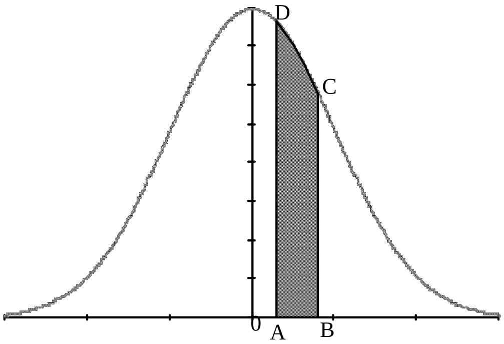
Deux évènements A et B sont dit aléatoirement indépendants si la probabilité de la réalisation de l'événement A et B est égal au produit de la probabilité de A par celle de B. Deux évènements A et B sont dit statistiquement indépendants si la fréquence des individus qui réalisent de l'événement A et B est égal au produit des fréquences de A et de B. Si nous considérons deux variables statistiques X et Y, par exemple X= « genre » et Y= « attitude à l'égard des mathématiques, l'événement A= « être une femme » et B= « aimer les mathématiques » peut être étudié du oint de vue de l'indépendance ou de la dépendance qui est la négation de l'indépendance.

Individu ou unité
L'individu ou l'unité statistique est un élément de la population.

Interpolation linéaire
Méthode d'estimation qui s'appuie sur la proportionnalité.

Intervalle interquartile
L'intervalle interquartile est l'intervalle [Q1 ;Q3]. Il représente approximativement les 50 % des réponses qui encadrent la médiane Q2. son amplitude, calculée par la différence Q3- Q1, mesure la dispersion autour de la médiane Q2.

Laplace Gauss LG(μ ;σ) (Variable de Laplace-Gauss ou variable normale)
Il s'agit d'une famille de variables continues mathématiquement bien connue dont l'importance théorique est remarquable. Chaque membre de cette famille est caractérisé par deux paramètres : (μ ; σ) qui ne sont autres que sa moyenne et son écart-type. La variable LG(0 ;1) centrée réduite est celle à partir de laquelle sont effectués les calculs concernant les autres variables LG(μ ;σ). L'histogramme d'une variable de Laplace-Gauss a une forme bien connue, appelée fréquemment « courbe de Gauss » ou courbe en cloche, dont l'usage par une analogie abusive se rencontre fréquemment.
La courbe de Gauss est mathématiquement définie par les formules suivantes :



fonction densité de fréquence

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$$

dont la fonction de répartition est

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} \exp(-\frac{t^2}{2})dt$$

La variable LG(0 ;1) joue un rôle très important pour d'autres variables dont les distributions peuvent être approchées par celle de LG(0 ;1).

Médiane
La médiane (notée Q2) est la modalité qui partage la population en deux parties de même effectif. La proportion des individus ayant choisi une modalité inférieure ou égale à Q2 est de 50 %.

Modalité
Les modalités sont les résultats possibles d'une variable qualitative.

Mode
Pour une variable quantitative discrète ou qualitative, le mode est une valeur de la variable correspondant à l'effectif maximum ou la fréquence maximum.

Pour une variable quantitative continue, le mode est une valeur de la variable correspondant à la densité de fréquence maximum.

Moyenne

La moyenne pour une variable quantitative discrète est la valeur de la variable X obtenue par l'une des deux procédures de calcul suivantes :

- sur la population d'effectif N, nous la désignons par la lettre grecque μ

$$\mu = \frac{1}{N}\sum_{i=1}^{i=N} o_i = \frac{1}{\mathrm{N}}\sum_{\mathrm{k=1}}^{\mathrm{k=p}} \mathrm{n_k x_k} = \sum_{k=1}^{k=p} f_k x_k$$

- sur l'échantillon d'effectif n , nous la désignons par la lettre latine *m*

$$m = \frac{1}{n}\sum_{i=1}^{i=n} o_i = \frac{1}{n}\sum_{k=1}^{k=p} n_k x_k = \sum_{k=1}^{k=p} f_k x_k$$

La moyenne pour une variable quantitative continue est la valeur de la variable X obtenue par l'une des deux procédures de calcul suivantes où c_k désigne les centres des intervalles:

- sur la population d'effectif N, nous la désignons par la lettre grecque μ

$$\mu = \frac{1}{N}\sum_{k=1}^{k=p} n_k c_k = \sum_{k=1}^{k=p} f_k c_k$$

- sur l'échantillon d'effectif n, nous la désignons par la lettre latine *m*

$$m = \frac{1}{n}\sum_{k=1}^{k=p} n_k c_k = \sum_{k=1}^{k=p} f_k c_k$$

Nuage (statistique) de points
Quand on dispose des n résultats (x_i ;y_i) issus de l'étude conjointe de deux variables quantitatives (X ;Y), on peut représenter graphiquement par n points M_i du plan repéré $(O, \vec{i}, \vec{j})$, dont les coordonnées sont (x_i ;y_i). L'ensemble géométrique ainsi obtenu est appelé « nuage de points ». Il s'agit alors d'interpréter cette forme géométrique pour rechercher le type de lien entre X et Y. La forme de base est évidemment la droite indiquant une liaison linéaire entre X et Y, c'est à dire : Y=aX+b

Population (statistique)
La population (ou univers statistique) est l'ensemble de tous les individus (ou unité statistique) concernés par la recherche.

Quartiles d'une variable quantitative ou d'une variable qualitative ordinale
Les Quartiles Q1, Q2, Q3 partagent une série statistique en 4 parties d'effectifs égaux. Q2 représente la médiane

Recensement
Le recensement est la collecte des informations de l'ensemble de tous les individus d'une population. Cette collecte est dite exhaustive. Le recensement national organisé par l'INSEE tous les 7 ans en est un exemple.

Région critique

La **région critique K** est l'ensemble des valeurs de la variable de décision qui conduisent à rejeter l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1. La région critique est déterminée par la relation : Prob{rejeter Ho / Ho vraie}= α.

Région d'acceptation

La **région d'acceptation A** est l'ensemble des valeurs de la variable de décision qui conduisent à ne pas rejeter l'hypothèse nulle H_0 au détriment de l'hypothèse alternative H_1.
La région critique K et la région d'acceptation A sont deux **ensembles complémentaires**. Dans cette perspective **construire un test** revient à construire une **variable de décision** explicite et à déterminer **une région critique**.

Représentativité
Cette notion est fortement perturbée par l'imprégnation du sens commun. L'idée de base est clairement d'obtenir une partie d'information par un échantillon qui « représente » au mieux la population, c'est à dire une partie qui apporte des informations suffisantes pour bien connaître le tout. Il y a une forme de représentativité qui est définie par le « modèle réduit ». Celle sur laquelle nous nous appuyons pour procéder aux opérations d'estimation ou de tests d'hypothèse est essentiellement fondée sur le caractère aléatoire de l'appartenance de l'individu à l'échantillon. En général c'est l'équiprobabilité qui est la propriété de référence mais des échantillons peuvent aussi être constitués aléatoirement à partir d'une appartenance non équiprobable des individus — ce qui importe est la connaissance de la probabilité d'appartenance de chaque individu à l'échantillon.

Sondage
Opération d'extraction d'un échantillon

Statistique descriptive bivariée
Cadre théorique pour l'analyse des propriétés fondée sur la description simultanée de deux variables conjointes.

Statistique descriptive univariée
Cadre théorique pour l'analyse des propriétés fondée sur la description d'une variable.

Statistique inférentielle
Cadre théorique permettant de conduire des raisonnements inductifs et déductifs pour une extension contrôlée des propriétés observées sur un échantillon à toute la population..

(Variable de) Student
La famille des variables de Student notée $T_{(n)}$ à n ddl est une famille bien connue mathématiquement tout comme la famille des variables de Laplace-Gauss. Chaque membre se distingue par son degré de liberté ddl = n. Dans la pratique elle est utilisée par l'intermédiaire de tables statistiques.

Tableau de contingence
Il s'agit du tableau statistique correspondant à l'étude conjointe de deux variables statistiques. Ce tableau donne les effectifs des couples de résultats de (X ;Y).

Test du Khi-deux d'adéquation
Le test du Khi-deux d'adéquation est une procédure mise en pratique pour comparer la distribution de fréquence d'une variable qualitative ou d'une variable quantitative observée sur un échantillon, à une distribution théorique connue sur la population.

Test du Khi-deux d'homogénéité
Le test du Khi-deux d'homogénéité est une procédure mise en pratique pour comparer des proportions relatives aux modalités d'une variable qualitative au sein de plusieurs populations

Test du Khi-deux d'indépendance
Le test du Khi-deux d'indépendance est une procédure mise en pratique pour décider du choix entre les deux hypothèses alternatives :
Ho les deux variables sont indépendantes
H1 les deux variables sont dépendantes

Test statistique, test d'hypothèse statistique
Procédure pour aider à la prise de décision entre deux hypothèses alternatives. Dans cette perspective, l'idée de vérité est remplacée par celle de vraisemblance. Un test statistique est alors une procédure logico-mathématique sur la quelle on s'appuie pour choisir entre deux hypothèses alternatives sur la base d'informations partielles issues d'un échantillon, celle qui est la plus vraisemblable.

		« Situation de nature » inconnue	
		Ho vraie H1 fausse	H1 vraie Ho fausse
Décision du chercheur	Rejeter Ho Conserver H1	Erreur de première espèce α	Correcte
	Conserver Ho Rejeter H1	Correcte	Erreur de seconde espèce β

		« Situation de nature » inconnue	
		Ho vraie H1 fausse	H1 vraie Ho fausse
Décision du chercheur	Rejeter Ho Conserver H1	Prob{Rejeter Ho sachant Ho vraie}= α	Prob{Conserver H1 sachant H1 vraie}= $1-\beta$
	Conserver Ho Rejeter H1	Prob{Conserver Ho sachant Ho vraie}= $1-\alpha$	Prob{Rejeter H1 sachant H1 vraie}= β

Variable statistique
La variable statistique qui modélise l'objet d'étude, est application mathématique qui à un individu (unité) statistique fait correspondre un résultat (quantité ou qualité) ; Ce concept fondamental renvoie au but principal de la science statistique qui est donné un cadre théorique à l'étude de la **variabilité**.

Variable statistique qualitative nominale

Une variable statistique est qualitative nominale si les résultats ne sont pas des quantités. Ses modalités sont alors codées à l'aide de mots ou d'expressions courtes ou parfois à l'aide de nombres qui n'autorisent cependant aucunement une interprétation en termes de grandeur. Par exemple la variable « genre » possède deux modalités « Masculin » et « Féminin », codées respectivement « 1 » et « 2 ».

Variable statistique qualitative ordinale

Une variable statistique est qualitative ordinale si les résultats qui ne sont pas des quantités, sont toutefois ordonnés. Ses modalités sont codées à l'aide de mots ou d'expressions courtes marquant une gradation ou parfois à l'aide de nombres. Dans ce dernier cas, l'ordre des nombres traduit la gradation. Par exemple la variable « goût pour la statistique » peut être modélisée avec les modalités suivantes :

Expression verbale	Je n'aime pas du tout	J'aime un peu	J'aime beaucoup	J'aime à la folie
Codage numérique	1	2	3	4

Variable quantitative continue

Les variables quantitatives continues sont celles qui ne sont pas discrètes. L'ensemble des résultats possibles est l'ensemble des nombres réels ou un sous-ensemble. Par exemple : le « poids » des cartables, la taille des élèves de quatrième, la distance entre le domicile et l'école, la durée du trajet entre le domicile et l'école.

Variable quantitative discrète
Les résultats sont alors de nature quantitative. L'ensemble des résultats possibles, que nous nommons valeurs est un ensemble de nombres qui possède une structure ayant des propriétés mathématiques remarquables habituelles autorisant l'usage de la comparaison, de la grandeur et des opérations arithmétiques ou algébriques. Ce que nous nommons variable quantitative discrète correspond à un ensemble fini ou dénombrable de résultats. Intuitivement chaque

résultat est isolable. Par exemple, les variables statistiques suivantes : nombre d'enfants par famille, nombre d'élèves par classe de seconde

Variable centrée réduite
Variable dont la moyenne est égale à 0 et l'écart-type est égal à 1. Si X est une variable de moyenne μ. et d'écart-type σ alors la variable Z obtenue par transformation de X par l'opération suivante $Z = \frac{X - \mu}{\sigma}$ est une variable centrée réduite. Cela revient à choisir la moyenne μ comme nouvelle origine des valeurs et l'écart-type σ comme nouvelle unité d'écart à la moyenne.

Variance
La variance ou moment centré d'ordre 2 d'une variable quantitative discrète est la valeur obtenue par l'une des deux procédures de calcul suivante :

- sur la population d'effectif N, nous la désignons par la lettre grecque σ^2

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{k=p} n_k (x_k - \mu)^2 = \left(\frac{1}{N}\sum_{k=1}^{k=p} n_k x_k^2\right) - \mu^2$$

- sur l'échantillon d'effectif n, nous la désignons par la lettre grecque σ_{ech}^2

$$\sigma_{ech}^2 = \frac{1}{n}\sum_{k=1}^{k=p} n_k (x_k - m)^2 = \left(\frac{1}{n}\sum_{k=1}^{k=p} n_k x_k^2\right) - m^2$$

La variance ou moment centré d'ordre 2 d'une variable quantitative continue est la valeur obtenue par l'une des deux procédures de calcul suivante où c_k sont les centres des intervalles :

- sur la population d'effectif N, nous la désignons par la lettre grecque σ^2

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{k=p} n_k (c_k - \mu)^2 = \left(\frac{1}{N}\sum_{k=1}^{k=p} n_k c_k^2\right) - \mu^2$$

- sur l'échantillon d'effectif n, nous la désignons par la lettre grecque σ_{ech}^2

$$\sigma_{ech}^2 = \frac{1}{n}\sum_{k=1}^{k=p} n_k (x_k - m)^2 = \left(\frac{1}{n}\sum_{k=1}^{k=p} n_k c_k^2\right) - m^2$$