Résumé Statistiques Descriptives / Inférentielles

BRADAI Zakaria

L2 Physique SPRINT

Table des matières

Chapitre I

Statistiques descriptives :

On parle de statistiques si on se restreint à l'observation et la description d'un phénomène, on parle de probabilités si on modélise des observations (i.e touver des lois mathématiques capables de générer les données que l'on observe). En statistiques descrptives on étudie des **individus** qui ont des caractéristiques que l'on appelle des **variables**, l'ensemble de tout les individus forme **la population** de taille N. En pratique on travaille sur des **échantillons** (de taille n), c-à-d une sélection d'individus parmi la population. On représente un échantillon via un data-set (tableau) où chaque ligne représente un individu et chaque colonne une variable.

Les 4 grands domaines des statistiques sont :

- *Les statistiques descriptives :* qui tentent de décrire et de résumer un jeu de données à l'aide de graphiques et de mesures sur une ou deux variables pas plus, on parle alors de statistiques descriptives **univariées** ou **bivariée**.
- *L'analyse multidimensionnelle* ou analyse exploratoire de données est le prolongement des statistiques descriptives à 3 variables ou plus.
- *Les statistiques inférentielles :* avec lesquelles on analyse un échantillon pour en déduire les caractéristiques globales de la population tout en contrôlant le risque lié à cette décision.
- *La modélisation statistique :* Il s'agit d'observer les caractéristiques d'un échantillon, puis de formaliser ces observations avec un modèle probabiliste àfin de la prédiction ou de la prévision.

En statistiques descriptives, on distingue 2 types de variables :

- **Variables quantitatives** dont les valeurs sont des nombres qui représentent une quantité, et dont les opérations arithmétiques sont possibles et ont un sens, elles peuvent être **discrêtes** ou **continues**.
- **Variables qualitatives** du nom des variables qui ne sont pas quantitatives, leur valeurs possibles sont des catégories que l'on appelle **modalités**. Une variable qualitative peut être **ordinale** (on peut ordonner les valeurs selon une certaine hiérarchie ou importance, par exemple les mentions au BAC) ou **nominale**.

Parmi les erreurs les plus récurrentes dans un jeu de données :

- Des valeurs manquantes. En pratique on supprime les ignes correspondantes s'il y'en a peu, ou au cotraire on supprime la variable entière s'il y'en a trop. On peut également assigner une valeur nous même, on parle d'**imputation** (par la moyenne par exemple).
- Des erreurs de formatage (ou erreurs lexicales), lorsqu'une valeur est incohérente par rapport au format ou par rapport à la façon dont la variable a été construite.
- Des doublons.
- Des valeurs extrêmes (ou *outiler*), c-à-d des valeurs trop importantes ou trop faibles par rapport à l'ensemble des valeurs d'une variable. On distingue cependant les outilers aberrants et atypiques. Il est d'usage de **citer les outilers dans ces analyses statistiques car la moyenne y est très sensible mais pas la médiane**.

Remarque : Cf. cours, représentation + graphiques...

A – Analyse univariée :

A.1 Notion de statisique :

Une **statistique** est un nombre calculé à partir d'une échantillon, on distingue une **mesure statistique** (ou **indicateur statistique**) qui est neutre (comme une moyenne par exemple) d'un **indice statistique** qui est une statistique construite à partir d'une certaine vision, à partir de connaissances d'un domaine.

A.2 Mesure d'une tendance centrale :

• **Moyenne empirique** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
• **Mode** x_m **:** modalité (pour une variable qualitative) ou valeur (pour une variable quantitative discrête) la plus fréquente.
• **Classe modale** $[x_m^-; x_m^+[$ **:** est la classe la plus fréquente.
La notion de mode en statistiques est différente de celle en probabilités.
Il arrive que l'on extrapole la définition du mode en l'assimilant au(x) pic(s) d'une distribution. Le mode n'est donc pas obligatoirement unique. Lorsqu'une distribution n'a qu'un seul pic, on parle de distribution **unimodale**. Il arrive aussi qu'une distribution présente deux ou plusieurs pics : on parle alors de distribution **bimodale** ou **plurimodale**.
• **Médiane :** pour n elle vaut $Med = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$, pour n impaire elle vaut $Med = x_{\left(\frac{n+1}{2}\right)}$.

A.3 Mesure d'une dispersion :

• **Variance empirique** (biaisée S^2 et non-biaisée S'^2).
• **Coefficient de variation :** $CV = \frac{S}{\bar{X}}$, son intérêt est de comparer moyenn et écart-type empiriques.
• **Quantile d'ordre** α **:** noté Q_α avec $\alpha \in [0;1]$, α (en %) des valeurs sont inférieures à Q_α et $1-\alpha$ (en %) y sont supérieures.
• **Boxplot :** elle utilise les quartiles (Q_1, $Q_2 = Med$ et Q_3), son intérêt est de fair figurer les outliers graphiquement. On appelle l'**écart inter-quartile** $IQ = Q_3 - Q_1$.
• **Ecart à la moyenne/médiane absolu :** $EMoyA = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{X}\right|$ ou $EMedA = \frac{1}{n}\sum_{i=1}^{n}|x_i - Med|$.

A.4 Mesures de forme :

• **Skewness empirique** γ_1 **:** Le skewness est **une mesure d'asymétrie**. L'asymétrie d'une distribution traduit la régularité ou non avec laquelle les observations se répartissent autour de la valeur centrale.

$$\boxed{\gamma_1 = \frac{\mu_3}{s^3}} \qquad / \quad \mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3$$

Si $\gamma_1 = 0$ la distribution est symétrique, si $\gamma_1 > 0$ elle est étalée à droite, enfin si $\gamma_1 < 0$ elle est étalée à gauche.
L'étude de l'asymétrie d'une distribution, c'est chercher qui de la médiane ou de la moyenne est la plus grande. Une distribution est dite symétrique si elle présente la même forme de part et d'autre du centre de la distribution, dans ce cas $x_m = Med = \bar{X}$. Une distribution est étalée à droite (ou oblique à gauche, ou présentant une asymétrie positive) si : $Mode < Med < \bar{X}$. De même, elle est étalée à gauche (ou oblique à droite) si : $Mode > Med > \bar{X}$.

• **Kurtosis empirique** γ_2 **:** Le kurtosis est une mesure d'**aplatissement**. L'aplatissement ne peut s'interpréter **que si la distribution est symétrique**.

$$\boxed{\gamma_2 = \frac{\mu_4}{s^4}} \qquad / \quad \mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^4$$

Si $\gamma_2 = 0$ la distribution **a le même aplatissement que la distribution normale.** Si $\gamma_2 > 0$ alors elle est moins aplatie que la distribution normale : les observations sont plus concentrées. Si $\gamma_2 < 0$ alors les observations sont moins concentrées : la distribution est plus aplatie.

A.5 Mesure de concentration :

Cf cours, courbe de Lorentz et indice de Gini.

B – Analyse bivariée :

L'analyse bivariée est une analyse menée entre deux variables.

Corrélation : dire que 2 variables sont corrélées signifie que si on connaît la valeur d'une variable, alors il est possible d'avoir *une indication* (plus ou moins précise) sur la valeur d'une autre variable. Mathématiquement l'étude d'une corrélation entre 2 variables revient à étudier la dépendance qu'il existerait entre les 2 événements ayatn généré ces variables.
Une erreur à ne **JAMAIS** commetre est de dire qu'il y a un lien de cause à effet lorsqu'on parle de corrélation. Il est possible que 2 variables soient corrélés sans qu'il n'y ait aucun lien entre elles, on parle de **corrélation fallacieuses** ou fortuites (spurious correlations).

Cf cours, **tableau de contingence.**

- Chaque valeur du tableau de contingence (hors de la ligne et de la colonne TOTAL) est appelée**effectif conjoint** n_{ij}.
- L'ensemble des effectifs conjoints est appelé **distribution conjointe empirique.**
- La dernière ligne et colonne (TOTAL) est appelée **distribution marginale empirique**.
- L'ensemble des effectifs conjoints d'une ligne / colonne est appelée **distribution conditionnelle empirique de** X/Y **étant donné que** $Y = ./X = .$ (Cf exemple cours).

Le **diagramme de dispersion** (*scatter plot*) est adapté pour visualiser la corrélation entre 2 variables quantitatives, on peut aussi découper les variables en classes et utiliser un plot de plusieurs *box plots*. Alors que le **tableau de contingence** est adapté pour visualiser la corrélation entre 2 variables qualitatives.

Corrélation empirique (estimateur sans biais) : $s(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ Propriétés de la covariance :

- **symétrie :** $s(X,Y) = s(Y,X)$
- **bilinéarité :** $Z = aU + bV \implies s(X,Z) = as(X,U) + bs(X,V)$

Coefficient de corrélation linéaire / de Pearson : "détecte les relations linéraires".

$$r_{X,Y} = \frac{s(X,Y)}{s_X, s_Y}$$

Chapitre II

Statistiques Inférentielles :

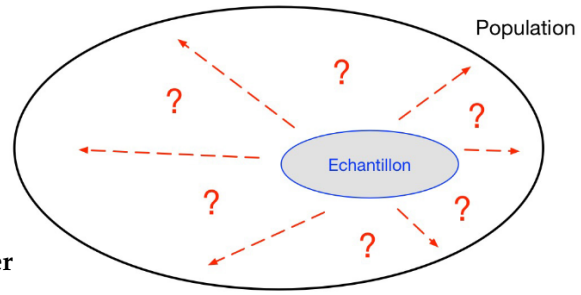
A – Introduction aux statistiques inférentielles :

On pose 2 problématiques :

- **Cas discret :** Un laboratoire dispose de 216 résultats de test d'un nouveau médicament, on se pose les questions suivantes :
  - — Quel est le taux de guérision théorique ?
  - — Comment estimer une fourchette du taux de guérison ?
  - — Ce nouveau médicament est-il significativement meuilleur que l'ancien qui avait un taux de guérison de 76% ?
- **Cas continu :** Un constructeur de voiture a mis en vente un nouveau moteur thermique, on se pose les questions suivantes :
  - — Quelle est la consommation théorique d'essence de ce moteur ?
  - — Comment estimer une fourchette de cette consommation ?
  - — Peut-on affirmer que la consommation est de 37 litres aux 1000 ?

Ainsi, nous avons abordé les 3 principaux thèmes des statistiques inférentielles à savoir :

* **L'estimation (ponctuelle).**
* **Les intervalles de confiance.**
* **Les test statistiques d'adéquation.**



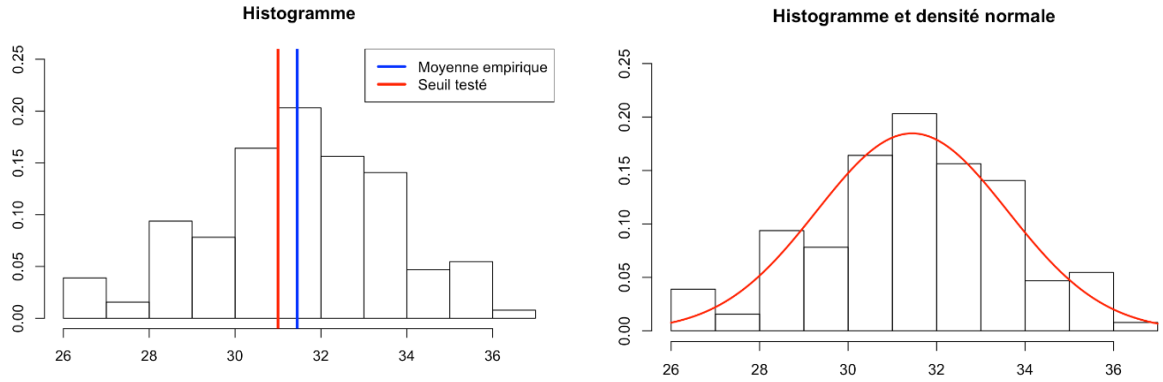
Pour rappel en *statistiques descriptives*, on souhaite décrire un échantillon mais il est impossible d'**inférer/extrapoler** ce qui est constaté sur un échantillon à la population statistique tout entière.

En d'autres termes, l'**inférence** consiste à conclure sur la population entière à partir d'un échantillon tout en controlant le risque lié à cette décision via un modèle probabiliste. On considère dans ce cours que les échantillons $(x_1, \dots, x_n)$ sont les réalisations de variables aléatoires $X_1, \dots, X_n$ **iid** pour **indépendantes et identiquement distribuées**, c-à-d **de même loi** (c-à-d que $\forall i \in [\![1; n]\!],\ x_i$ est le fruit d'un tirage aléatoire de X_i).

L'une des difficultés de l'inférence est la variabilité qui se divise 2 types :

- **Variabilité intrinsèque** (du phénomène), pour reprendre les exemples précédents les patiens n'ont pas le mêm patrimoine génétique et les voitures n'ont pas toutes roulées sur le même terrain.
- **Variabilité due à l'échantillonage**, on obtiendra sans nul doute des résultats différents (pas radicalement opposés) sur d'autres échantillons.

Pour le cas discret précédent, le modèle probabiliste adapté est une loi de Bernouillo $\mathcal{B}(p)$ avec p le taux de guérison ou la proportion de patients guéris, pour le cas continu l'hypothèse probabiliste n'est pas certaine, on peut supposer $X_i \hookrightarrow \mathcal{N}(\mu, \sigma^2)$. On pourra vérifier cette hypothèse par une vérification théorique via un test d'adéquation statistique, ou par une vérification pratique en regardant l'adéquation potentielle de la distribution empirique avec la loi utilisée.



Vérification pratique : on constate que l'écart entre l'histogramme, **primo estimation de la densité de probabilité**, et la densité gaussienne standard sont plutôt proches (ici quasiment superposables).

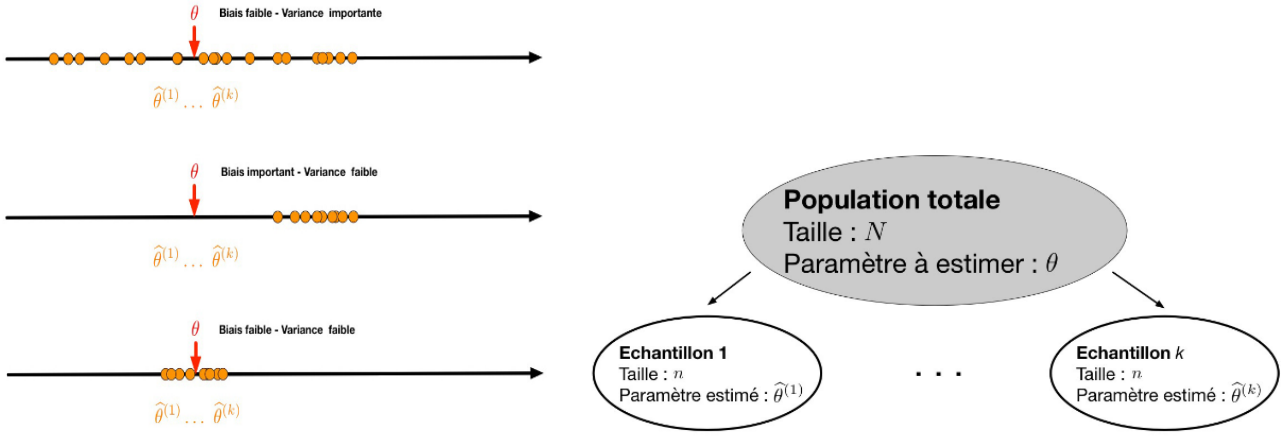
2) Réalisation d'une estimation ponctuelles :

Paramètre θ : grandeurs dont la conaissance seule revient à connaitre la loi de probabilité entière.
Dans l'exemple discret $\theta = p$ et dans l'exemple continu $\theta = (\mu, \sigma^2)$.

Estimateur $\hat{\theta} = f(X_1, \ldots, X_n)$: fonction des observations **qui prend ses valeurs dans le domaine de définition du/des paramètres** θ.

La qualité d'un estimateur réside dans les propriétés qu'on attend, à savoir :

- **L'exhaustivité :** c-à-d si l'estimateur capte toute l'information contenue dans l'échantillon pour estimer le paramètre θ.
- **Consistance/Convergence :** $\hat{\theta} \underset{n \to +\infty}{\to} \theta_{\text{réel}}$, c-à-d que l'estimateur se rapproche asymptotiquement de la vraie valeur de paramètre et que sa qualité est meuilleur pour n grand.
- **Biais et variance faibles :** $Biais(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$, le biais d'un estimateur est la différence entre la moyenne empirique que l'on aurait pû obtenir sur plusieurs échantillons et la vraie valeur du paramètre θ, la variance est la dispertion des des estimateurs obtenus $\hat{\theta}^{(i)}$ sur chacun des échantillons. On souhaiterait dans l'absolu qu'un estimateur soit sans biais (du moins pour des grands échantillons) et de variance faible.
- **Risque quadratique faible :** $MSE(\hat{\theta}, \theta) = Var(\hat{\theta}) + \Big(Biais(\hat{\theta}, \theta)\Big)^2$



La traduction mathématique des propriétés précédentes est :

- **L'exhaustivité :** "La loi de probabilités de l'échantillon $(X_1, \ldots, X_n)$ conditionnellement à lestimateur considéré $\hat{\theta}$ est indépendante du paramètre θ.
- **Consistance/Convergence :** signifie la **convergence en probabilités** :

$$\hat{\theta} \xrightarrow{\mathbb{P}} \theta \quad \iff \quad \forall \varepsilon > 0, \lim_{n \to +\infty} \mathbb{P}\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0$$

  L'estimateur $\hat{\theta}$ est dit **fortement convergent** s'il **converge presque sûrement vers** θ i.e :

$$\mathbb{P}\left(\lim_{n \to +\infty} \hat{\theta} = \theta\right) = 1$$

  La convergence presque-sûre $\longrightarrow$ la convergence en probabilités, en pratique on utilise la loi forte des grands nombres (ref Annexe???) pour la démontrer.
- **Biais/Variance :** $\hat{\theta}$ est une variable aléatoire.
  - $\hat{\theta}$ est **sans biais** : $\mathbb{E}[\hat{\theta}] = \theta$
  - $\hat{\theta}$ est **asymptotiquement sans biais** : $\mathbb{E}[\hat{\theta}] \underset{n \to +\infty}{\to} \theta$
- **Risque quadratique :** $R(\hat{\theta}, \theta) = Var(\hat{\theta}) + \left(Biais(\hat{\theta})\right)^2 = \mathbb{E}[\left(\hat{\theta} - \theta\right)^2]$
  - estimateur **sans biais** : $R(\hat{\theta}, \theta) = Var(\hat{\theta}, \theta)$
  - estimateur **asymptotiquement sans biais** : $R(\hat{\theta}, \theta) \underset{n \to +\infty}{\to} Var(\hat{\theta}, \theta)$

A.1 Estimation d'une proportion :

En reprenant l'exemple discret, intuitivement la fréquence empirique des guérisons sur l'échantillon parraît correspondre au vrai taux de guérison.

Plus formellement, la **fréquence empirique** $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ (moyenne empirique de 0 et de 1) est un estimateur **consistant** et **sans biais** de la proportion p.

A.2 Estimation d'une moyenne :

La **moyenne empirique** $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ est un estimateur **consistant** et **sans biais** de la moyenne μ.

A.3 Estimation d'une variance :

La **variance empirique non biaisée** $\widehat{\sigma^2} = \widehat{\sigma^2}_{\text{non biaisée}} = S^{2\prime} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})$ est un estimateur **consistant** et **sans biais** de la variance σ^2.

La **variance empirique biaisée** $\widehat{\sigma^2} = \widehat{\sigma^2}_{\text{biaisée}} = S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})$ est un estimateur **consistant** et **asymptotiquement sans biais** de la variance σ^2. De ce fait on estimera une variance avec S'^2.

A.4 Méthode des moments :

Cette méthode consiste à estimer les paramètres recherchés θ en égalisant certains moments ordinaires théoriques avec leurs estimations empiriques ce qui donne un estimateur **consistant**, l'égalisation est justifiée par la loi des grands nombres.
Pour rappel l'expression du moment ordinaire m_k d'ordre k d'una variable aléatoire X est :

$$m_k = \mathbb{E}[X^k] = \int X^k f(X) \mathrm{d}X$$

On pose $G = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}$ que l'on exprime en fonction des paramètres $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$.

$$G = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix} = \begin{pmatrix} f_1(\theta_1, \ldots, \theta_n) \\ \vdots \\ f_n(\theta_1, \ldots, \theta_n) \end{pmatrix} \Longrightarrow \theta = \begin{pmatrix} {f_1}^{-1}(m_1, \ldots, m_n) \\ \vdots \\ {f_n}^{-1}(m_1, \ldots, m_n) \end{pmatrix}$$

Enfin en vertu de la loi faible des grands nombres, on remplace les moments ordinaires théoriques par leurs estimations empiriques $m_k \longleftarrow \widehat{m_k} = \frac{1}{n} \sum_{i=1}^{n} {x_i}^k$.

A.5 Méthode du maximum de vraisemblance :

• **Cas discret :** On définit la vraisemblance d'une échatillon $(x_1, \ldots, x_n)$ de paramètres θ comme suit :

$$\mathcal{L}(\theta | x_1, \ldots, x_n) = \mathbb{P}(x_1, \ldots, x_n; \theta)$$

$$\mathcal{L}(\theta | x_1, \ldots, x_n) = \mathbb{P}(x_1, \ldots, x_n; \theta) \overset{\text{indépendance}}{=} \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \overset{\text{même loi}}{=} \prod_{i=1}^{n} \mathbb{P}(X = x_i) = \prod_{i=1}^{n} \mathbb{P}(x_i)$$

$$\boxed{\mathcal{L}(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathbb{P}(X = x_i) = \prod_{i=1}^{n} \mathbb{P}(x_i)}$$

• **Cas continu :** $f(x; \theta)$ est la densité de probabilité du modèle probabiliste choisi.

$$\mathcal{L}(\theta | x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta) \overset{\text{indépendance}}{=} \prod_{i=1}^{n} f(X_i = x_i) \overset{\text{même loi}}{=} \prod_{i=1}^{n} f(X = x_i) = \prod_{i=1}^{n} f(x_i)$$

$$\boxed{\mathcal{L}(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(X = x_i) = \prod_{i=1}^{n} f(x_i)}$$

La fonction de vraisemblance $\mathcal{L}$ mesure la probabilité que les observations proviennent effectivement d'un échantillon de loi paramétrée par θ, trouver le maximim de vraisemblance revient à **trouver le paramètre le plus vraisemblable pour notre échantillon.**

L'intérêt de cette méthode, au-delà de trouver un estimateur dans un cas moins intuitif, est de trouver un estimateur qui est muni d'excellentes qualités dans de très nombreux cas : **consistant, asymptotiquement sans biais et de variance minimale.** La plupart du temps, cet estimateur est **unique et se détermine explicitement.**

Remarque : Cf Wiki intérêt et limitations des méthodes.

B – Déterminer des intervalles de confiance :

On dit qu'un intervalle de confiance I est **bilatère** s'il encadre le paramètre θ à gauche et à droite, il est dit **unilatère** dans le cas contraire. On appelle α le **risque** et $1-\alpha$ le **niveau de confiance** de l'intervalle I.

$$\mathbb{P}\left(IC^{-}(X_1,\ldots,X_n) \leq \theta \leq IC^{+}(X_1,\ldots,X_n)\right)$$

On voit qu'il y a un compromis largeur-niveau de confiance (quand la largeur diminue, le niveau de confiance diminue également). En pratique on cherche un niveau de confiance suffisamment grand avec un intervalle pas trop large. Classiquement on cherche des niveaux de confiance de 90% ou 95% ($\alpha = 10\%$ ou $\alpha = 5\%$).

On notera par la suite I_α **l'intervalle de confiance bilatère de niveaud de confiance** $1-\alpha$**.**

B.1 Intervalle de confiance sur une proportion :

$$I_\alpha = \left[\bar{X} - \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\ ;\ \bar{X} + \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right]$$

$\bar{X}$ étant la fréquence empirique, $\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$ un estimateur de la variance d'une loi de Bernoulli. On remarque que plus n la taille de l'échantillon augmente, plus la largeur de I_α pour un α fixé diminue, plus on est confiant. Enfin $\phi_{1-\frac{\alpha}{2}}$ est le **quantile d'ordre** $1-\dfrac{\alpha}{2}$ **d'une loi binomiale**. En pratique on l'approxime par le quantile d'une loi normale pour des échantillons de grande taille ($n \geq 30$) ce qui facilite les calculs car on utilise une loi continue, dans le cadre de cette approximation on parle d'**intervalle de confiance asymptotique**.

B.2 Intervalle de confiance sur une moyenne :

$$I_\alpha = \left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S'}{\sqrt{n}}\ ;\ \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S'}{\sqrt{n}}\right]$$

Où $t_{n-1,1-\frac{\alpha}{2}}$ est le **quantile d'ordre** $1-\dfrac{\alpha}{2}$ **d'une loi de Student** $T(n-1)$ **à** $n-1$ **degrés de liberté**.

On considère un échantillon **iid** de loi normale $\mathcal{N}(\mu,\sigma^2)$, ou **un grand échantillon (en pratique** $n \geq 30$**) iid non gaussien** (en vertu du thérème central limite).

B.3 Intervalle de confiance sur une variance :

$$I_\alpha = \left[\frac{(n-1)S'^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\ ;\ \frac{(n-1)S'^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

Où $\chi^2_{n-1,\frac{\alpha}{2}}$ est le **quantile d'ordre** $\dfrac{\alpha}{2}$ **d'une loi du** $\chi^2(n-1)$ **à** $n-1$ **degrés de liberté** (La loi du χ^2 est asymétrique).

Remarque : annexe lois de Student, de Fisher et du χ^2 + démonstration CF IPAD NOTES ...

C – Test statistiques :

On parle ausse de **test paramétriques** car il pré-supposent que l'échantillon à étudier suit une certaine loi décrite par des paramètres θ.

Reprenons l'exemple discret :

$$\begin{cases} H_0 : & p & = & 0.76 & \text{(hypothèse nulle)} \\ H_1 : & p & \neq & 0.76 & \text{(hypothèse alternative)} \end{cases}$$

Si le rejet de H_0 est fait à mauvias escient il sera considéré comme le plus coûteux pour le laboratoire (répercussion néfastes).
L'hypthèse alternative H_1 indique dans quelles conditions on rejette H_0, elle fournit les informations de forme de la **région critique**.
région critique W **:** région pour laquelle on rejette H_0, elle est basée sur la statistique de test. Pour déterminer entièrement W, il faut *connaître la loi (éventuellement asymptotique) de la statistique de test sous* H_0 .

C.1 Risques associés aux décisions :

- **Risque de première espèce :** risque de rejeter H_0 alors qu'elle est vraie, c'est le risque dont on veut se prémunir en premier lieu.
  On fixe un **niveau de test** qui représente le **risque de première espèce maximale** que l'on note α (usuellement $\alpha = 5\%$ ou $\alpha = 10\%$).

- **Risque de seconde espèce :** risque de non rejet de H_0 alors qu'elle est fausse, il faut minimiser β et donc maximiser $1 - \beta$.

		Vérité	
		H_0 vraie	H_0 fausse
Décision	Rejet de H_0	α	1-β
	Non rejet de H_0	1-α	β

On représente en vert les bonnes décision et en rouge les mauvaises. α correspond donc à **la probabilité de faire un erreur de type** I **en supposant** H_0 **vraie**, et β correspond à **la probabilité de faire un erreur de type** II **en supposant** H_0 **fausse.**

C.2 Risque et $p-$value :

Lors de la réalisation des tests statistiques, on sera amené à calculer pour chaque valeur du risque α un quatile $\phi = f(\alpha)$. Pour éviter de re-calculer cette quantité à chaque fois que l'on change la valeur du risque α on emploi la méthode de la $p-$value qui revient à calculer une probabilité et dont le principe est la suivant :

$$p - \text{value} < \alpha \implies H_0 \textbf{ est rejetée}$$

La $p-$value correspond à **la probabilité d'obtenir le résultat observé si** H_0 **est vraie.**
Elle est d'autant plus sensible que la taille n de l'échantillon augmente.

C.3 Types de tests :

• Test bilatère :

$$\begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \neq & p_0 \end{cases}$$

• **Test unilatère droite et gauche :**

$$\begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \geq & p_0 \end{cases} \qquad \begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \leq & p_0 \end{cases}$$

C.4 Tester une proportion :

$$\begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \neq & p_0 \end{cases} \qquad \begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \geq & p_0 \end{cases} \qquad \begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \leq & p_0 \end{cases}$$

— Statistique de test : $Z = \sqrt{n}\left(\dfrac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\right)$

— Loi de la statistique de test : **convergence asymptotique** $Z \hookrightarrow \mathcal{N}(0,1)$

— Région critique :

  — test bilatère : $W = \left\{|Z| > \phi_{1-\frac{\alpha}{2}}\right\} = \left\{Z < \phi_{\frac{\alpha}{2}} \text{ ou } Z > \phi_{1-\frac{\alpha}{2}}\right\}$ [1]

  — test unilatère à droite : $W = \{Z < \phi_\alpha\}$

  — test unilatère à gauche : $W = \{Z > \phi_{1-\alpha}\}$

— $p-$value :

  — test bilatère : $p - \text{value} = 2\int_{x=|Z|}^{+\infty} \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{x^2}{2}\right) \mathrm{d}x = \int_{x=|Z|}^{+\infty} \sqrt{\dfrac{2}{\pi}} \exp\left(-\dfrac{x^2}{2}\right) \mathrm{d}x$

  — test unilatère à droite : $p - \text{value} = \int_{-\infty}^{Z} \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{x^2}{2}\right) \mathrm{d}x$

  — test unilatère à gauche : $p - \text{value} = \int_{Z}^{+\infty} \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{x^2}{2}\right) \mathrm{d}x$

C.5 Tester l'égalité de 2 proportions :

$$\begin{cases} H_0 : & p_1 & = & p_2 \\ H_1 : & p_1 & \neq & p_2 \end{cases} \qquad \begin{cases} H_0 : & p_1 & = & p_2 \\ H_1 : & p_1 & \geq & p_2 \end{cases} \qquad \begin{cases} H_0 : & p_1 & = & p_2 \\ H_1 : & p_1 & \leq & p_2 \end{cases}$$

— Statistique de test : $Z = \dfrac{\bar{X_1} - \bar{X_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{n_1+n_2}{n_1 n_2}\right)}}$ / $\widehat{p} = \dfrac{\bar{X_1} + \bar{X_2}}{n_1 + n_2}$

— Loi de la statistique de test : $Z \hookrightarrow \mathcal{N}(0,1)$.

— Région critique :

  — test bilatère : $W = \left\{|Z| > \phi_{1-\frac{\alpha}{2}}\right\} = \left\{Z < \phi_{\frac{\alpha}{2}} \text{ ou } Z > \phi_{1-\frac{\alpha}{2}}\right\}$

  — test unilatère à droite / gauche : $W_{\text{droite}} = \{Z < \phi_\alpha\}$; $W_{\text{gauche}} = \{Z > \phi_{1-\alpha}\}$

— $p-$ value :

  — test bilatère : $p - \text{value} = 2\int_{x=|Z|}^{+\infty} \mathcal{N}(x,0,1)\mathrm{d}x$

  — test unilatère à droite : $p - \text{value} = \int_{-\infty}^{Z} \mathcal{N}(x,0,1)\mathrm{d}x$

  — test unilatère à gauche : $p - \text{value} = \int_{x=Z}^{+\infty} \mathcal{N}(x,0,1)\mathrm{d}x$

C.6 Tester une moyenne théorique :

$$\begin{cases} H_0 : & \mu & = & \mu_0 \\ H_1 : & \mu & \neq & \mu_0 \end{cases} \qquad \begin{cases} H_0 : & \mu & = & \mu_0 \\ H_1 : & \mu & \geq & \mu_0 \end{cases} \qquad \begin{cases} H_0 : & \mu & = & \mu_0 \\ H_1 : & \mu & \leq & \mu_0 \end{cases}$$

— Statistique de test : $T = \sqrt{n}\left(\dfrac{\bar{X} - \mu_0}{S'}\right)$ / $S' = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}$

— Loi de a statistique de test : $T \hookrightarrow \mathcal{T}(n-1)$ (loi de Student à $n-1$ degrés de liberté).

— Région critique :

1. Pour une distribution symétrique, $\phi_{\frac{\alpha}{2}} = -\phi_{1-\frac{\alpha}{2}}$

— test bilatère : $W = \left\{|T| > t_{n-1,1-\frac{\alpha}{2}}\right\} = \left\{T < t_{n-1,\frac{\alpha}{2}} \text{ ou } T > t_{n-1,1-\frac{\alpha}{2}}\right\}$

— test unilatère à droite : $W = \{T < t_{n-1,\alpha}\}$

— test unilatère à gauche : $W = \{T > t_{n-1,1-\alpha}\}$

— $p-$value :

— test bilatère : $p - \text{value} = 2\int_{x=|T|}^{+\infty} \mathcal{T}(x, n-1)\,\mathrm{d}x$

— test unilatère à droite : $p - \text{value} = \int_{x=-\infty}^{T} \mathcal{T}(x, n-1)\,\mathrm{d}x$

— test unilatère à gauche : $p - \text{value} = \int_{x=T}^{+\infty} \mathcal{T}(x, n-1)\,\mathrm{d}x$

C.7 Tester une variance théorique :

$$\left\{\begin{array}{lcl} H_0 : & \sigma^2 & = \sigma_0^2 \\ H_1 : & \sigma^2 & \neq \sigma_0^2 \end{array}\right. \qquad \left\{\begin{array}{lcl} H_0 : & \sigma^2 & = \sigma_0^2 \\ H_1 : & \sigma^2 & > \sigma_0^2 \end{array}\right. \qquad \left\{\begin{array}{lcl} H_0 : & \sigma^2 & = \sigma_0^2 \\ H_1 : & \sigma^2 & < \sigma_0^2 \end{array}\right.$$

— Statistique de test : $K = (n-1)\left(\dfrac{S'^2}{\sigma_0^2}\right)$

— Loi de la statistique de test : $K \hookrightarrow \chi^2(n-1)$.

— Région critique :

— test bilatère : $W = \left\{K < \chi^2_{n-1,\frac{\alpha}{2}} \text{ ou } K > \chi^2_{n-1,1-\frac{\alpha}{2}}\right\}$ [2]

— test unilatère à droite / gauche : $W_{\text{droite}} == \left\{K < \chi^2_{n-1,\alpha}\right\} \quad ; \quad W_{\text{gauche}} = \left\{K > \chi^2_{n-1,1-\alpha}\right\}$

— $p-$value :

— test bilatère : $p - \text{value} = 2\min\left(\int_{x=0}^{K} \chi^2(x, n-1)\mathrm{d}x\ ,\ \int_{x=K}^{+\infty} \chi^2(x, n-1)\mathrm{d}x\right)$

— test unilatère à droite : $p - \text{value} = \int_{x=0}^{K} \chi^2(x, n-1)\mathrm{d}x$

— test unilatère à gauche : $\int_{x=K}^{+\infty} \chi^2(x, n-1)\mathrm{d}x$

C.8 Comparaison de 2 échantillons gaussiens iid :

On réaslie d'abord un test d'égalité des variances, puis s'il s'avère concluant (si on ne rejette pas $H_0^{(1)}$) on réalise un test d'égalité des moyennes.

• **Test d'égalité des variances :**

$$\left\{\begin{array}{lcl} H_0^{(1)} : & \sigma_1^2 & = \sigma_2^2 \\ H_1^{(1)} : & \sigma_1^2 & \neq \sigma_2^2 \end{array}\right.$$

— Statistique de test : $F = \dfrac{S_1'^2}{S_2'^2}$

— Loi de la statistique de test : $F \hookrightarrow \mathcal{F}(n_1 - 1, n_2 - 1)$ (loi de Fisher).
Intuitivement on s'attend à avoir $F \approx 1$).

— Région critique : $W = \left\{F < f_{n_1-1,n_2-1,\frac{\alpha}{2}} \text{ ou } F > f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right\}$

— $p - \text{value} = 2\min\left(\int_{x=0}^{F} \mathcal{F}(x, n_1-1, n_2-1)\mathrm{d}x\ ;\ \int_{x=F}^{+\infty} \mathcal{F}(x, n_1-1, n_2-1)\mathrm{d}x\right)$

• **Test d'égalité des moyennes :** (à faire uniquement si on ne rejette pas $H_0^{(1)}$) :

$$\left\{\begin{array}{lcl} H_0^{(2)} : & \mu_1 & = \mu_2 \\ H_1^{(2)} : & \mu_1 & \neq \mu_2 \end{array}\right.$$

2. La loi du χ^2 contrairement aux lois normales et de Student, n'est pas symétrique, tout comme la loi de Fisher.

— Statistique de test : $T = \dfrac{\sqrt{n_1+n_2-2}(\bar{X}_1+\bar{X}_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{(n_1-1){S'_1}^2+(n_2-1){S'_2}^2}}$

— Loi de la statistique de test : $T \hookrightarrow \mathcal{T}(n_1+n_2-2)$

— Région critique : $W = \left\{|T| > t_{n_1+n_2-2,1-\frac{\alpha}{2}}\right\} = \left\{T < t_{n_1+n_2-2,\frac{\alpha}{2}} \text{ ou } T > t_{n_1+n_2-2,1-\frac{\alpha}{2}}\right\}$

— $p-$ value $= 2\int_{x=|T|}^{+\infty} \mathcal{T}(x, n_1+n_2-2)\mathrm{d}x$

Conclusion : enfin, au début d'un test statistique on pose les hypothèses H_0 et H_1 et on indique le risque α, et on conclut par les phrases "H_0 est rejetée avec un risque de $\alpha(\%)$" ou "H_0 n'est pas rejetée avec un risque de $\alpha(\%)$".

Dans l'exemple du problème du moteur, nous avons supposé que les observations $(x_1, \ldots, x_n)$ sont ma réalisation d'un variable aléatoire $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$. En pratique il faut d'abord faire un test d'adéquation pour qu'on suite on sera plus certain dans les test qu'on applique.

C.9 Test de Welch - égalité des moyennes avec inégalité des variances :

Si l'on souhaite tester l'égalité des moyennes de 2 échantillons **gaussiens** dont **les variances sont inégales et inconnues**, on réalise le test t dit de Welch.

$$\begin{cases} H_0 : & \mu_1 & = & \mu_2 \\ H_1 : & \mu_1 & \neq & \mu_2 \end{cases}$$

— Statistique de test : $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{s'_1}^2}{n_1}+\frac{{s'_2}^2}{n_2}}} \quad / \quad s'^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

— Loi de la statistique de test : convergence asymptotique $T \hookrightarrow \mathcal{T}(\nu)$ avec :

$$\nu = \frac{\left(\frac{{s'_1}^2}{n_1}+\frac{{s'_2}^2}{n_2}\right)^2}{\frac{{s'_1}^4}{n_1^2(n_1-1)}+\frac{{s'_2}^4}{n_2^2(n_2-1)}}$$

— Région critique : $W = \left\{|T| > t_{\nu,1-\frac{\alpha}{2}}\right\}$

— $p-$ value $= \int_{x=|T|}^{+\infty} \mathcal{T}(x, \nu)\mathrm{d}x$

C.10 Test de Kolmogorov-Smirnov - cas continu :

Ce test n'est pas très puissant en général, on privilégiera le test de Shapiro-Wilk.
On suppose $X \hookrightarrow \mathcal{F}_0$ avec $\mathcal{F}_0$ une loi continue quelconque, on note $\mathcal{F}$ la loi continu de X à valeurs dans $\mathbb{R}$.

On note D_n la statistique de test appelée la statistique de Kolmogorov.

$$D_n = \sup_{x\in\mathbb{R}} |F_{emp}(x) - F_0(x)|$$

Avec F_{emp} la fonction de répartition empirique et F_0 la fonction de répartition théorique. F_{emp} est un estimateur de F_0.
Plus la statistique de Kolmogorov D_n est grande, plus on est amène à rejeter H_0.

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \frac{c}{\sqrt{n}}\right] \underset{n\to+\infty}{\to} \alpha(c) = 2\sum_{r=1}^{+\infty}(-1)^{r-1}\exp(-2c^2r^2) \quad \forall c > 0$$

Le terme $\alpha(c)$ vaut 0.05 pour $c = 1.36$.

C.11 Test de Shapiro-Wilk - cas continu :

Il s'agit du test d'adéquation à une loi normale le plus utilisé aujourd'hui sur des échantillons **iid** de petites tailles.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$x_{(i)}$: $i$ème statistique d'ordre i.e $i$ème plus petit nombre de l'échantillon.

$m = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}$ avec m_i l'espérence de x_i, V : matrice de variance-covariance de l'échantillon.

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

C.12 Test d'adéquation du χ^2 - cas discret :

On suppose $\forall i \in [\![1; n]\!] X_i \hookrightarrow \mathcal{D}_i$ c-à-d que l'on **dispose de** n **variables aléatoires discrètes**.

On dipose des n réalisations de variables aléatoires discrètes $x_1, \ldots, x_n$ et de k valeurs distinctes $a_1, \ldots, a_k$. On note p_i la probabilité ou la fréquence empirique de la valeur a_i.

On doit vérifier l'adéquation des probabilités ou fréquences observées $p = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}$ et théoriques $p_0 = \begin{pmatrix} p_{0,1} \\ \vdots \\ p_{0,k} \end{pmatrix}$.

$$\begin{cases} H_0 : & p & = & p_0 \\ H_1 : & p & \neq & p_0 \end{cases}$$

On construit une *pseudo-distance* nommée χ^2 et notée ξ_n :

$$\xi_n = n \sum_{i=1}^{k} \frac{(p_i - p_{0,i})^2}{p_{0,i}}$$

On peut montrer que $\xi_n \hookrightarrow \chi^2(k-1)$.
On peut donc faire une $p-$value, on fixe un risque α et on rejette H_0 si $\xi_n > \chi^2_{k-1,1-\alpha}$
Région critique : $W = \left\{\xi_n > \chi^2_{k-1,1-\alpha}\right\}$

$p - \text{value} = \int_{x=K}^{+\infty} \chi^2(x, n-1)\mathrm{d}x$

C.13 Test d'adéquation du χ^2 - cas continu :

Le théorème de Gauss-Markov énonce que parmi les estimateurs linéaires non -biaisés, l'estimateur par moindres carrés présente une variance minimale.

On mesure 2 grandeurs $X = (x_1, \ldots, x_n)$ et $Y = (y_1, \ldots, y_n)$ ainsi que leurs incertitudes associées $\sigma_X = (\sigma_{x_1}, \ldots, \sigma_{x_n})$ et $\sigma_Y = (\sigma_{y_1}, \ldots, \sigma_{y_n})$. On cherche à montrer la relation $Y = f(X, \theta)$ avec f ayant p paramètres.

$$\begin{cases} H_0 : & Y & = & f(X, \theta) \\ H_1 : & Y & \neq & f(X, \theta) \end{cases}$$

Statistique de test : $\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i, \theta)}{\sigma_i}\right)^2$ avec : $\sigma_i = \sqrt{\sigma_{y_i}^2 + \left(\left.\frac{\partial f}{\partial x}\right|_{x_i} . \sigma_{x_i}\right)^2}$

Loi de la statistique de test : $\chi^2(\theta) \hookrightarrow \chi^2(n-p)$
Région critique : $W = \left\{\chi^2(\theta) < \chi^2_{1-\alpha}\right\}$

$p - \text{value} = \int_{x=\chi^2(\theta)}^{+\infty} \chi^2(x, n-p)\mathrm{d}x$ – Estimateur rapide du *Goodness of fit :* $\frac{\chi^2(\theta)}{n-p} \approx 1$