Année universitaire 2019-2020

STATISTIQUES INFÉRENTIELLES
L'essentiel de la première année

Semestre 7

Auteur : Florent ARNAL
Adresse électronique : florent.arnal@u-bordeaux.fr
Site : http://flarnal.e-monsite.com

Table des matières

Chapitre I

Généralités autour des variables d'échantillonnage

Définition 1 : On considère $X_1, X_2, \cdots, X_n$ des variables indépendantes et identiquement distribuées (i.i.d).
On définit les variables $\bar{X}, \hat{S}^2, F$ appelées variables d'échantillonage associées respectivement à la moyenne, la variance et la proportion par

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

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{SCE}{n-1}$$

$$F = \frac{X}{n}$$

Propriété 1 : Distribution de $\bar{X}$ (Cas de distributions normales, écart-type connu)
Si les variables $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu;\sigma)$ alors

$$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Propriété 2 : Distribution de $\bar{X}$ (Application du TCL dans de distributions quelconques, écart-type connu)
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées avec $n \geqslant 30$, $\mathbb{E}(X_i) = \mu$ et $\mathbb{V}(X_i) = \sigma^2$ alors
la loi de $\bar{X}$ converge vers $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.

Ce propriété découle du TCL qui affirme que $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0;1)$. En divisant par n, il vient $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ ce qui permet d'affirmer que

$$\boxed{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0;1)}$$

Propriété 3 : Distribution associée à $\bar{X}$ (Écart-type de la population inconnu)
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu;\sigma)$ alors

$$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$$

PROPRIÉTÉ 4 : Distribution associée à la variance
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu; \sigma)$ alors

$$\frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{SCE}{\sigma^2} \sim \chi^2(n-1)$$

PROPRIÉTÉ 5 : Considérons une certaine population dans laquelle la proportion d'individus ayant une propriété donnée est égale à p.

- $X \sim \mathcal{B}(n, p)$;
- Pour n suffisamment grand, la loi de X peut être approchée par la loi $\mathcal{N}\left(np; \sqrt{np(1-p)}\right)$;
- Pour n suffisamment grand, la loi de $F = \dfrac{X}{n}$ peut être approchée par la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$.

Chapitre II

Estimation

II.A Estimation ponctuelle

Définition 2 : Soit θ un paramètre d'un modèle.
Une variable Y est un estimateur sans biais de θ si

$$\mathbb{E}(Y) = \theta$$

Une estimation ponctuelle de ce paramètre θ, notée $\widehat{\theta}$, est donnée par la réalisation y de la variable Y sur un échantillon donné. On a donc :

$$\widehat{\theta} = y$$

Propriété 6 : Estimation ponctuelle d'une moyenne
- $\bar{X}$ est un estimateur non biaisé de μ.
- Une estimation ponctuelle $\widehat{\mu}$ de la moyenne μ d'une population est donnée par

$$\widehat{\mu} = \bar{x}$$

En effet : $\mathbb{E}(\bar{X}) = \mu$.

Propriété 7 : Estimation ponctuelle d'une variance
- $\dfrac{SCE}{n-1}$ est un estimateur non biaisé de σ^2.
- Une estimation ponctuelle $\widehat{\sigma}^2$ de la variance σ^2 d'une population est donnée par

$$\widehat{\sigma}^2 = \frac{SCE}{n-1} = \hat{s}^2$$

En effet : $\dfrac{SCE}{\sigma^2} \sim \chi^2(n-1)$ et $\mathbb{E}(\chi^2(n-1)) = n-1$ donc $\mathbb{E}\left(\dfrac{SCE}{\sigma^2}\right) = n-1$.
Ainsi : $\mathbb{E}\left(\dfrac{SCE}{n-1}\right) = \sigma^2$.

Remarque 1 : La variance empirique corrigée $\hat{S}^2 = \dfrac{SCE}{n-1}$ est un estimateur non biaisé de σ^2 alors que la variance empirique $S^2 = \dfrac{SCE}{n}$ est un estimateur biaisé de σ^2.

Remarque 2 : $\hat{s}$ est une estimation biaisée de σ.

PROPRIÉTÉ 8 : Estimation ponctuelle d'une proportion
Considérons une certaine population dans laquelle la proportion d'individus ayant une propriété donnée est égale à p.
- $F = \dfrac{X}{n}$ est un estimateur sans biais de p.
- Une estimation ponctuelle de p est donnée par

$$\widehat{p} = f$$

En effet, le variable X dénombrant les individus d'un échantillon de taille n possédant cette propriété suit la loi binomiale $\mathcal{B}(n, p)$.
On a donc : $\mathbb{E}(F) = \mathbb{E}\left(\dfrac{X}{n}\right) = \dfrac{1}{n}\mathbb{E}(X) = \dfrac{1}{n} \times np = p$.

II.B Estimation par intervalle de confiance

DÉFINITION 3 : Estimer un paramètre θ par intervalle (symétrique en probabilité), au niveau de confiance $1-\alpha \in]0;1[$, à partir d'un estimateur Y revient à déterminer un réel positif η tel que

$$\mathbb{P}\left(Y - \eta \leq \theta \leq Y + \eta\right) = 1 - \alpha$$

L'intervalle $[y - \eta; y + \eta]$ est une estimation par intervalle de confiance de θ au niveau de confiance $1 - \alpha$.

REMARQUE 3 : Dans le cas d'une distribution normale d'écart-type connu σ, une estimation par intervalle de confiance de la moyenne μ au niveau de confiance 0,95 est

$$\left[\overline{x} - 1,96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right]$$

Ainsi, si l'on considèrait tous les intervalles de confiance de la moyenne associés à une expérience aléatoire donnée, 95 % d'entre eux contiendraient la moyenne μ de la population.
Il est donc très vraisemblable que la moyenne μ de la population appartienne à l'intervalle ci-dessus mais on ne peut associer aucune probabilité d'appartenance ...

Chapitre III

Généralités sur les tests statistiques

III.A Un peu d'histoire : Neyman, Pearson et Fisher

III.A.1 Les tests selon Fisher : 1 hypothèse et 1 probabilité

Dans la théorie des tests de Fisher, une seule hypothèse est testée. Elle est appelée "hypothèse nulle" dans le sens "to be nullified" c'est-à-dire à réfuter et notée $\mathcal{H}_0$.
L'hypothèse $\mathcal{H}_0$ n'est jamais "prouvée" mais peut être rejetée.
La conclusion est basée sur le calcul d'une probabilité conditionnelle (p-valeur sous $\mathcal{H}_0$) correspondant à la probabilité d'avoir une observation égale ou pire à celle que l'on a obtenue. Si cette probabilité est jugée suffisamment faible, on considère qu'on a réussi à montrer la fausseté de l'hypothèse nulle, on la rejette et le résultat est déclaré significatif. Si p est trop élevé, on suspend le jugement et le résultat est déclaré non significatif.
Il n'y a qu'un seul risque d'erreur (α) correspondant à la probabilité de rejeter $\mathcal{H}_0$ à tort. Pour juger de p, on raconte que le seuil de 5 % avait sa préférence car il percevait 5 % de royalties sur ses publications. La p-valeur est centrale. Plus elle est petite, plus on a de preuves contre $\mathcal{H}_0$.
• $p = 0,049$: l'observation est "limite"
• $p = 0,001$: l'observation est "probante"
• $p = 0,049$ et $p = 0,051$: les résultats sont voisins.

Critique :
La p-valeur correspond à la probabilité d'avoir une observation pire (ou égale) à celle que l'on a obtenue (sous $\mathcal{H}_0$).
$\mathcal{H}_0$ peut être rejetée (alors qu'elle est vraie) parce qu'elle est "en contradiction" avec des résultats qui n'ont pas été observés.

III.A.2 Les tests selon Neyman-Pearson : 2 hypothèses et 1 région critique

En 1928 et 1933, Neyman et Pearson introduisent la notion d'hypothèse alternative $\mathcal{H}_0$, induisant l'apparition d'un second risque (de deuxième espèce β). Basée sur la règle de décision par rapport à α (du ressort du chercheur), p non présentée, une seule hypothèse
Si le seuil est fixé à 5% alors
• $p = 0,049$ et $p = 0,001$ conduisent à la même conclusion : rejet de $\mathcal{H}_0$
• $p = 0,049$ et $p = 0,051$ conduisent à des conclusions différentes.

Critique :
Cette méthode, basée sur la notion de région critique (zone de rejet $\mathcal{H}_0$) ne tient pas compte du "degré" de preuve (lié à p).

III.A.3 Les tests des nos jours : l'approche N-P-F

De nos jours, on utilise une méthode hybride basée sur les construction de Neyman, Pearson et Fisher. Les différentes étapes d'un test sont les suivantes :

1. **Hypothèses d'un test**
   La première étape d'un test consiste à déterminer l'hypothèse à tester.
   Cette hypothèse, notée H_0, est appelée hypothèse nulle. Il s'agit, en général, d'une égalité.
   On définit ensuite l'hypothèse qui sera retenue si on rejette $\mathcal{H}_0$.
   Cette hypothèse est appelée l'hypothèse alternative $\mathcal{H}_1$ et se présente souvent sous la forme :
   soit " $\ldots.\neq.\ldots$ " ; on dit que le test est bilatéral.
   soit " $\ldots.>.\ldots$ " ou " $\ldots.<.\ldots.$ " ; on dit alors que le test est unilatéral.

2. **Seuil de signification et choix du modèle**
   Le niveau de signification d'un test est fixé fixé a priori, est fréquemment égal à 5 %.
   Quant à la statistique de test (variable de décision) et la loi associée, elle est définie suivant le paramètre considéré.
   La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse $\mathcal{H}_0$ est vraie.
   Par exemple, dans le cas d'un test de conformité d'une moyenne ($H_0 : \mu = \mu_0$), on prendra, sous réserve de validité : $\dfrac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \hookrightarrow \mathcal{N}(0;1)$.

3. **Calcul de la p-valeur (p-value)**
   Une règle de rejet, utilisée par le logiciel R, consiste à calculer la probabilité que la statistique de test soit égale à la valeur observée ou encore plus extrême, tout en supposant que l'hypothèse nulle $\mathcal{H}_0$ est vraie :
   on appelle cette probabilité la p-valeur (p-value en anglais) voire probabilité critique.
   Elle correspond à la **probabilité d'avoir une observation égale ou "pire" que celle que l'on a obtenue. Dans le cas d'un test bilatéral, la p-value correspond au double de celle qui aurait été obtenue lors de la mise en place d'un test unilatéral.** Nous voyons donc que la p-valeur est une probabilité calculée a posteriori, en fonction des données.

4. **Conclusion**
   Si p-valeur $\leqslant 0,05$ alors il y avait moins de 5% de chance que la statistique de test prenne une valeur "pire" que la valeur observée ce qui conduit à remettre en question l'hypothèse initiale. On rejette donc $\mathcal{H}_0$. Plus généralement :
   on rejette $\mathcal{H}_0$ (au profit de $\mathcal{H}_1$) lorsque p-value $\leqslant \alpha$, et le test est effectué au risque α.
   Dans le cas contraire, on ne rejette pas $\mathcal{H}_0$.
   Si l'hypothèse $\mathcal{H}_0$ est rejetée alors que le seuil de signification considéré était égal à **5 %**, on dit qu'on rejette $\mathcal{H}_0$ de manière significative.

III.B Risques de première et de deuxième espèce

Tous les règles de décision acceptent un risque α correspondant au risque de rejeter à tort l'hypothèse $\mathcal{H}_0$, c'est-à-dire le risque de rejeter l'hypothèse $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie. Ce risque s'appelle aussi le *risque de première espèce*.
La règle de décision du test comporte également un deuxième risque, à savoir de celui de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ alors que c'est l'hypothèse $\mathcal{H}_1$ qui est vraie. C'est le *risque de deuxième espèce*.

Les deux risques peuvent se définir ainsi :

$$\begin{aligned} \alpha &= \mathbb{P}(\text{rejeter}\mathcal{H}_0 \,|\, \mathcal{H}_0\text{vraie}) = \text{probabilité de commettre une erreur de première espèce.} \\ \beta &= \mathbb{P}(\text{ne pas rejeter}\mathcal{H}_0 \,|\, \mathcal{H}_1\text{vraie}) = \text{probabilité de commettre une erreur de deuxième espèce.} \end{aligned}$$

Le risque de première espèce α est choisi a priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative $\mathcal{H}_1$ et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse $\mathcal{H}_1$ que l'on suppose vraie.
Les risques liés aux tests d'hypothèses peuvent se résumer ainsi :

	$\mathcal{H}_0$ est vraie	$\mathcal{H}_0$ est fausse
$\mathcal{H}_0$ n'est pas rejetée	$1-\alpha$	β
$\mathcal{H}_0$ est rejetée	α	$1-\beta$

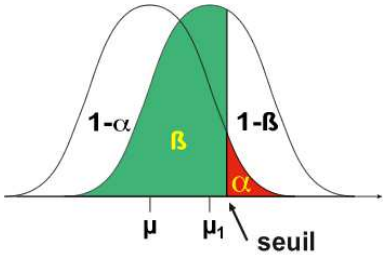
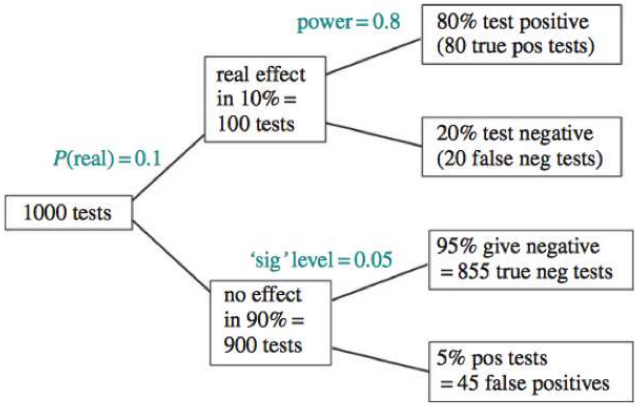


FIGURE III.1 – Cas d'un test de conformité d'une moyenne

DÉFINITION 4 : Puissance d'un test
$1-\beta$ définit la *puissance du test* à l'égard de la valeur du paramètre dans l'hypothèse alternative $\mathcal{H}_1$.
La puissance du test représente la probabilité de rejeter l'hypothèse nulle $\mathcal{H}_0$ lorsque l'hypothèse vraie est $\mathcal{H}_1$.
Plus β est petit, plus le test est puissant.

III.B.1 Comment utiliser les tests ?

Se limiter à un simple calcul de p-value pour décider d'un éventuel effet est trop réducteur et source de beaucoup d'erreurs.
Par exemple, supposons que sur 1000 tests, seuls 10 % aient un effet réel. On fixe $\alpha = 5\%$ et $\beta = 20\%$ soit $1-\beta = 80\%$.



Ainsi, $80 + 45 = 125$ tests s'avèreront positifs (rejet de $\mathcal{H}_0$) et seulement 80 seront de "vrais" positifs. La probabilité, lorsqu'on a déclaré un effet positif (rejet de $\mathcal{H}_0$) qu'il n'y ait pas d'effet est

$$\frac{45}{125} \simeq 0,36$$

Près de $\frac{1}{3}$ des effets annoncés comme significatifs ne le sont pas ... Il est donc important d'accompagner un test de mesures ou analyses telles que :

- Calcul de la puissance d'un test (ajustement éventuel de la taille d'échantillon)
- Détermination d'intervalles de confiance
- Calculs de la taille de l'effet.

III.B.1.a Puissance d'un test

La puissance d'un test correspond à la probabilité de mettre en évidence un effet lorsque celui existe.
Il s'agit donc de la capacité d'un test à prendre la bonne décision lorsqu'il existe un effet.

1. Pour un même risque α et une même taille d'échantillon, on constate que, si *l'écart* Δ entre la valeur du paramètre posée en $\mathcal{H}_0$ et celle supposée dans l'hypothèse vraie $\mathcal{H}_1$ augmente, le risque β diminue ce qui induit une augmentation de $1-\beta$.
2. Une diminution de la *variabilité* (cf. écart-type) peut également induire une augmentation de la puissance du test.
3. Enfin, une *augmentation de la taille du ou des échantillons* aura pour effet de donner une meilleure précision.
   Le test est alors plus puissant.

Il est recommandé de choisir une taille d'échantillon conduisant, a priori, à une **puissance de test au moins égale à 80 %**. Avec R, on peut utiliser des fonctions implémentées de base ('power.t.test' ou 'power.anova.test') ou accessibles via le package 'pwr'.

III.B.1.b Taille de l'effet

Pour un grand échantillon, un effet minime (et sans réelle signification) suffira à faire rejeter l'hypothèse $\mathcal{H}_0$.
On peut donc introduire la notion de *taille de l'effet* désignant à quel degré un phénomène donné est présent dans la population (*Revue des sciences de l'éducation, volume 31, 2009*).
Ainsi, **ne pas rejeter l'hypothèse nulle revient à considérer que la taille de l'effet est nulle**.

Soit Δ l'écart entre la moyenne de la population et une valeur cible, ou entre les moyennes de deux populations.
La taille de l'effet est déterminé par

$$\text{Taille de l'effet} = \frac{\Delta}{\sigma}$$

où σ correspond à l'écart-type des populations.
On peut utiliser des niveaux définis par Cohen pour qualifier un effet.
- $d = 0.2$: Effet faible
- $d = 0.5$: Effet moyen
- $d = 0.8$: Effet fort

Chapitre IV

Tests usuels

IV.A Tests paramétriques de conformité

On considère des échantillons de taille n.

Paramètre	**Conditions**	**Statistique de test et loi**
Moyenne μ	Distribution normale et écart-type σ connu	$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ $*$
	Distribution normale et écart-type inconnu	$\frac{\bar{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$ $**$
	Distribution quelconque et grand échantillon ($n \geqslant 30$)	$\bar{X} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$
Variance σ^2	Distribution normale	$\frac{SCE}{\sigma^2} \sim \chi^2(n-1)$
Proportion p		$X \sim \mathcal{B}(n,p)$
	Grand échantillon	$F \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$

$*$ On rappelle que $\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ conduit à $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0;1)$.

$**$ $\hat{S}^2 = \frac{SCE}{n-1}$.

IV.B Tests paramétriques de comparaison

On considère deux échantillons de tailles respectives n_1 et n_2.

Paramètres	**Conditions**	**Statistique de test et loi**
Variances	Distributions normales et indépendantes	$\frac{\hat{S}_1^2}{\hat{S}_2^2} \sim \mathcal{F}(n_1-1, n_2-1)$
Moyennes	Distributions normales, indépendantes, variances connues	$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(0; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
	Distributions normales, indépendantes et homoscédastiques	$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{CM}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}\,(n_1 + n_2 - 2\,)$ $*$
	Distributions normales, indépendantes et homoscédastiques	$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathcal{T}\,(\nu)$ $**$
	Grands échantillons indépendants	$\bar{X}_1 - \bar{X}_2 \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
	Échantillons appariés, distribution des différences normale	$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$
Proportions	Grands échantillons indépendants	$F_1 - F_2 \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0; \sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$ $***$

$*$ $CM = \dfrac{SCE_1 + SCE_2}{n_1 + n_2 - 2}$.
Sa réalisation est une "bonne" estimation de la variance commune des deux populations.

$**$ Test t de Welch

$***$ $\hat{p}$ représente une estimation de la proportion commune d'individus pésentant le caractère étudié dans les deux populations.

IV.C Tests du Khi-deux, test exact de Fisher et test sur la médiane

- **Test d'indépendance**
  Ce test est utilisé pour tester l'indépendance de deux variables (caractères) X (prenant p modalités) et Y (prenant q modalités) étudiés sur une population à partir d'observations réalisées sur un échantillon.

- **Test d'homogénéité de plusieurs populations**
  Ce test est utilisé pour comparer la distribution d'une variable qualitative (à p modalités) sur q populations indépendantes (à partir d'observations réalisées sur un échantillon).
  Il peut donc être utilisé pour comparer plusieurs proportions sur des échantillons indépendants.

  La statistique de test et la loi utilisée sont, sous $\mathcal{H}_0$,

$$\sum_{i,j} \frac{(N_{ij} - N_{ij}^{th})^2}{N_{ij}^{th}} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\left((p-1)(q-1)\right)$$

- **Test d'ajustement à une loi théorique**
  Ce test est utilisé pour tester si un échantillon dont on a la distribution des effectifs (observés avec répartition en p classes) peut provenir d'une loi donnée (théorique).
  La statistique de test et la loi utilisée sont, sous $\mathcal{H}_0$,

$$\sum_{i} \frac{(N_i - N_i^{th})^2}{N_i^{th}} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\left(p-1\right)$$

  où $N_i^{th} = Np_i$, p_i correspondant à la probabilité d'avoir l'évènement i et $N = \sum_i N_i$.

- **Test exact de Fisher**
  Le test exact de Fisher est un test d'indépendance (comme le test du Khi-deux) pouvant être utilisé pour tester l'indépendance de deux caractères A et B prenant chacun 2 modalités (A_1, A_2 et B_1, B_2) Il présente l'avantage d'utiliser une loi exacte (d'où son nom). Les données peuvent se présenter par un tableau de contingence comme ci-dessous :

	A_1	A_2	$Total$
B_1	a	b	$a+b$
B_2	c	d	$c+d$
$Total$	$a+c$	$b+d$	$a+b+c+d=n$

  La probabilité de cette configuration est donnée, sous l'hypothèse nulle d'absence d'association, par la loi hypergéométrique :

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

  La p-value s'obtient en ajoutant les probabilités de chacun des tableaux aussi éloignés ou plus éloignés de l'indépendance que la table observée.
  Il s'agit alors d'un test bilatéral.

- **Test sur la médiane**
  Ce test permet de tester l'hypothèse selon laquelle deux populations ont la même médiane.
  On note M_e la médiane lorsqu'on regroupe les deux populations.
  Le test est basé sur un test d'indépendance lorsqu'on regroupe les données ainsi :

	$Valeurs > M_e$	$Valeurs \leq M_e$	$Total$
Echantillon 1	a	b	n_1
Echantillon 2	c	d	n_2
$Total$	$a+c$	$b+d$	$n_1+n_2=n$

  On peut alors utiliser le test **exact de Fisher** voire un test du **Khi-deux**.

IV.D ANOVA à 1 facteur

Dans cette partie, on présente le cas de dispositifs sans facteurs emboités, à effets fixes (Anova de type I).

DÉFINITION 5 : Un facteur correspond à une variable qualitative (catégorielle) dont les différentes valeurs prises sont appelés niveaux, modalités voire catégories.
On distingue deux types de facteurs :

- les facteurs provoqués (introduits volontairement) ;
- les facteurs aléatoires : inhérents au milieu (terrain, environnement de l'essai) appelés facteurs contrôlés lorsque le dispositif expérimental utilisé les prend en compte.

On considère un facteur prenant p niveaux.
Le niveau i contient n_i observations. On note $N = \sum_{i=1}^{p} n_i$.
On note $\bar{x}_i$ la moyenne observée pour le i-ème niveau et $\bar{x}$ la moyenne globale.
On note x_{ij} la j-ème répétition pour le niveau i.

Répétition	Modalité 1	$\cdots$	Modalité i	$\cdots$	Modalité p
1	x_{11}	$\cdots$	x_{i1}	$\cdots$	x_{p1}
2	x_{12}	$\cdots$	$\cdots$	$\cdots$	x_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_{1j}	$\cdots$	x_{ij}	$\cdots$	x_{pj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

On a

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

L'Anova à un facteur correspond donc à un modèle linéaire qui s'écrit ainsi

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où

- μ est la moyenne (globale) associée aux p populations.
- X_{ij} est une variable aléatoire (réponse quantitative). Les X_{ij} sont indépendantes.
- ε_{ij} est une variable aléatoire d'erreur (non observées). On a :
  $\varepsilon_{ij} \sim \mathcal{N}(0;\sigma)$, σ étant un paramètre inconnu (à estimer).
- α_i correspond à l'effet associé à la modalité i avec $\sum_{i=1}^{p} \alpha_i = 0$. Une estimation de cet effet est donnée par

$$\hat{\alpha}_i = \bar{x}_i - \bar{x}$$

Le modèle peut également s'écrire

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

où $\mu_i = \mu + \alpha_i$ dont une estimation est $\hat{\mu}_i = \bar{x}_i$.

L'analyse de variance permet de tester l'hypothèse de nullité de tous les paramètres α_i du modèle alors que l'utilisation du modèle linéaire permet de tester la nullité de chaque paramètre du modèle. On a :

$\mathcal{H}_0 : \alpha_1 = \cdots = \alpha_p = 0$ qui s'écrit également $\mathcal{H}_0 : \mu_1 = \cdots = \mu_p$
et
$\mathcal{H}_1$: "un des paramètres α_i au moins n'est pas nul" qui s'écrit également
$\mathcal{H}_1$: "deux moyennes au moins sont différentes".

DÉFINITION 6 : On appelle résidu toute valeur notée e_{ij} définie par

$$e_{ij} = x_{ij} - \bar{x}_i$$

PROPRIÉTÉ 9 : Les résidus sont de somme et de moyenne nulle (par niveau et globalement).

En effet :
$\bar{x}_i$ correspond à la moyenne du niveau i dont les observations sont notées x_{ij}. Ainsi

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

On a donc $\sum_{j=1}^{n_i} e_{ij} = \sum_{j=1}^{n_i} x_{ij} - \sum_{j=1}^{n_i} \bar{x}_i = n_i\bar{x}_i - n_i\bar{x}_i = 0$.

Le modèle linéaire d'Anova nécessite que $\varepsilon_{ij} \sim \mathcal{N}(0;\sigma)$ i.e. $X_{ij} \sim \mathcal{N}(\mu_i;\sigma)$.
En conséquence, les variables ε_{ij} doivent

- avoir la même variance (hypothèse d'homoscédasticité),
- être indépendantes,
- être distribuées normalement.

Ces hypothèses peuvent être vérifiées à l'aide des graphiques diagnostiques.

$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})$ avec $e_{ij} = x_{ij} - \bar{x}_i$ donc

$$x_{ij} - \bar{x} = e_{ij} + (\bar{x}_i - \bar{x})$$

En élevant au carré et en sommant sur j, il vient
$\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{n_i} e_{ij}^2 + \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 + \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})\, e_{ij}$. Or :
$(\bar{x}_i - \bar{x})$ ne dépend pas de j donc $\sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})\, e_{ij} = (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} e_{ij} = 0$ car la somme des résidus est nulle. Ainsi :

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{n_i} e_{ij}^2 + \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2$$

En posant

- $SCE_{fact} = SCE_{inter} = \sum_{i,j} (\bar{x}_i - \bar{x})^2$ (source de variations factorielle, expliquée)
- $SCE_{res} = SCE_{intra} = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$ (source de variations résiduelle)
- $SCE_{totale} = \sum_{i,j} (x_{ij} - \bar{x})^2$ (source de variations totale) on a

$$SCE_{totale} = SCE_{fact} + SCE_{res}$$

DÉFINITION 7 : On définit la notion de Carré Moyen, noté CM, par

$$CM = \frac{SCE}{\text{ddl}}$$

où ddl correspond aux degrés de libertés (associés au modèle).

On a

- $CM_{res} = \dfrac{SCE_{res}}{N-p} \sim \chi^2(N-p)$.
- Sous $\mathcal{H}_0$, $CM_{tot} = \dfrac{SCE_{tot}}{N-1} \sim \chi^2(N-1)$.
- En outre, si $X \sim \chi^2(a)$ et $Y \sim \chi^2(b)$ avec X et Y indépendantes alors $X+Y \sim \chi^2(a+b)$. Ainsi, sous $\mathcal{H}_0$, $CM_{fact} = \dfrac{SCE_{fact}}{p-1} \sim \chi^2(p-1)$.

On en déduit que

Sous $\mathcal{H}_0$,

$$F_{obs} = \frac{CM_{fact}}{CM_{res}} \sim \mathcal{F}(p-1; N-p)$$

On rejette $\mathcal{H}_0$ si **le bruit expliqué est "bien plus grand" que le bruit résiduel c'est-à-dire lorsque** $f_{obs} >> 1$. L'Anova est donc, par nature, un test **unilatéral à droite**. On a donc

$$p - \text{valeur} = \mathbb{P}(F_{obs} > f_{obs})$$

IV.E Arbre de choix autour de la recherche d'un effet factoriel

Cette partie est en lien avec une partie du cours de 2A.