Échantillonnage & Estimation | T.S.

Vue d'ensemble : Résumé de toute une scolarité d'échantillonnage et d'estimation en un coup d'œil

- Soit p la proportion réelle dans la population totale = la probabilité qu'un individu choisi au hasard dans la population ait la caractéristique à laquelle on s'intéresse. C'est la fréquence théorique.
- Soit f la fréquence observée de cette caractéristique dans l'échantillon.

	Échantillonnage ***p connue ou bien on veut tester une hypothèse sur la valeur de p*** Utilisation = Prise de décision à partir d'un échantillon : On regarde si la valeur observée pour f est « raisonnable » *(Pour voir s'il y a discrimination contre les femmes dans une entreprise, pour voir si un jury a vraiment été choisi au hasard...etc)*	**Estimation** ***p inconnue*** *(et on n'a aucune hypothèse sur la valeur de p)* Utilisation = Estimation : On estime p au moyen de f . *(pour estimer la proportion p de gens qui vont voter pour un candidat à partir des résultats d'un sondage...etc)*
En seconde	Si $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$ alors pour au moins 95% des échantillons, la fréquence **observée *f*** vérifie $f \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ *Intervalle de **fluctuation** au seuil de 95%*	Si $n \geqslant 25$ et $0{,}2 \leqslant f \leqslant 0{,}8$ alors pour au moins 95% des échantillons, la fréquence **théorique** vérifie $p \in \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ *Intervalle de **confiance*** au seuil de 95%

Le tirage au hasard dans la population d'un individu qui peut présenter le caractère C avec la probabilité p est une épreuve de Bernoulli de paramètre p où le succès est l'issue « avoir C ». Le prélèvement au hasard d'un échantillon de taille n dans cette population s'assimile à un schéma de Bernoulli. La variable aléatoire qui compte le nombre de succès, c'est à dire le nombre d'individus présentant le caractère C, suit la loi binomiale $\mathscr{B}(n, p)$. Voilà pourquoi les méthodes de Première et Terminale se focalisent sur les loi binomiales.

En 1S (si on a le temps de finir le programme)	On calcule les probabilités d'avoir k succès sur n tirages pour la loi binomiale et on en déduit l'intervalle de fluctuation au seuil de 95%.	
En TS *(seulement le Best of)*	Si $n \geqslant 30$, $n\,p \geqslant 5$ et $n(1-p) \geqslant 5$, alors pour environ 95% des échantillons, on peut considérer que la fréquence **observée *f*** vérifie $f \in \left[p - 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ ↑ ***Intervalle de fluctuation asymptotique au seuil de 95%*** (Le 1,96 vient de la loi normale, voir P9)	*Même résultat qu'en seconde* : Pour au moins 95% des échantillons, la fréquence **théorique *p*** vérifie $p \in \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ ↑ ***Intervalle de confiance au seuil de 95%*** mais sous les conditions $n \geqslant 30$, $n\,f \geqslant 5$ et $n(1-f) \geqslant 5$
Un jour peut-être		*Résultat symétrique de celui de TS sur l'intervalle de fluctuation :* Pour environ 95% des échantillons, la fréquence **théorique *p*** vérifie $p \in \left[f - 1{,}96 \frac{\sqrt{f(1-f)}}{\sqrt{n}} ; f + 1{,}96 \frac{\sqrt{f(1-f)}}{\sqrt{n}} \right]$ *Intervalle de confiance au seuil de 95%* ↴ sous les conditions $n \geqslant 30$, $n\,f \geqslant 5$ et $n(1-f) \geqslant 5$

Remarque : Nous ne mentionnerons plus jamais le dernier résultat mais vu la symétrie des résultats d'échantillonnage et d'estimation vus en seconde, on ne peut pas s'empêcher de se demander si le résultat d'échantillonnage de TS n'aurait pas son symétrique en estimation : La réponse est donc « *oui, mais le résultat en question n'est pas su programme de TS* ». Savoir que ce résultat existe satisfait cependant notre sens de l'esthétique.

Introduction : Ce chapitre permet surtout de répondre à deux types de questions :

♠ Exemple d'introduction 1. Vu le sondage, le candidat sera-t-il élu ou non ? (Oral)

Avant une élection, un sondage indique qu'un candidat a 56 % d'intentions de votes. Mais si un sondage sur 4 personnes ne nous dit pas la même chose que s'il s'agit d'un sondage sur 1000 personnes. Remarquons aussi que la proportion de votants favorables au candidats dépend de l'échantillon choisi. Peut-on utiliser l'échantillon pour obtenir un encadrement des intentions de votes dans la population ? Quelle est la fiabilité de cet encadrement ?

♠ Exemple d'introduction 2. Y a-t-il discrimination ou non ? (Oral)

Au vu de la proportion de femmes ou de noirs dans une entreprise, comment établir s'il y a discrimination ou non ? A partir de quelle valeur la proportion observée est-elle anormalement faible et révèle une discrimination ? Peut-on être sûr(e) à 100% que la proportion est anormale? Alors « sûr » dans quel sens ?

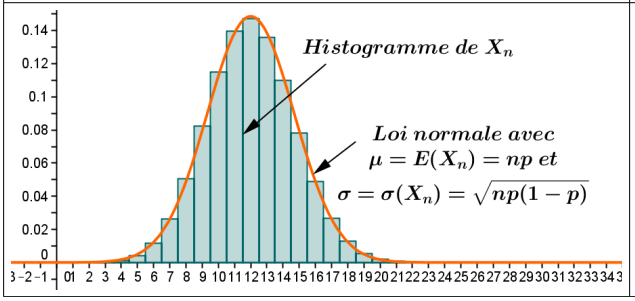
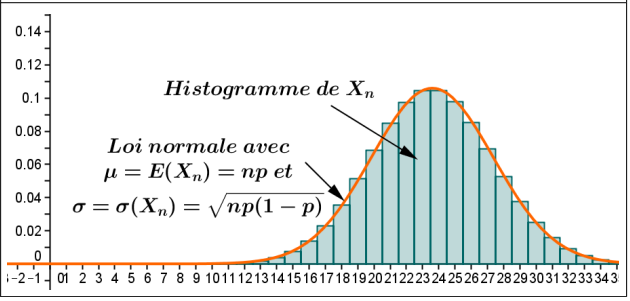
I. Approximation de la loi binomiale par la loi normale : Théorème de Moivre-Laplace

A. Pour quoi faire ? Pour que les calculs soient gérables

Exemple : On lance $n=10\,000$ fois une pièce équilibrée. La variable aléatoire X_n qui compte le nombre de fois où on a obtenu Pile sur n lancers suit une loi binomiale. Tout ceci ne pose aucun problème théorique mais le calcul de $P(4\,900\leqslant X_{10^4}\leqslant 5\,100)=\sum_{k=4900}^{5100}\binom{10000}{k}(0{,}5)^{10\,000}$ n'est pas faisable à la machine car les « k parmi n » sont trop grands. On voudrait donc approximer cette probabilité par une expression qui donne un calcul faisable (à la machine).

B. Quelques figures qui aident à comprendre intuitivement le théorème de Moivre-Laplace

■ **Avec des variables aléatoires ni centrées ni réduites** : On a représenté sur chacune des figures ci-dessous l'histogramme de variables aléatoires X_n qui suivent la loi binomiale $\mathscr{B}(n;p)$ ainsi que la densité de probabilité des lois normales de même moyenne et de même écart-type que X_n, càd de moyenne $\mu_n=np$ et d'écart-type $\sigma_n=\sqrt{np(1-p)}$.

$n=30$ et $p=0{,}4$	$n=60$ et $p=0{,}4$
	

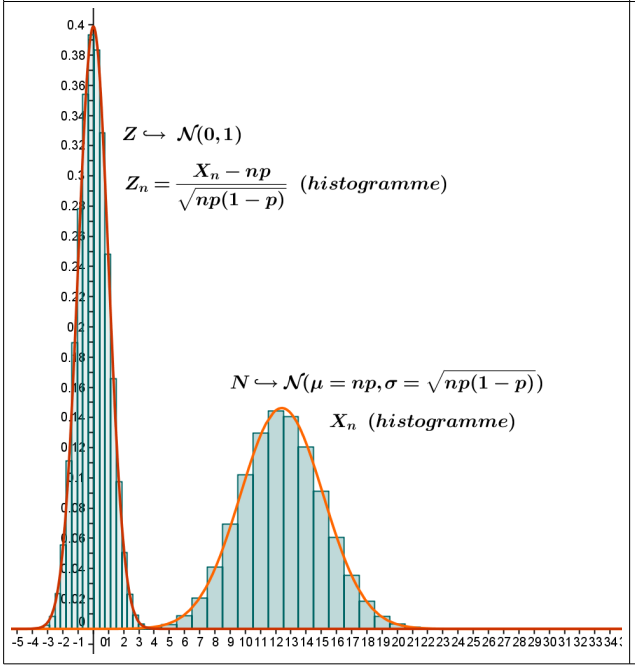
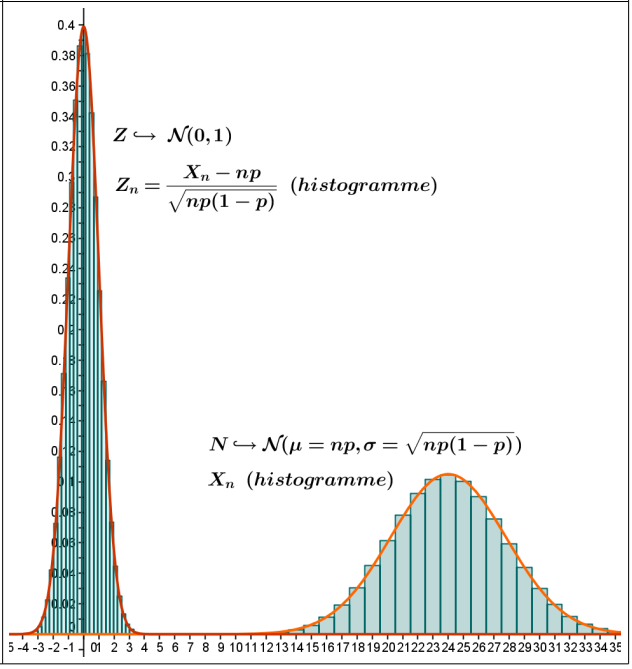
Observations :

- Dans les deux cas, il semble que la courbe en cloche donne une bonne approximation de l'histogramme.
- Les deux courbes obtenues sont des courbes en cloches mais elles sont assez différentes l'une de l'autre.

Idée : Pour toujours se ramener à la même courbe en cloche quelle que soit la loi binomiale de laquelle on part, il suffit de centrer et réduire. Faisons-le :

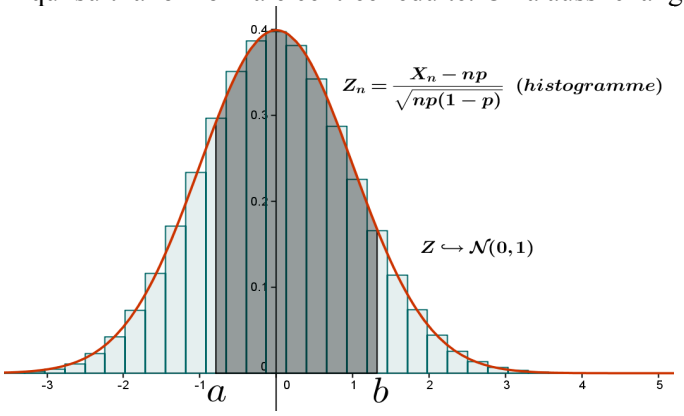
■ **Avec des variables aléatoires centrées et réduites : C'est le théorème de Moivre-Laplace**

- Comme précédemment, on a représenté sur chacune des figures ci-dessous l'histogramme d de variables aléatoires X_n qui suivent la loi binomiale $\mathscr{B}(n;p)$ ainsi que la densité de probabilité des lois normales de même moyenne et de même écart-type que X_n, càd de moyenne $\mu_n=np$ et d'écart-type $\sigma_n=\sqrt{np(1-p)}$.
- Par rapport aux figures précédentes, on a rajouté l'histogramme de la variable aléatoire Z_n obtenue en centrant et réduisant X_n (*sur les deux figures, c'est l'histogramme de gauche, le plus haut des deux*) et la densité de probabilité de la loi normale centrée réduite.

$n=30$ et $p=0{,}4$	$n=60$ et $p=0{,}4$
	

Observations :
Il semble que quelle que soit la loi binomiale dont on part, l'histogramme de la variable aléatoire Z_n obtenue en centrant et réduisant X_n (*sur les deux figures, c'est l'histogramme de gauche, le plus haut des deux*) est plus ou moins le même quand n est assez grand et qu'il peut être approximé par la densité de probabilité de la loi normale centrée réduite. La variable Z_n « tend »vers une variable universelle indépendante de p !

■ **Utilisation pour calculer une probabilité** :
Contrairement aux figures précédentes, on n'a représenté que les variables centrées et réduites : on a donc l'histogramme de la variable aléatoire Z_n obtenue en centrant et réduisant X_n (*sur les deux figures, c'était l'histogramme de gauche, le plus haut des deux*) et la densité de probabilité d'une variable aléatoire Z qui suit la loi normale centrée réduite. On a aussi changé l'échelle des figures.



La probabilité $P\left(a \le Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} \le b\right)$ s'obtient en additionnant l'aire des rectangles correspondants de l'histogramme de Z_n.

Sur la figure il semble que cette aire est proche de l'aire sous la courbe de densité de la loi normale si n est assez grand. Cette aire (en gris foncé) est égale à

$$P(a \le Z \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Le théorème de Moivre-Laplace formule tout ceci de façon plus précise :

C. Énoncé du théorème de Moivre-Laplace

[P1] Théorème de Moivre-Laplace. (*admis*)
Soit n un entier naturel non nul et p un réel dans $]0;1[$. Soit (X_n) une suite de variables aléatoires où X_n suit une loi binomiale $\mathcal{B}(n;p)$.
Alors pour tous réels a et b, $\lim\limits_{n\to+\infty} P\left(a \le \frac{X_n - np}{\sqrt{np(1-p)}} \le b\right) = P(a \le Z \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ où Z suit la loi normale centrée réduite $\mathcal{N}(0;1)$.

D. Application pratique (admis)

[P2] Application pratique. On considère que la limite est pratiquement atteinte lorsqu'on a simultanément $n \geqslant 30, np \geqslant 5$ et $n(1-p) \geqslant 5$. (*Autrement dit, n doit être suffisamment grand et p ni trop proche de 0 ni trop proche de 1*). Dans ces conditions, on considère que pour tous réels a et b , $P\left(a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b\right) \approx P(a \leq Z \leq b)$ où Z suit la loi normale centrée réduite $\mathcal{N}(0;1)$.

Remarque [3]. On est pas obligé de repasser à chaque fois par les lois centrées réduites: En effet soit X_n une variable aléatoire suivant une loi binomiale $\mathcal{B}(n;p)$. L'espérance de X_n est $\mu_n = np$ et son écart-type $\sigma_n = \sqrt{np(1-p)}$. Étant donnés deux réels *a* et *b*, on a

$$P(a \leq X_n \leq b) = P\left(\frac{a-\mu_n}{\sigma_n} \leq \frac{X_n - \mu_n}{\sigma_n} \leq \frac{b-\mu_n}{\sigma_n}\right) \overset{(i)}{\approx} P\left(\frac{a-\mu_n}{\sigma_n} \leq Z \leq \frac{b-\mu_n}{\sigma_n}\right) = P(a \leq \sigma_n Z + \mu_n \leq b) = P(a \leq N \leq b)$$

où Z suit la loi normale centrée réduite $\mathcal{N}(0;1)$ et $N = \sigma_n Z + \mu_n$ suit la loi normale de même espérance et écart type que X_n càd que *N* suit $\mathcal{N}(\mu_n; \sigma_n^2)$.

(i) Par [P2], l'application pratique du Thm de Moivre Laplace ci-dessus.

En résumé,

[P4] Corollaire de l'application pratique. Si les conditions $n \geqslant 30, np \geqslant 5$ et $n(1-p) \geqslant 5$ sont toutes réalisées, on a $P(a \leq X_n \leq b) \approx P(a \leq N \leq b)$ où *N* suit une loi normale de même espérance et même écart-type que X_n .

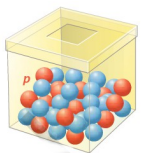
Cette formulation du résultat ne figure pas dans le programme mais elle éclaire les choses. Il faudra donc dans les sujets de bac repasser à chaque fois par les lois centrées réduites, ce qui revient à refaire le petit raisonnement ci-dessus.

Exercice 3. On lance un dé parfaitement équilibré 180 fois et on veut connaître la probabilité d'obtenir entre 27 et 36 fois la face numérotée 6.

II. Échantillonnage : Intervalle de fluctuation asymptotique

A. Intervalle de fluctuation asymptotique au seuil 1-α

■ Exemple 4. On dispose d'une urne contenant un très grand nombre de boules rouges et bleues. On sait que la proportion de boules rouges dans l'urne est égale à $p = 0{,}4$.
Si on tire successivement avec remise, *n* boules dans l'urne ($n \in \mathbb{N}^*$), et si on appelle X_n la variable aléatoire dénombrant les boules rouges tirées, alors X_n suit une loi binomiale $\mathcal{B}(n;p)$.



Si on tire plusieurs fois de suite un échantillon de 50 boules dans l'urne, la fréquence d'apparition des boules rouges ne sera pas à chaque fois exactement égale à 0,4 : Il y aura des *fluctuations autour de 0,4*. On s'attend à avoir d'autant plus de fluctuation que le nombre de boules tiré est petit. Pour mesurer ces fluctuations, on peut par exemple chercher à trouver un intervalle qui contient la fréquence de 90% des échantillons. Le paragraphe ci-dessous explique comment trouver un tel intervalle dans le cas général.

Définition 5. Intervalle de fluctuation asymptotique ou non
Soit X_n une variable aléatoire qui compte les succès dans un schéma de Bernoulli de paramètre (n, p) . X_n suit une loi binomiale $\mathcal{B}(n;p)$. La variable aléatoire $F_n = \frac{X_n}{n}$ mesure la fréquence des succès.

▪ [**2nde**] Un intervalle I est un ***intervalle de fluctuation au seuil de*** $1-\alpha$ si $P(F_n \in I) \geqslant 1-\alpha$.

▪ [**TS**] Un intervalle I_n défini à partir de n et p est un ***intervalle de fluctuation asymptotique***[1] ***au seuil de*** $1-\alpha$ si $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1-\alpha$.

[1] *Asymptotique car il fait intervenir une limite.*

■ **Cas général (ROC exigible)**

• Soit α un réel de $]0;1[$. Si on veut obtenir un intervalle qui contient la fréquence de 95% des échantillons, on prend $1-\alpha=0{,}95=95\%$ càd $\alpha=0{,}05=5\%$; si on veut obtenir un intervalle qui contient la fréquence de 99% des échantillons, on prend $1-\alpha=0{,}99=99\%$ càd $\alpha=0{,}01=1\%$ ...etc.

• On a vu que si une variable aléatoire Z suit une loi normale contrée réduite, il existe un unique réel u_α tel que $P(-u_\alpha\leq Z\leq u_\alpha)=1-\alpha$.

• Soit X_n une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n;p)$ où p est un réel de $]0;1[$ (X_n a donc pour moyenne $E(X_n)=np$ et pour écart-type $\sigma(X_n)=\sqrt{np(1-p)}$). On sait d'après le théorème de Moivre-Laplace que $\lim\limits_{n\to+\infty} P\left(-u_\alpha\leq\dfrac{X_n-np}{\sqrt{np(1-p)}}\leq u_\alpha\right)=P(-u_\alpha\leq Z\leq u_\alpha)=1-\alpha$

Or la condition sur X_n peut se réécrire sous forme d'un encadrement de la fréquence. En effet,

$-u_\alpha\leq\dfrac{X_n-np}{\sqrt{np(1-p)}}\leq u_\alpha \overset{(i)}{\Leftrightarrow} -u_\alpha\sqrt{np(1-p)}\leq X_n-np\leq u_\alpha\sqrt{np(1-p)}$ *(i) En multipliant l'inégalité par* $\sqrt{np(1-p)}>0$

$\overset{(ii)}{\Leftrightarrow} p-u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\leq\dfrac{X_n}{n}\leq p+u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ *(ii) En divisant l'inégalité par n puis en ajoutant p.*

On en déduit :

[P6] Théorème et définition.

Soit u_α l'unique réel tel que $P(-u_\alpha\leq Z\leq u_\alpha)=1-\alpha$ où Z suit une loi normale centrée réduite.

Soit (X_n) une suite de variables aléatoires où X_n suit une loi binomiale $\mathscr{B}(n;p)$. La fréquence de succès $F_n=\dfrac{X_n}{n}$ vérifie $\lim\limits_{n\to+\infty} P\left(p-u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\leq F_n\leq p+u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right)=1-\alpha$

Autrement dit, l'intervalle $\boldsymbol{I_n=\left[p-u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p+u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]}$ contient la fréquence $F_n=\dfrac{X_n}{n}$ avec une probabilité qui tend vers $1-\alpha$ lorsque n tend vers l'infini.

Un tel intervalle s'appelle un ***intervalle de fluctuation asymptotique de F_n au seuil $1-\alpha$***.

♠ Retour à l'exemple 4. Rappelons que l'on tire 50 boules de l'urne décrite ci-dessus, et que l'on souhaite déterminer un intervalle de fluctuation au seuil 0,9 (c'est à dire avec $\alpha=0{,}1$). A l'aide de la calculatrice, on trouve $u_{0,1}\approx 1{,}645$ (à 10^{-3} près). On obtient pour intervalle de fluctuation asymptotique :

$$I_{50}=\left[0{,}4-u_{0,1}\frac{\sqrt{0{,}4\times 0{,}6}}{\sqrt{50}}\,;\,0{,}4+u_{0,1}\frac{\sqrt{0{,}4\times 0{,}6}}{\sqrt{50}}\right]\approx[0{,}286\,;\,0{,}514]$$

Ceci signifie que si on effectue un très grand nombre de fois le tirage (avec remise) de 50 boules dans cette urne (on considère alors que chaque tirage de 50 boules constitue *un* échantillon), la fréquence d'apparition d'une boule rouge est comprise entre 0,286 et 0,514 pour environ 90% des échantillons.

Pour 200 tirages, au même seuil 0,9, on obtient $I_{200}\approx[0{,}343\,;\,0{,}457]$. L'amplitude de l'intervalle, pour un même seuil, a été divisée par 2 en passant de 50 à 200 tirages. Remarquons donc que pour diviser par 2 l'amplitude de l'intervalle il faut non pas doubler la taille de l'échantillon mais la quadrupler (logique vu le $\sqrt{n}$ dans la formule.

B. Un cas particulier important : Intervalle de fluctuation asymptotique au seuil de 95%

[P7] Intervalle de fluctuation asymptotique au seuil de 95%.

Pour une variable aléatoire X_n suivant une loi binomiale $\mathscr{B}(n;p)$, si $n\geqslant 30, np\geqslant 5$ et $n(1-p)\geqslant 5$, alors la fréquence de succès F_n fluctue avec une probabilité d'environ 95% dans l'**intervalle de fluctuation asymptotique au seuil de 95%** qui est $I_n=\left[p-1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p+1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

Dem: On a déjà vu que pour la loi normale, avec $\alpha = 0{,}05$ *on a* $u_{0{,}05} = 1{,}96$. *Ainsi, en partant de P2 (pour savoir ce qu'on peut déduire sous les conditions* $n \geqslant 30 ; np \geqslant 5$ et $n(1-p) \geqslant 5$ *) et en refaisant les même manipulations d'inégalités que pour prouver P6, on obtient la propriété ci-dessus.*

C. Application à la prise de décision : Test d'hypothèse

■ Exemple 5. On admet que la poportion de femmes dans la population française est de 51,4%. On dénombre dans une classe de terminale 16 filles et 24 garçons. Peut-on conclure que les filles sont sous-représentées dans cette classe?

■ **Cas général**: Dans une certaine population, on **suppose** que la proportion d'un certain caractère est p et que l'échantillon observé a été choisi au hasard. En général, on est sûr(e) d'un de ces deux faits; celui sur lequel on a un doute et que l'on veut tester s'appelle l'***hypothèse initiale***.

Démarche: On suppose que l'hypothèse initiale est vraie et on regarde si, dans ce contexte, ce qu'on observe avait une chance "raisonnable" de seproduire ou s'il s'agit d'un événement "très improbable". Dans ce dernier cas, on on se dit que l'hypothèse initiale était probablement[2] fausse. Cela ressemble un peu au raisonnment par l'absurde: "*Si l'hypothèse initiale était vraie, voilà ce que l'on devrait observer (La fréquence de l'echantillon devrait être dans un certain intervalle). Or on ne l'oberve pas, donc l'hypothèse initiale était probablement fausse.*"

Plus précisément:

[P8] Règle de prise de décision au seuil de 5%.

▪ Si $n \geqslant 30, np \geqslant 5$ et $n(1-p) \geqslant 5$ mais que pourtant la **fréquence observée n'est PAS dans l'intervalle de fluctuation asymptotique au seuil de 95%**, défini rappellez-vous par $I_n = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$, cela indique que :

dans la populations initiale la proportion n'etait pas égale à p,
ou l'échantillon n'a pas été choisi au hasard,
ou l'on est tombé sur un des rares[3] échantillons pour lesquels cela se produit.

Dans ce cas, ***on rejette l'hypothèse initiale au seuil de 5%*** (càd en prenant un risque de 5% de la rejeter à tort).

▪ Si la fréquence observée est dans l'intervalle de fluctuation asymptotique au seuil de 95%, on ne peut pas rejeter l'hypothèse initiale donc on l'accepte (jusqu'à avoir éventuellement un jour des raisons suffisantes de la rejeter).

♠ Exemple 6 avec rédaction-type pour ce genre d'exercices. Algues toxiques

Dans un pays, 10 % des plages étaient atteintes par des algues toxiques. On a modifié le processus de rejets chimiques : on admet que le nouveau processus de rejet, très différent du précédent, pourrait modifier cette proportion.
On prend un échantillon aléatoire de 150 prélèvements, on constate que 18 présentent des traces d'algues toxiques. Peut-on penser que le nouveau traitement a un impact sur le pourcentage de plages polluées ?

Corrigé . Utilisation classique de l'intervalle de fluctuation: rejet ou non d'une hypothèse sur une proportion.
On veut tester l'hypothèse p =10%. On étudie un échantillon de 150 plages donc $n=500$. Pour un échantillon de cette taille, l'intervalle de l'intervalle de fluctuations asymptotique au seuil de 95% est $IFA_{150} = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] = \left[0{,}1 - 1{,}96\frac{\sqrt{0{,}1\times 0{,}9}}{\sqrt{150}} ; 0{,}1 + 1{,}96\frac{\sqrt{0{,}1\times 0{,}9}}{\sqrt{150}} \right] \approx [5{,}2\% ; 14{,}8\%]$.
Comme les conditions $n \geqslant 30, np \geqslant 5$ et $n(1-p) \geqslant 5$ sont remplies, pour environ 95% des échantillons de 150 plages tirés au hasard, la fréquence observée devrait être dans cet intervalle. Or la fréquence observée lors de l'étude $f = \frac{18}{150} = 0{,}12$ appartient à cet intervalle. On en déduit qu'au seuil que 95% on ne peut pas rejeter l'hypothèse initiale : le hasard peut à lui seul expliquer la différence entre les valeurs 10% de p et 12% de la fréquence.

[2] *« Probablement » car on travaille au seuil de 95% et pas 100%.*
[3] *« rares » car avec le choix du seuil de 95%, il y a au plus 5% des échantillons pour lesquels f n'est pas dans l'intervalle de fluctuations.*

D. Lien avec l'intervalle de fluctuation vu en seconde

■ Remarquons tout d'abord que l'intervalle de fluctuation asymptotique I_n **est contenu dans l'intervalle de fluctuation « simplifié » vu en seconde** : $J_n=\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$.

En effet, en faisant l'étude sur $]0,\ 1[$ du trinôme $p\mapsto p(1-p)$, on constate que $\forall\ p\in]0,1[,\ p(1-p)\leqslant\frac{1}{4}$ d'où $\forall\ p\in]0,1[,\ \sqrt{p(1-p)}\leqslant\sqrt{\frac{1}{4}}=\frac{1}{2}$.

De plus $1{,}96<2$ donc en multipliant membre à membre ces deux inégalités où tous les termes dont positifs on obtient $\forall\ p\in]0,1[,\ 1{,}96\sqrt{p(1-p)}<2\times\frac{1}{2}=1$ d'où $I_n\subset J_n$.

[P9] Au moins 95% des fréquences dans l'intervalle de fluctuation « simplifié » pour n assez grand

p étant fixé dans $]0;1[$, il existe $n_0\in\mathbb{N}$ tel que $\forall\ n\geqslant n_0,\ P\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right)\geqslant 95\%$.

Cela signifie pour n assez[4] grand, l'intervalle de fluctuation « simplifié » $J_n=\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$ contient la fréquence $F_n=\frac{X_n}{n}$ avec une probabilité **supérieure ou égale à 95%**.

Démonstration **ROC exigible**: *On prend* $1-\alpha=P(-2\leq Z\leq 2)\approx 95{,}4\%$ *(où Z est suit une loi normale centrée réduite[5]) d'où* $u_\alpha=2$. *On introduit une nouvelle suite d'intervalles aléatoires* $K_n=\left[p-2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p+2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$. *Par le même raisonnement que précédemment, on prouve que* $K_n\subset J_n$ *puisque* $\forall\ p\in]0,1[,\ 2\sqrt{p(1-p)}\leqslant 2\times\frac{1}{2}=1$. *(En fait, on a même* $I_n\subset K_n\subset J_n$*).*

• *Or [P6] nous dit que pour p fixé dans* $]0;1[$ *l'intervalle* K_n *contient la fréquence* $F_n=\frac{X_n}{n}$ *avec une probabilité qui tend vers* $1-\alpha\approx 95{,}4\%$ *lorsque n tend vers l'infini. Finalement,* $\lim\limits_{n\to+\infty} P(F_n\in K_n)\approx 0{,}954$.

• *Puisque* $\lim\limits_{n\to+\infty} P(F_n\in K_n)\approx 0{,}954$, *par définition de la limite d'une suite, l'intervalle ouvert* $]0{,}95\,;0{,}96[$ *contient tous les termes de la suite à partir d'un certain rang* n_0. *Ceci entraîne que* $\forall\ n\geqslant n_0,\ P(F_n\in K_n)>95\%$. *Or* $K_n\subset J_n$ *donc* $\forall n n\geqslant n_0,\ P(F_n\in J_n)\geqslant P(F_n\in K_n)>95\%$.

III. Estimation : Intervalle de confiance de p

On l'utilise quand p, **la proportion dans l'ensemble de la population, n'est PAS connue** et que l'**on veut l'estimer** à l'aide d'un échantillon. C'est ce qui se passe pour les sondages: On observe la fréquence du caractère dans l'échantillon et on en déduit un encadrement de la fréquence du caractère dans la population.

[P10] Théorème de l'intervalle de confiance avec un niveau de confiance de plus de 95%

p étant fixé dans $]0;1[$, il existe $n_0\in\mathbb{N}$ tel que $\forall\ n\geqslant n_0, P\left(F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\right)\geqslant 95\%$.

Cela signifie pour n assez grand, dans **au moins 95%** des cas, la proportion ***p*** dans la population (totale) vérifie $p\in\left[f-\frac{1}{\sqrt{n}}\,;\ f+\frac{1}{\sqrt{n}}\right]$ où f est la fréquence observée dans l'échantillon. On dit que $\left[f-\frac{1}{\sqrt{n}}\,;\ f+\frac{1}{\sqrt{n}}\right]$ est un ***intervalle de confiance de la proportion inconnue p avec un niveau de confiance d'au moins 95%***

En pratique, on utilise cet intervalle de confiance dès que $n\geqslant 30, nf\geqslant 5$ et $n(1-f)\geqslant 5$.

[4] *Condition pas vraiment commode en termes d'applications pratiques.*

[5] *Rappelez-vous les valeurs remarquables données dans le cours sur la loi normale (ou sortez votre calculatrice): Quand on prend deux écart-types de part et d'autre de la moyenne, on a une probabilité d'environ 95,4 %.*

Dem: On peut réecrire la condition dans P9 : $p-\frac{1}{\sqrt{n}} \leqslant F_n \leqslant p+\frac{1}{\sqrt{n}} \Leftrightarrow F_n-\frac{1}{\sqrt{n}} \leqslant p \leqslant F_n+\frac{1}{\sqrt{n}}$.

Remarques:

- La précision de l'estimation, mesurée par l'amplitude de l'intervalle de confiance, est $\frac{2}{\sqrt{n}}$.
- La précision de l'estimation ne dépend que de la taille de l'échantillon, pas de celle de la population.
- En pratique on peut utiliser cet intervalle dès que $n \geqslant 30, nf \geqslant 5$ et $n(1-f) \geqslant 5$. Les hypothèse à vérifier pour utiliser le théorème portent sur f, la proportion observée dans l'échantillon (*et pas sur p. Heureusement! vu que p est inconnu, ce serait impossible à vérifier!*) et on obtient un encadrement de p.
- Un intervalle de confiance étant un intervalle numérique (= non aléatoire), il est incorrect de conclure la détermination d'un intervalle de confiance par une phrase du type « *p a une probabilité de 0,95 d'être entre $f-\frac{1}{\sqrt{n}}$ et $f+\frac{1}{\sqrt{n}}$* » car il n'y a plus d'aléatoire à ce stade : p est dans cet intervalle ou il ne l'est pas (le hic est qu'on ne sait pas s'il y est ou non). Il est en revanche convenable d'écrire : *« $\left[f-\frac{1}{\sqrt{n}};\ f+\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion inconnue p au niveau de confiance d'au moins 0,95 »*.

Table des matières

Sources : Le cours de M. Mugnier, le cours de Pierre Lux, le manuel math'x de 1S, le manuel Repères de TS, les commentaires de certains des membres de la liste de diffusion *mathlyc* (notamment Christian Vassard, Paul Beaurivage, Jean-Luc Giaco ? et bien d'autres) …. et mes cogitations (intenses pour ce chapitre!).

<u>NB</u>: Comment masquer ou afficher les discussions entre profs, les démonstrations, les exercices ou les paragraphes en préparation...etc (Les variables que j'utilise dépendent des documents)

- Dans la version Open Office de ce document, les **corrigés** par exemples (s'ils existent) sont visibles sauf quand la variable CORR prend la valeur M (« M » pour « Masqué »). Une variable est un champ particulier (de type texte) et se crée de la même façon : « Insérer » puis « champs ». Attention ! Il faut placer la variable AVANT les sections qu'elle pilote. Dans ce document, il y a aussi une variable DEM pour les démonstrations et une variable DP (Discussions entre profs) qui masquent les paragraphes correspondants quand elles valent M.
- Les différentes variables se pilotent en haut du document (Elles sont en grisé dans le titre du .odt et invisibles dans le pdf). Mettre le curseur à gauche de la bande grise, puis cliquer à droite et aller dans « champs ».
- Pour créer une section à masquer, sélectionner le texte à masquer, puis « insertion », puis «section » puis cliquer sur masquer : La condition s'écrit : CORR==« M » (Il faut les guillemets autour du M, un double égal et pas d'espaces).
- Pour faire réapparaître la section, changer la valeur de CORR à une autre valeur que M.
- Idem pour la variable EP (En Préparation) qui permet de masquer les exercices qui ne sont pas finis ou que j'envisage de mettre dans le DS. Elle vaut pour le moment EP=M et les sections correspondantes sont masquées quand EP=M.
- Dans ce document il y a aussi une variable DP = Discussion entre Profs, qui est un copier-coller des interventions des uns et des autres sur la liste « mathsLyc ».
- Quand un exercice ou un paragraphe est prêt on peut supprimer la section correspondante (pour qu'il soit visible tout le temps) avec « Format » puis « Sections »
- Évidemment dans le pdf cela ne marche pas, c'est tout l'intérêt....

Une mise en route du chapitre inoubliable !

Intro : La phrase « 60% de libanais dans mes tricheurs. »induit une réaction viscérale. Or il manque deux infos pour analyser cette donnée.

Activité L'affaire Castaneda contre Partida

http://dutarte.perso.neuf.fr/statistique/Jury%20et%20discrimation%20TS.htm