Loi des grands nombres et concentration

Introduction :

Ce cours s'inscrit dans la continuité du précédent sur les sommes de variables aléatoires, et il est fondé sur une propriété, très utile « techniquement » en probabilité et en statistique, appelée inégalité de Bienaymé-Tchebychev.
Celle-ci donne un moyen de contrôler l'écart entre les valeurs prises par une variable aléatoire et son espérance, en fonction de sa variance. Plus précisément, elle donne une majoration de la probabilité que l'écart soit grand. Elle sera donc l'objet de la première partie de ce cours.
Dans la deuxième partie, nous parlerons de l'inégalité de concentration, qui est obtenue grâce à l'inégalité de Bienaymé-Tchebychev appliquée à l'échantillon d'une variable aléatoire. Cette inégalité permet de déterminer la taille d'un échantillon en fonction de la précision (de l'écart) et du risque.
Elle permet également de donner une démonstration de la loi des grands nombres, qui sera étudiée dans la troisième partie.

1 Inégalité de Bienaymé-Tchebychev

Commençons donc par exprimer l'inégalité de Bienaymé-Tchebychev.

Propriété

Inégalité de Bienaymé-Tchebychev :

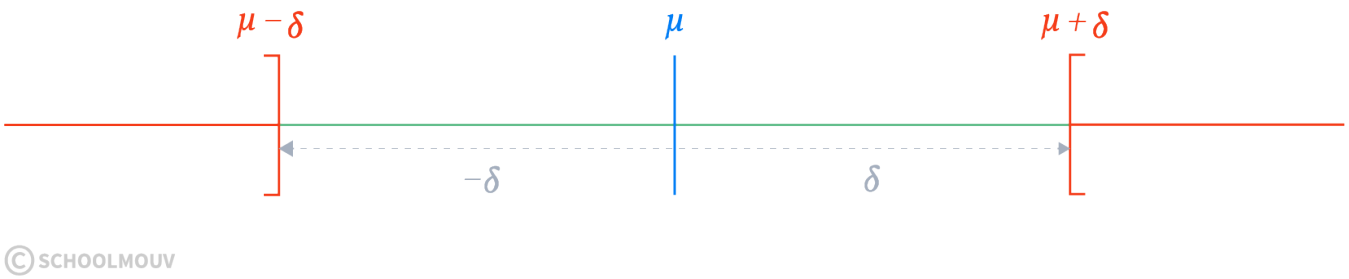
Soit X une variable aléatoire d'espérance μ et de variance V.
Pour tout réel δ strictement positif, on a :

$$p(|X - \mu| \geq \delta) \leq \frac{V}{\delta^2}$$

➜ La probabilité que l'écart entre X et son espérance soit supérieur ou égal à δ, est inférieure ou égale à $\frac{V}{\delta^2}$.

Donnons une petite représentation graphique pour bien comprendre ce que nous dit cette inégalité ; nous pourrons ainsi savoir quand et comment l'utiliser.

1. $|X - \mu|$ exprime la distance, ou l'**écart**, de X à μ.



Représentation de la distance de X à μ

Et nous nous intéressons au cas où cet écart est supérieur ou égal à δ :

$$|X - \mu| \geq \delta \Leftrightarrow \begin{cases} X - \mu \leq -\delta \\ \text{ou :} \\ X - \mu \geq \delta \end{cases}$$

$$\Leftrightarrow \begin{cases} X \leq \mu - \delta \\ \text{ou :} \\ X \geq \mu + \delta \end{cases}$$

$|X - \mu| \geq \delta$ signifie donc que X appartient alors aux intervalles représentés par les demi-droites rouges.

➜ L'inégalité de Bienaymé-Tchebychev majore ainsi la probabilité que l'écart entre X et μ soit supérieur à δ.

2. Considérons maintenant l'événement suivant :

$$|X - \mu| < \delta \Leftrightarrow -\delta < X - \mu < \delta$$
$$\Leftrightarrow \mu - \delta < X < \mu + \delta$$

Nous voyons sur le schéma que cela correspond au cas où X appartient à l'intervalle ouvert $]\mu - \delta\ ;\ \mu + \delta[$ (en vert).

Nous nous rendons aussi compte qu'il s'agit de l'événement contraire de $|X - \mu| \geq \delta$. Nous pouvons donc écrire :

$$p(|X - \mu| \geq \delta) = 1 - p(|X - \mu| < \delta)$$

À retenir

Nous obtenons donc :

$$1 - p(|X - \mu| < \delta) \leq \frac{V}{\delta^2}$$
$$\Leftrightarrow p(|X - \mu| < \delta) \geq 1 - \frac{V}{\delta^2}$$
$$\Leftrightarrow p(\mu - \delta < X < \mu + \delta) \geq 1 - \frac{V}{\delta^2}$$
$$\Leftrightarrow \ p(X \in]\mu - \delta \ ; \ \mu + \delta[) \geq 1 - \frac{V}{\delta^2}$$

➜ Grâce à l'inégalité de Bienaymé-Tchebychev, nous pouvons aussi minorer par $1 - \frac{V}{\delta^2}$ la probabilité que l'écart entre X et μ soit inférieur à un nombre donné δ.

Nous allons prendre un exemple pour voir comment interpréter un résultat obtenu avec l'inégalité de Bienaymé-Tchebychev.

Exemple

La consommation d'eau quotidienne, en litres, d'un·e Français·e pris·e au hasard dans la population est donnée par la variable aléatoire C d'espérance $\mu = 130$ et de variance $V = 750$.
Nous nous intéressons à la probabilité que l'écart entre C et l'espérance soit supérieur à 50 litres.

On applique l'inégalité de Bienaymé-Tchebychev avec $\delta = 50$:

$$p(|C - \mu| \geq 50) \leq \frac{V}{50^2} \Leftrightarrow p(|C - 130| \geq 50) \leq \frac{750}{50^2} = 0,3$$

Pour bien comprendre ce que cela signifie, donnons l'équivalence suivante :

$$|C - 130| \geq 50 \Leftrightarrow \begin{cases} C - 130 \leq -50 \\ \text{ou :} \\ C - 130 \geq 50 \end{cases}$$
$$= \begin{cases} C \leq 80 \\ \text{ou :} \\ C \geq 180 \end{cases}$$

➜ Ainsi, la probabilité qu'un·e Français·e pris·e au hasard ait une consommation inférieure à 80 litres ou supérieure à 180 litres est inférieure ou égale à $0,3$.

➜ Autrement dit, la probabilité qu'un·e Français·e pris·e au hasard ait une consommation comprise entre 80 et 180 litres est supérieure à $0,7$.

➜ Ou encore, si nous choisissons au hasard un·e Français·e, il y a au moins $70\,\%$ de chance pour qu'il ou elle consomme entre 80 et 180 litres d'eau par jour.

De ce que nous venons de dire, retenons les points suivants, auxquels il faudra penser lors de la résolution d'exercices.

À retenir

1. Pour majorer une probabilité d'un événement du type $|X - E(X)| \geq \delta$, on pense à appliquer l'inégalité de Bienaymé-Tchebychev.

➜ On calcule, si elles ne sont pas données, l'espérance et la variance de la variable aléatoire.

2. De manière équivalente, on peut avoir une minoration de la probabilité que X appartienne à un intervalle centré en $E(X)$.

➜ On peut interpréter concrètement les résultats dans les termes de l'énoncé en regardant l'intervalle $]E(X) - \delta\ ;\ E(X) + \delta[$.

Allons maintenant un peu plus loin : nous savons que l'écart-type σ d'une variable aléatoire X donne une indication sur la dispersion de X autour de son espérance μ.

➜ Il est donc logique d'étudier des écarts entre X et son espérance inférieurs ou supérieurs à quelques σ.

C'est ce que nous allons faire dans l'exemple suivant, en nous servant de l'inégalité de Bienaymé-Tchebychev.

Exemple

Soit la variable aléatoire X qui suit une loi binomiale de paramètres $n = 150$ et $p = 0,45$.

1. Commençons par calculer son espérance μ, sa variance V et son écart-type σ grâce aux propriétés de la loi binomiale que nous avons découvertes dans le cours précédent :

$$\begin{aligned}\mu &= np \\ &= 150 \times 0,45 \\ &= 67,5\end{aligned}$$

$$\begin{aligned}V &= np(1-p) \\ &= 150 \times 0,45 \times 0,55 \\ &= 37,125\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{V} \\ &= \sqrt{37,125}\end{aligned}$$

2 Appliquons l'inégalité de Bienaymé-Tchebychev avec $\delta = 2\sigma$, en remarquant que $V = \sigma^2$:

$$\begin{aligned}p(|X-\mu| \geq 2\sigma) &\leq \frac{\sigma^2}{(2\sigma)^2} \\ \Leftrightarrow p(|X-67,5| \geq 2\sqrt{37,125}) &\leq \frac{\sqrt{37,125}^2}{(2\sqrt{37,125})^2} \\ &\leq \frac{1}{4} = 0,25\end{aligned}$$

➜ Cela signifie que la probabilité que l'écart entre X et μ soit supérieur à 2σ est inférieure à $0,25$, ce qui revient à dire que la probabilité que l'écart entre X et μ soit inférieur à 2σ est supérieure à $0,75$.

3 Continuons maintenant d'appliquer cette inégalité avec des écarts de 3σ, 4σ, 5σ.

$$\begin{aligned}p(|X-67,5| \geq 3\sqrt{37,125}) &\leq \frac{\sqrt{37,125}^2}{(3\sqrt{37,125})^2} \\ &\leq \frac{1}{9} \approx 0,111\end{aligned}$$

$$\begin{aligned}p(|X-67,5| \geq 4\sqrt{37,125}) &\leq \frac{\sqrt{37,125}^2}{(4\sqrt{37,125})^2} \\ &\leq \frac{1}{16} = 0,0625\end{aligned}$$

$$\begin{aligned}p(|X-67,5| \geq 5\sqrt{37,125}) &\leq \frac{\sqrt{37,125}^2}{(5\sqrt{37,125})^2} \\ &\leq \frac{1}{25} = 0,04\end{aligned}$$

C Nous voyons que la probabilité d'avoir un écart supérieur à quelques σ ira en diminuant si on augmente le nombre de δ.

À retenir

Ainsi, de cet exemple, nous pouvons déduire que des écarts entre X et μ supérieurs à quelques σ deviennent improbables.

Attention

Nous avons aussi vu que la probabilité que l'écart entre X et μ soit supérieur à 2σ est majorée par $0,25$.
Toutefois, si l'on simule une telle expérience, on se rend compte que la probabilité d'avoir un écart supérieur à 2σ est souvent majorée par $0,05$.

➜ L'inégalité de Bienaymé-Tchebychev est vraie, mais elle n'est pas optimale car on peut encadrer/majorer plus précisément cette probabilité.

2 Inégalité de concentration

Nous allons maintenant appliquer l'inégalité de Bienaymé-Tchebychev à la moyenne d'un échantillon d'une variable aléatoire.

Considérons un échantillon de taille n de la variable aléatoire X, d'espérance μ et de variance V.
Et considérons M_n la variable aléatoire moyenne de cet échantillon.

En appliquant l'inégalité de Bienaymé-Tchebychev à M_n, avec δ un réel strictement positif, nous obtenons :

$$p(|M_n - E(M_n)| \geq \delta) \leq \frac{V(M_n)}{\delta^2}$$

Or, par les propriétés vues dans le cours précédent, nous avons :

$$E(M_n) = \mu$$
$$V(M_n) = \frac{V}{n}$$

Nous avons donc :

$$p(|M_n - \mu| \geq \delta) \leq \frac{\frac{V}{n}}{\delta^2}$$
$$\leq \frac{V}{n\delta^2}$$

➜ Nous obtenons une **inégalité** dite **de concentration**.

Propriété

Inégalité de concentration :

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de taille n de la variable aléatoire X, d'espérance μ et de variance V.
Soit M_n la variable aléatoire moyenne de cet échantillon :

$$M_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Alors, pour tout réel δ strictement positif, on a :

$$p(|M_n - \mu| \geq \delta) \leq \frac{V}{n\delta^2}$$

➜ À partir de cette inégalité, on dit qu'on obtient pour M_n une précision de δ et un risque de $\frac{V}{n\delta^2}$.

Comme dans la première partie, nous pouvons considérer l'événement contraire de $|M_n - \mu| \geq \delta$, à savoir : $|M_n - \mu| < \delta$.

À retenir

L'inégalité de concentration peut alors être écrite de la façon suivante :

$$p(|M_n - \mu| < \delta) \geq 1 - \frac{V}{n\delta^2}$$
$$\Leftrightarrow p(\mu - \delta < M_n < \mu + \delta) \geq 1 - \frac{V}{n\delta^2}$$
$$\Leftrightarrow p(M_n \in]\mu - \delta\ ;\ \mu + \delta[) \geq 1 - \frac{V}{n\delta^2}$$

Concrètement, l'inégalité de concentration nous dit :

- que nous pouvons majorer la probabilité que l'écart entre M_n et μ soit supérieur à un nombre donné δ ;

- que nous pouvons minorer la probabilité que l'écart entre M_n et μ soit inférieur à un nombre donné δ.

L'inégalité de concentration permet notamment de déterminer la taille d'un échantillon en fonction d'une précision et d'un risque fixés.
L'exemple suivant va nous donner une méthodologie pour résoudre de tels problèmes.

Exemple

Un institut politique a pour projet de constituer un échantillon de personnes tirées au sort parmi celles inscrites sur les listes électorales. Il s'intéresse plus particulièrement à leur participation au premier tour des dernières élections.
Sachant que la participation était de $45\,\%$ et afin d'avoir une représentativité la plus fidèle du corps électoral, l'institut souhaite connaître la taille minimale de l'échantillon à constituer pour être sûr, au moins à $95\,\%$, que le taux de participation en son sein sera compris entre $41\,\%$ et $49\,\%$.
Nous considérons en outre que le nombre d'inscrits sur les listes électorales est suffisamment grand pour que la constitution de l'échantillon soit assimilée à un tirage avec remise.

1. Soit la variable aléatoire X qui, pour chaque personne tirée au sort, prend la valeur 1 si elle a voté et 0 si elle s'est abstenue.

La variable X suit une loi de Bernoulli, de paramètre $p = 0,45$.

➜ Nous en déduisons l'espérance μ et la variance V :

$$\begin{aligned} \mu &= p \\ &= 0,45 \\ V &= p(1-p) \\ &= 0,45 \times 0,55 \\ &= 0,2475 \end{aligned}$$

2. Traduisons mathématiquement l'énoncé.

Considérons d'abord $(X_1,\ X_2,\ \ldots,\ X_n)$ un échantillon de taille n de la variable aléatoire X.
Considérons ensuite M_n, la variable aléatoire moyenne de l'échantillon.

➜ Elle donnera le taux de participation dans cet échantillon.

Nous souhaitons donc trouver n tel que la probabilité que M_n soit comprise entre $0,41$ ($41\,\%$) et $0,49$ ($49\,\%$) est supérieure ou égale à $0,95$ ($95\,\%$) :

$$p(0,41 < M_n < 0,49) \geq 0,95$$

3 Cela doit nous faire penser immédiatement à l'inégalité de concentration.

Il faut donc commencer par faire apparaître l'écart entre M_n et l'espérance $\mu = 0,45$ de X :

$$\begin{aligned} 0,41 < M_n < 0,49 &\Leftrightarrow 0,41 - \mu < M_n - \mu < 0,49 - \mu \\ &\Leftrightarrow 0,41 - 0,45 < M_n - 0,45 < 0,49 - 0,45 \\ &\Leftrightarrow -0,04 < M_n - 0,45 < 0,04 \\ &\Leftrightarrow |M_n - 0,45| < 0,04 \end{aligned}$$

➜ Nous obtenons donc :

$$p(0,41 < M_n < 0,49) = p(|M_n - 0,45| < 0,04) \geq 0,95$$

4 Ensuite, pour nous rapprocher de l'inégalité de concentration, nous allons nous intéresser à l'événement contraire : $|M_n - 0,45| \geq 0,04$:

$$p(|M_n - 0,45| < 0,04) = 1 - p(|M_n - 0,45| \geq 0,04)$$

Nous obtenons alors :

$$\begin{aligned} & p(|M_n - 0,45| < 0,04) \geq 0,95 \\ \Leftrightarrow\ & 1 - p(|M_n - 0,45| \geq 0,04) \geq 0,95 \\ \Leftrightarrow\ & p(|M_n - 0,45| \geq 0,04) \leq 0,05 \end{aligned}$$

➜ M_n est obtenu avec une précision de $0,04$ et un risque de $5\,\%$.

5 Nous pouvons donc maintenant appliquer l'inégalité de concentration (avec $\delta = 0,04$) :

$$p(|M_n - 0,45| \geq 0,04) \leq \frac{V}{n \times 0,04^2}$$

Pour que cette probabilité soit inférieure à $0,05$, il suffit que $\frac{V}{n \times 0,04^2}$ soit inférieur à $0,05$.
On résout donc cette dernière inégalité pour estimer la taille n de l'échantillon qui correspond aux conditions de l'énoncé.

$$\begin{aligned} \frac{V}{n \times 0,04^2} \leq 0,05 &\Leftrightarrow \frac{0,2475}{n \times 0,04^2} \leq 0,05 \\ &\Leftrightarrow n \geq \frac{0,2475}{0,04^2 \times 0,05} \\ &\Leftrightarrow n \geq 3\,093,75 \end{aligned}$$

➜ n doit donc être un entier supérieur ou égal à $3\,093,75$, soit :

$$n \geq 3\,094$$

C Pour être sûr au moins à $95\,\%$ que, dans l'échantillon, le taux de participation au premier tour des élections les plus récentes sera compris entre $41\,\%$ et $49\,\%$, l'institut devra constituer un échantillon d'au moins $3\,094$ personnes.

Vérifions notre résultat en prenant un échantillon de $3\,100$ personnes. L'inégalité de concentration donne alors :

$$p(|M_{3\,100} - 0,45| \geq 0,04) \leq \frac{0,2475}{3\,100 \times 0,04^2}$$

Ce qui est équivalent à :

$$p(|M_{3\,100} - 0,45| < 0,04) \geq 1 - \frac{0,2475}{3\,100 \times 0,04^2} \approx 0,9501$$

➜ Dans un échantillon de $3\,100$ personnes, le taux de participation sera compris entre $41\,\%$ et $49\,\%$ avec un degré de confiance d'au moins $95\,\%$.

Remarquons que, si l'institut considérait que l'échantillon était trop grand à constituer, pour des raisons économiques ou d'organisation, il devrait accepter soit d'agrandir l'intervalle possible pour la moyenne – au risque d'avoir une majorité de votants dans l'échantillon, ce qui ne refléterait pas la réalité –, soit d'avoir un degré de confiance moindre.

Astuce

Dans l'exemple ci-dessus, nous sommes revenus à la majoration de la probabilité que l'écart soit grand, et ce afin de bien montrer toutes les étapes du raisonnement.
Nous aurions toutefois pu utiliser la deuxième écriture que nous avons donnée pour arriver directement à :

$$p(|M_n - \mu| < 0,04) \geq 1 - \frac{0,2475}{n \times 0,04^2}$$

Si on veut que cette probabilité soit supérieure à $0,95$, il suffit que $1 - \frac{0,2475}{n \times 0,04^2}$ soit supérieur à $0,95$.

➜ C'est cette dernière inéquation que l'on résout, et elle est équivalente à celle que nous avions obtenue :

$$1 - \frac{0,2475}{n \times 0,04^2} \geq 0,95 \Leftrightarrow \frac{0,2475}{n \times 0,04^2} \leq 0,05$$

3 Loi des grands nombres

Ce que nous venons de voir nous ramène à la loi des grands nombres, qui a été évoquée en seconde mais dont nous allons donner une expression plus formelle.

Propriété

Loi des grands nombres :

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de taille n de la variables aléatoire X, d'espérance μ.
Soit M_n la variable aléatoire moyenne de cet échantillon.
Alors, pour tout réel strictement positif δ :

$$\lim_{n \to +\infty} p(|M_n - \mu| \geq \delta) = 0$$

Donnons une démonstration de cette loi grâce à l'inégalité de concentration.

Démonstration

Soit V la variance de la variable aléatoire X.
Nous avons alors :

$$p(|M_n - \mu| \geq \delta) \leq \frac{V}{n\delta^2}$$

Par quotient des limites, V et δ étant des constantes :

$$\lim_{n \to +\infty} \frac{V}{n\delta^2} = 0$$

Or, $p(|M_n - \mu| \geq \delta)$ est une probabilité, elle est donc supérieure ou égale à 0.
Nous avons ainsi l'encadrement suivant :

$$0 \leq p(|M_n - \mu| \geq \delta) \leq \frac{V}{n\delta^2}$$

➜ Finalement, par le théorème d'encadrement (ou des gendarmes), nous obtenons :

$$\lim_{n \to +\infty} p(|M_n - \mu| \geq \delta) = 0$$

À retenir

Concrètement, la loi des grands nombres, fondamentale en probabilité et en statistique, nous dit que la moyenne d'un échantillon d'une variable aléatoire X se rapproche d'autant plus de son espérance que la taille n de l'échantillon est grande.

Pour conclure ce cours, précisons que nous avons donné la loi « faible » des grands nombres.

- Il existe une loi « forte » des grands nombres, que nous n'avons pas abordée car elle dépasse le cadre de la terminale, et vous la découvrirez sans doute durant vos études supérieures.

Conclusion :

Ce cours nous a permis de découvrir des formules d'une importance majeure en probabilité : l'inégalité de Bienaymé-Tchebychev, dont nous avons déduit l'inégalité de concentration.
Et nous avons conclu notre chapitre avec la loi fondamentale de la théorie des probabilités : la loi des grands nombres. Celle-ci ouvre la porte à des applications très nombreuses, notamment en statistique, par exemple pour élaborer des sondages les plus fiables possibles.