PR12

Statistiques inférentielles

Version du 27-01-2023 à 10:33

1. Loi faible des grands nombres

Définition 1 – Moyenne empirique

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles.

On appelle moyenne empirique, la variable aléatoire notée $\overline{X_n}$ définie par $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$.

□

Théorème 1 – Loi faible des nombres nombres

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes, admettant une même espérance m et une même variance σ^2.

En notant $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$ la moyenne empirique, on a :

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\left[\left|\overline{X_n} - m\right| \geq \varepsilon\right]\right) \underset{n\to+\infty}{\longrightarrow} 0$$

On dit que $\overline{X_n}$ converge en probabilité vers la variable constante égale à m.

Interprétation

Lorsque l'on répète un grand nombre de fois la même expérience, la moyenne des résultats de ces expériences a une grande probabilité d'être proche de l'espérance de notre expérience.

□

Éléments de preuve:

La linéarité de l'espérance donne que : $\mathbb{E}\left(\overline{X_n}\right) =$

De plus, les variables aléatoires X_1, …, X_n étant supposées indpendantes, on a :

$$\mathbb{V}\left(\overline{X_n}\right) =$$

D'apès l'inégalité de Bienaymé-Tchebychev on a alors :

2. Intervalle de confiance

Introduction – Des statistiques aux probabilités

On souhaite indiquer sur l'emballage d'une ampoule basse consommation sa durée de vie moyenne « sans trop d'erreur ». Comment s'y prendre ?

Contrôle qualité

On a naïvement branché et allumé 250 ampoules, et obtenus les résultats statistiques suivant :

- Durée de vie moyenne : 450 jours ;
- Écart-type de la série statistique : 70 jours.

Il va falloir attendre un certain temps...et il n'y a pas vraiment de certitude sur ce que l'on peut annoncer.

Modélisation

Le prélèvement au hasard d'une ampoule dans la production et regarder sa durée de vie peut être assimilée à la définition d'une variable aléatoire X dont on ne connaît pas particulièrement sa loi.
Lorsque l'on procède au prélèvement de 250 ampoules, cela revient donc à considérer 250 variables aléatoires X_1, ..., X_{250}.

On peut légitimement penser que X et X_1, ..., X_{250} suivent la même loi.

Les calculs du contrôle qualité

La première ampoule a duré 473 jours.
La deuxième ampoule a duré 424 jours.
etc.
La 250e ampoule a duré 463 jours.
et donc la durée moyenne $\overline{m}$ de vie d'une ampoule a été :

$$\begin{aligned}\overline{m} &= \frac{473+424+\ldots+463}{250}\\ &= 450\end{aligned}$$

Modélisation

On note $\overline{X}_{250}$ la variable aléatoire définie par :

$$\overline{X}_{250} = \frac{X_1+\ldots+X_{250}}{250}$$

et on s'intéresse à $\mathbb{E}\left(\overline{X}_{250}\right)$.
On remarquera en particulier que par linéarité de l'espérance, et puisque toutes les variables aléatoires X_1, ..., X_{250} sont de même loi et donc de même espérance, que $\mathbb{E}\left(\overline{X}_{250}\right) = \mathbb{E}\left(X_1\right)$.

Ce que l'on veut indiquer sur la boîte

Compte-tenu de l'étude précédente, le responsable production veut écrire sur l'emballage que la durée de vie moyenne d'une de ses ampoules est de 450 jours.

Le fait-il en prenant un risque supérieur à 5% ?

Modélisation

L'inégalité de Bienaymé-Tchebychev permet d'écrire que :

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\left[\left|\overline{X}_{250} - \mathbb{E}\left(X_1\right)\right| > \varepsilon\right]\right) \leq \frac{\mathbb{V}\left(X_1\right)}{250\varepsilon^2}$$

ce qui permet d'obtenir que :

$$\mathbb{P}\left(\left[\overline{X}_{250} - \varepsilon \leq \mathbb{E}\left(X_1\right) \leq \overline{X}_{250} + \varepsilon\right]\right) > 1 - \frac{\mathbb{V}\left(X_1\right)}{250\varepsilon^2}$$

Réglagles de l'outil de production

Le responsable production affirme que l'outil de production ne se dérègle globalement pas lorsqu'il est en fonctionnement, et ainsi que l'écart-type de la série statistique précédente reste constant à 70 s'il relance une production de 250 ampoules.

Modélisation

Avec l'hypothèse que $\mathbb{V}\left(X_1\right) = 70^2$ et en choisissant ε tel que $1 - \frac{\mathbb{V}\left(X_1\right)}{250\varepsilon^2} = 0,95$, on aura donc que puisque sur l'échantillon observé on a $\mathbb{E}\left(\overline{X}_{250}\right) = 450$:

$$\mathbb{P}\left(\left[450 - \varepsilon \leq \mathbb{E}\left(X_1\right) \leq 450 + \varepsilon\right]\right) > 0,95$$

Avec $\varepsilon \approx 0,051$, on a alors :

$$\mathbb{P}\left(\left[449{,}949 \leq \mathbb{E}\left(X_1\right) \leq 450{,}051\right]\right) \geq 0,95$$

□

Théorème 2 – Localisation de l'espérance

Soient X une variable aléatoire et $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et de même loi que X.

La moyenne empirique $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$ des n variables aléatoires $(X_1, \ldots, X_n)$ est telle que $E\left(\overline{X_n}\right) = \mathbb{E}(X)$ et $\mathbb{V}\left(\overline{X_n}\right) = \dfrac{\mathbb{V}(X)}{n}$.

Pour tout $\alpha \in]0;1[$ la probabilité que l'intervalle $\left[\overline{X_n} - \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}}; \overline{X_n} + \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}}\right]$ contienne $\mathbb{E}(X)$ est supérieure ou égale à $1-\alpha$.

Cela revient à dire que : $\mathbb{P}\left(\left[\overline{X_n} - \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}} \leq \mathbb{E}(X) \leq \overline{X_n} + \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}}\right]\right) \geq 1-\alpha$

Intervalle de confiance

L'intervalle $\left[\overline{X_n} - \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}}; \overline{X_n} + \sqrt{\dfrac{\mathbb{V}(X)}{n\alpha}}\right]$ est appelé intervalle de confiance pour $\mathbb{E}(X)$ au seuil ou au niveau $1-\alpha$.

□

Éléments de preuve:

La moyenne empirique $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$ des n variables aléatoires $(X_1, \ldots, X_n)$ est par linéarité :

$$\mathbb{E}\left(\overline{X_n}\right) =$$

et par indépendance on a que : $\mathbb{V}\left(\overline{X_n}\right) =$.

D'après l'inégalité de Bienaymé-Tchebychev, on peut écrire : $\forall \varepsilon > 0,\ \mathbb{P}\left(\left[\left|\overline{X_n} - \mathbb{E}(X)\right| \geq \varepsilon\right]\right) \leq \dfrac{\mathbb{V}\left(\overline{X_n}\right)}{\varepsilon^2}$

ou encore que : $\forall \varepsilon > 0,\ \mathbb{P}\left(\left[\left|\overline{X_n} - \mathbb{E}(X)\right| \geq \varepsilon\right]\right) \leq$

Comme $\left[\left|\overline{X_n} - \mathbb{E}(X)\right| > \varepsilon\right] \subset \left[\left|\overline{X_n} - \mathbb{E}(X)\right| \geq \varepsilon\right]$, on a $\mathbb{P}\left(\left[\left|\overline{X_n} - \mathbb{E}(X)\right| > \varepsilon\right]\right) \leq$

Ainsi :

$$\left(\mathbb{P}\left(\left[\left|\overline{X_n} - \mathbb{E}(X)\right| > \varepsilon\right]\right) \leq \frac{\mathbb{V}(X)}{n\varepsilon^2}\right) \quad \Leftrightarrow$$

Comme $\left[\left|\overline{X_n} - \mathbb{E}(X)\right| \leq \varepsilon\right] =$, pour $\varepsilon =$, on en déduit que :

$$\mathbb{P}\left(\left[\overline{X_n} - \sqrt{\frac{\mathbb{V}(X)}{n\alpha}} \leq \mathbb{E}(X) \leq \overline{X_n} + \sqrt{\frac{\mathbb{V}(X)}{n\alpha}}\right]\right) \geq 1-\alpha$$

3. Cas d'une variable de Bernoulli

Introduction – Statistiques et probabilités

Une partie du travail du statisticien consiste, à partir de données observées sur des échantillons d'une population pour un caractère donné, à proposer un modèle théorique permettant d'étudier la répartition de ce caractère sur l'ensemble de la population.
La plupart du temps cela conduit à proposer un modèle probabiliste, autrement dit à assimiler la répartition de ce caractère dans la population à un phénomène aléatoire qui suit une loi de probabilité donnée.

Quelle proportion ?

On dispose d'une urne contenant des boules rouges et des boules blanches, mais on ne connaît pas la composition de l'urne. En 1000 tirages avec remise, on obtient 300 boules rouges et 700 boules blanches.

À combien peut-on estimer « sans trop se tromper » la proportion de boules rouges dans l'urne ?

□

Proposition 1 – Majoration de la variance d'une loi de Bernoulli

Si X suit une loi de Bernoulli de paramètre $p \in [0;1]$, alors $\mathbb{V}(X) \leq \dfrac{1}{4}$.

□

Proposition 2 – Intervalle de confiance pour une variable aléatoire de Bernoulli

Si X suit une loi de Bernoulli et $(X_n)_{n\geq 1}$ est une suite de variables aléatoires de même loi que X, alors pour tout $n \in \mathbb{N}^*$ et $\alpha \in]0;1[$ $\mathbb{P}\left(\left[\overline{X_n} - \dfrac{1}{2\sqrt{n\alpha}} \leq \mathbb{E}(X) \leq \overline{X_n} + \dfrac{1}{2\sqrt{n\alpha}}\right]\right) \geq 1-\alpha$.

Interprétation

Dans le cas où $n = 1000$, au seuil de confiance de 90 %, on a :

$$\mathbb{P}\left(\left[\overline{X_{1000}} - \frac{1}{2\sqrt{1000\times 0,1}} \leq \mathbb{E}(X) \leq \overline{X_{1000}} + \frac{1}{2\sqrt{1000\times 0,1}}\right]\right) \geq 0,9$$

ce qui donne comme $\dfrac{1}{2\sqrt{1000\times 0,1}} = \dfrac{1}{20}$: $\mathbb{P}\left(\left[\overline{X_{1000}} - 0,05 \leq \mathbb{E}(X) \leq \overline{X_{1000}} + 0,05\right]\right) \geq 0,9$.

Dans ce cas, au seuil de confiance de 90 %, l'incertitude sur $\mathbb{E}(X)$ est de 5 %.

□

Éléments de preuve:

4. Variables aléatoires bornées ou à variance bornée

Proposition 3 – Majoration de la variance pour une variable aléatoire bornée et intervalle de confiance

Si X est une variable aléatoire bornée par $M \geq 0$, alors X admet une variance qui est majorée par M^2.

□

Éléments de preuve:
L'existence d'un moment d'ordre 2, d'une espérance et d'une variance est alors assurée par le caractère borné de X en revenant à la définition de ces objets.

Puisque $|X| \leq M$, on a et par croissance de l'espérance que .

On en déduit alors que :

Proposition 4 – Intervalle de confiance pour une variable aléatoire bornée

Si X est bornée par M et $(X_n)_{n\geq 1}$ est une suite de variables aléatoires de même loi que X,
alors pour tout $n \in \mathbb{N}^*$ et $\alpha \in]0;1[$ $\mathbb{P}\left(\left[\overline{X_n} - \frac{M}{\sqrt{n\alpha}} \leq \mathbb{E}(X) \leq \overline{X_n} + \frac{M}{\sqrt{n\alpha}}\right]\right) \geq 1 - \alpha.$

□

Éléments de preuve:
On adapte la démonstration faite pour le cas d'une loi de Bernoulli.