

---

## TP8 - Intervalle de confiance pour le paramètre d'une loi de Bernoulli et test de conformité à la moyenne

---

### 1 Simuler une variable aléatoire de loi de Bernoulli

1. Importer le module `random`. Que fait la commande `random()` ? Comment simuler une variable aléatoire qui suit une loi de Bernoulli de paramètre  $p$  à partir de cette fonction ?
2. Simuler avec une fonction `Bern(p)` une réalisation d'une variable aléatoire  $X$  qui suit une loi de Bernoulli de paramètre  $p$ .
3. On imagine maintenant que l'on dispose d'une fonction pour simuler la variable aléatoire  $X$ , dont on sait qu'elle suit une loi de Bernoulli, mais qu'on ne connaît pas le paramètre  $p$ . Que pourrait-on calculer pour estimer ce paramètre ?

### 2 La moyenne empirique

**Définition 2.0.1.** Soit  $n \in \mathbb{N}^*$  fixé et soient  $(X_1, X_2, \dots, X_n)$  des variables aléatoires mutuellement indépendantes qui suivent une même loi de Bernoulli de paramètre  $p$ .  
La **moyenne empirique** associée est :

$$M_n = \frac{1}{n} \sum_{k=1}^n X_k.$$

4. Pour tout  $n \in \mathbb{N}^*$ , calculer l'espérance et la variance de  $M_n$ . Qu'en déduit-on ?
5. On simule la variable aléatoire  $X$  grâce à la fonction `Bern(1/3)`. Programmer une fonction `M(n)` qui simule une réalisation de la variable aléatoire  $M_n$ .
6. Tester cette fonction pour différentes valeurs de  $n$ . Le résultat est-il proche de celui attendu ?

### 3 Intervalle de confiance

On cherche maintenant à donner non plus simplement une approximation de  $p$  mais un intervalle dans lequel on est sûr à 95% de trouver  $p$ . C'est ce qu'on appelle un **intervalle de confiance** du paramètre  $p$ . Pour ça, on a besoin d'avoir une valeur approchée de la variance de  $X$ .

**Définition 3.0.1.** Soient  $n \in \mathbb{N}^*$  et  $(X_1, X_2, \dots, X_n)$  des variables aléatoires mutuellement indépendantes qui suivent une même loi de Bernoulli de paramètre  $p$ . La **variance empirique**  $S_n^2$  associée est la variable aléatoire :

$$S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - M_n)^2 \quad \text{où} \quad M_n = \frac{1}{n} \sum_{k=1}^n X_k$$

7. Montrer que pour tout  $n \in \mathbb{N}^*$ , on a  $S_n^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - (M_n)^2$ .
8. Pour tout  $n \in \mathbb{N}^*$ , déterminer l'espérance de  $S_n^2$ . Montrer que  $\lim_{n \rightarrow +\infty} E(S_n^2) = \text{Var}(X)$ .
9. Programmer une fonction `S2(n)` qui simule une réalisation de la variable aléatoire  $S_n^2$ .

**Théorème 3.0.1.** Soit  $(X_n)_{n \geq 1}$  une suite de variables aléatoires mutuellement indépendantes qui suivent une même loi de Bernoulli de paramètre  $p$ . Pour tout  $n \in \mathbb{N}^*$ , on pose :

$$M_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{et} \quad S_n = \sqrt{\frac{1}{n} \sum_{k=1}^n (X_k - M_n)^2}$$

On a :

$$\lim_{n \rightarrow +\infty} \mathbf{P} \left( M_n - 1.96 \frac{S_n}{\sqrt{n}} < p < M_n + 1.96 \frac{S_n}{\sqrt{n}} \right) = 0.95.$$

Ainsi, l'**intervalle de confiance** du paramètre  $p$  est :

$$\left] M_n - 1.96 \frac{S_n}{\sqrt{n}}; M_n + 1.96 \frac{S_n}{\sqrt{n}} \right[.$$

10. Programmer une fonction **Intervalle(n)** qui donne l'intervalle de confiance du paramètre  $p$  à  $n$  fixé.
11. Tester ce programme pour  $n = 100$ .
12. Quel ordre de grandeur faut-il choisir pour  $n$  pour diminuer la largeur de l'intervalle par 2 ?
13. Quel ordre de grandeur faut-il choisir pour  $n$  pour avoir une approximation de  $p$  à  $10^{-3}$  près ?

## 4 Test de conformité à la moyenne

Soient  $X$  une variable aléatoire d'espérance  $\mu$  (inconnue) et  $\mu_0 \in \mathbb{R}$ . On voudrait savoir si l'espérance de  $X$  (i.e.  $\mu$ ) est égale à  $\mu_0$  ou non. Pour cela, on fait ce qu'on appelle un **test de conformité sur la moyenne**. On émet pour cela l'hypothèse suivante :

$$(H_0) : \mu = \mu_0$$

et on travaille sous cette hypothèse. On se demande comment faire pour accepter ou rejeter cette hypothèse avec un certain degré de certitude.

Le protocole du test est le suivant :

1. On émet l'hypothèse  $(H_0) : \mu = \mu_0$
2. On pose  $\alpha = 0,05$  : on prend un risque de 5% de rejeter l'hypothèse alors qu'elle était vraie.
3. On calcule la valeur de  $\frac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$  pour l'échantillon choisi.
4. Si la valeur obtenue n'appartient pas à  $[-1.96, 1.96]$ , on rejette l'hypothèse.  
Sinon on l'accepte.

Il se base sur le théorème suivant :

**Proposition 4.0.1.** Soit  $(X_n)_{n \geq 1}$  une suite de variables aléatoires mutuellement indépendantes et de même loi, admettant une espérance  $\mu$  et une variance  $\sigma^2$  non nulle. Pour tout  $n \in \mathbb{N}^*$ , on pose :

$$M_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{et} \quad S_n = \sqrt{\frac{1}{n} \sum_{k=1}^n (X_k - M_n)^2}$$

On a alors :

$$\lim_{n \rightarrow +\infty} \mathbf{P} \left( \left| \frac{M_n - \mu}{\frac{S_n}{\sqrt{n}}} \right| > 1.96 \right) = 0.05.$$

Dans un centre avicole, des études antérieures ont montré que la masse d'un œuf choisi au hasard peut être considérée comme la réalisation d'une variable aléatoire gaussienne  $X$  de moyenne  $m$  et de variance  $\sigma^2$ . On admet que les masses des œufs sont indépendantes les unes des autres. On prend un échantillon de  $n = 36$  œufs que l'on pèse. Les mesures sont données dans le tableau suivant :

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| 50,34 | 52,62 | 53,79 | 54,99 | 55,82 | 57,67 |
| 51,41 | 53,13 | 53,89 | 55,04 | 55,91 | 57,99 |
| 51,51 | 53,28 | 54,63 | 55,12 | 55,95 | 58,10 |
| 52,07 | 53,30 | 54,76 | 55,24 | 57,05 | 59,30 |
| 52,22 | 53,32 | 54,78 | 55,28 | 57,18 | 60,58 |
| 52,38 | 53,39 | 54,93 | 55,56 | 57,31 | 63,15 |

Vous trouverez la liste correspondante dans le fichier Liste.py sur CDP.

14. Calculer avec Python  $M_n$  et  $S_n$  pour cette série statistique.
15. Tester si la moyenne de cette variable aléatoire  $X$  est égale à  $\mu_0 = 55.95$ .

## 5 Test de Student de la moyenne

Soient  $X$  une variable aléatoire suivant une loi normale d'espérance  $\mu$  (inconnue) et  $\mu_0 \in \mathbb{R}$ . On voudrait savoir si l'espérance de  $X$  (*ie.*  $\mu$ ) est égale à  $\mu_0$  ou non. Pour cela, on fait ce qu'on appelle un **test de conformité Student**. On émet l'hypothèse suivante :

$$(H_0) : \mu = \mu_0$$

et on travaille sous cette hypothèse. On se demande comment faire pour accepter ou rejeter cette hypothèse avec un certain degré de certitude.

Le protocole du test est le suivant :

1. On émet l'hypothèse  $(H_0) : \mu = \mu_0$
2. On se donne un réel  $\alpha \in ]0, 1[$  assez petit (en pratique, on a souvent  $\alpha = 0,05$ ).
3. On calcule la valeur de  $\frac{M_n - \mu_0}{\frac{S_n^*}{\sqrt{n}}}$  pour l'échantillon choisi avec

$$M_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{et} \quad S_n^* = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (X_k - M_n)^2}$$

4. Si la valeur obtenue n'appartient pas à  $[-t, t]$ , on rejette l'hypothèse.  
Sinon on l'accepte. ( $t$  est calculé à partir de  $\alpha$  et de  $n$ )

On reprend l'exemple précédent sur le poids des œufs.

16. Calculer avec Python  $S_n^*$  pour cette série statistique.
17. Tester avec le test de Student si la moyenne de cette variable aléatoire  $X$  est égale à  $\mu_0 = 55.95$  avec  $\alpha = 0.05$ , sachant que cela donne une valeur de  $t \approx 2.03$  pour  $n = 36$ .
18. Comparer ce résultat avec celui obtenu pour l'autre test de la moyenne.

## 6 Pour aller plus loin

On veut comprendre d'où vient de lien entre le 1.96 annoncé dans les parties 3 et 4 et la certitude à 95%.

19. Montrer que  $P\left(M_n - 1.96 \frac{S_n}{\sqrt{n}} < \mu < M_n + 1.96 \frac{S_n}{\sqrt{n}}\right) = P(-1.96 \leq M_n^* \leq 1.96)$   
où  $M_n^* = \sqrt{n} \frac{M_n - \frac{1}{3}}{S_n}$ .

On cherche à déterminer la valeur de  $u$  pour laquelle

$$P(-u \leq M_n^* \leq u) = 0.95.$$

Notre objectif est de montrer que  $u \approx 1/96$ .

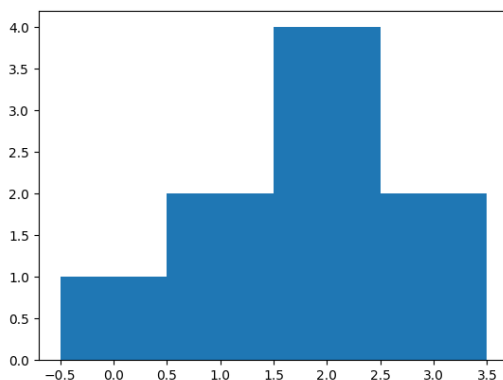
20. Définir une fonction `Mstar(n)` qui réalise une simulation de  $M_n^*$ .  
21. Définir une fonction `Liste(n)` qui crée une liste de 10000 simulations de  $M_n^*$ .  
22. Grâce au programme suivant, réaliser un histogramme associé à la variable aléatoire  $M_n^*$  pour  $n = 500$ .

```
1 n=500
2 L=Liste(n)
3 b=np.linspace(-3.5,3.5,100) #On dessine 100 rectangles
4 res=plt.hist(L,b,density=False)
5 plt.show()
```

Si tout va bien, vous devez constater que la variable aléatoire  $M_n^*$  est symétrique.

23. En déduire que  $P(-u \leq M_n^* \leq u) = 0.95 \iff P(0 \leq M_n^* \leq u) = 0.475$ .

Dans la suite, nous aurons besoin des informations contenues dans `res`. Le vecteur `res` contient dans sa première composante les hauteurs des rectangles, autrement dit les effectifs, et dans sa deuxième composante les coordonnées des bases des rectangles. Pour l'histogramme suivant, on aurait :



```
>>> res[0]
array([1., 2., 4., 2.])

>>> res[1]
array([-0.5, 0.5, 1.5, 2.5, 3.5])
```

24. On a simulé 10000 réalisations de  $M_n^*$ , donc environ 5000 sont au-dessus de 0. On veut savoir à quel moment l'effectif atteint 95% de l'effectif total de 5000. Pour ça, compléter la fonction suivante.

```
1 S=...
2 k=49 #On part de 0
3 while S<... :
4     S+=res[0][k] #On calcule les effectifs cumulés jusqu'à atteindre 95% du total
5     k+=1
6 print(...)
```

25. Répéter plusieurs fois cette opération et calculer la moyenne des valeurs obtenues lors des répétitions. Le résultat que vous trouvez doit être proche de 1.96.