

## Statistiques inférentielles

Faire une estimation sur une population à partir d'un échantillon.

### Loi des moyennes

Dans une population de moyenne  $\mu$  et d'écart-type  $\sigma$ , on prélève un échantillon de taille  $n$ .

On s'intéresse à la moyenne obtenue dans l'échantillon. L'échantillon était pris au hasard, la moyenne est également aléatoire. On note  $\bar{X}_n$  la variable aléatoire représentant la moyenne obtenue dans l'échantillon.

$\bar{X}_n$  suit **approximativement** une loi  $\mathcal{N}$  de moyenne  $\mu$  et écart-type  $\frac{\sigma}{\sqrt{n}}$



Dans cette loi comme dans d'autres à suivre, il n'est pas nécessaire que la population suive une loi normale! La loi  $\mathcal{N}$  apparaît en moyenne quelle que soit la loi suivie par  $X$ . Bien sûr, l'approximation est meilleure si  $n$  plus grand.

Dans le cas particulier où  $X$  suit une loi  $\mathcal{N}$ , alors le loi de  $\bar{X}_n$  est aussi  $\mathcal{N}$  (pas une approximation dans ce cas)

### Loi des fréquences

Dans la population, une proportion  $p$  d'individus possède une certaine caractéristique. On prélève au hasard un échantillon de taille  $n$ .

La fréquence d'individus ayant la caractéristique dans l'échantillon est aléatoire, on la note  $F$ .

$F$  suit approximativement la loi  $\mathcal{N}$  de moyenne  $p$  et d'écart-type  $\sqrt{\frac{p(1-p)}{n}}$ .

**Remarque 1 :** La loi des fréquences est un cas particulier de la loi des moyennes. En effet, on pourrait associer une variable  $X$  à tout individu prélevé au hasard. On dirait que  $X = 1$  si l'individu a la caractéristique demandée et  $X = 0$  sinon. Dans ce cas,  $X$  serait aléatoire, aurait une variable  $\mu = p$  et un écart-type  $\sigma = \sqrt{p(1-p)}$ . La fréquence  $F$  ne serait rien d'autre que  $\bar{X}_n$ . On retrouve alors la même chose pour les deux lois.

**Remarque 2 :** quand on dit par exemple qu'un intervalle de fluctuation asymptotique à 95 % est donné par  $p \pm 1,96\sqrt{\frac{p(1-p)}{n}}$ , cela coïncide avec la loi des fréquences.

### Estimation ponctuelle

Dans les lois précédentes, on connaissait la population et on faisait une prédiction sur le résultat d'un échantillon.

À présent, c'est l'inverse : on a un échantillon et on souhaite estimer quelque chose pour la population.

En gros, l'estimation ponctuelle consiste à estimer que la population a les mêmes caractéristiques que l'échantillon.

- Si on a obtenu la moyenne  $\bar{x}_e$  et l'écart-type  $\sigma_e$  dans l'échantillon, on estime que la population a la moyenne  $\bar{x}_e$  et l'écart-type  $s = \sigma_e \sqrt{\frac{n}{n-1}}$ .
- Si on a obtenu la fréquence  $f$  dans l'échantillon, on estime que la fréquence dans la population est  $p = f$ .

**Remarque 3 :** Si j'ai obtenu la moyenne  $\bar{x}_e = 12$  dans l'échantillon, j'estime que la moyenne est aussi 12 dans la population. Quand on dit que l'on **estime**, on a l'idée d'un à peu près, d'une marge d'erreur. L'estimation ponctuelle ne permet pas de chiffrer cette marge d'erreur.

**Remarque 4 :** J'ai noté  $\bar{x}_e$ , en minuscule. En effet, on parle ici de la moyenne dans l'échantillon **après** avoir fait le prélèvement. Ce n'est donc plus une variable aléatoire.

**Remarque 5 :** La formule pour l'écart-type nous dit que l'écart-type obtenu dans un échantillon a tendance à être plus petit que l'écart-type dans la population. En effet, l'écart-type représente la diversité. Il y aura plus de diversité dans la population que dans un échantillon. Il est d'usage de noter  $s$  l'écart-type estimé. Vos calculatrices, dans le mode statistiques, calculent deux écarts-types. Le plus petit est  $\sigma_e$  et le plus grand est  $s$ . Il n'est pas rare qu'on fasse simple en disant  $s = \sigma_e$ . D'ailleurs, par exemple si  $n = 100$ ,  $\sqrt{\frac{n}{n-1}} \approx 0,995 \approx 1$ . Ça ne fait donc pas une grande différence.

### Estimation par intervalle de confiance

L'intervalle de confiance ajoute ce qui manque à l'estimation ponctuelle : une idée de la marge d'erreur de l'estimation.

Dans l'échantillon de taille  $n$ , on a obtenu la moyenne  $\bar{x}_e$  et l'écart-type  $\sigma_e$ . La moyenne dans la population est estimée par l'intervalle de confiance :

$$I_C = \left[ \bar{x}_e - t \frac{\sigma_e}{\sqrt{n-1}} ; \bar{x}_e + t \frac{\sigma_e}{\sqrt{n-1}} \right], \begin{cases} t = 1,645 \text{ pour } I_C \text{ à } 90\% \\ t = 1,96 \text{ pour } I_C \text{ à } 95\% \\ t = 2,575 \text{ pour } I_C \text{ à } 99\% \end{cases}$$

De la même façon, dans l'échantillon de taille  $n$ , on a constaté la fréquence  $f$ . L'intervalle de confiance pour la fréquence  $p$  dans la population est :

$$I_C = \left[ f - t \sqrt{\frac{f(1-f)}{n-1}} ; f + t \sqrt{\frac{f(1-f)}{n-1}} \right], \begin{cases} t = 1,645 \text{ pour } I_C \text{ à } 90\% \\ t = 1,96 \text{ pour } I_C \text{ à } 95\% \\ t = 2,575 \text{ pour } I_C \text{ à } 99\% \end{cases}$$

**Remarque 6 :** l'intervalle de confiance pour la moyenne tient compte de l'estimation  $\sigma_{pop} = \sigma_e \sqrt{\frac{n}{n-1}}$  et du fait que  $\bar{x}_e$  est la réalisation d'une variable  $\bar{X}_n$  qui a l'écart-type  $\frac{\sigma_{pop}}{\sqrt{n}}$  selon la loi des moyennes.

**Remarque 7 :**  $I_C$  pour les fréquences est une adaptation du  $I_C$  pour les moyennes comme vu dans la remarque 2. En effet, si dans l'échantillon on donne la valeur 1 à ceux qui ont la caractéristique demandée et 0 aux autres, on aura une moyenne  $\bar{x}_e = f$  et un écart-type  $\sigma_e = \sqrt{f(1-f)}$ . Le raisonnement est alors le même et le  $n-1$  apparaît pour la même raison.

### Test de validité

On émet une hypothèse comme « la moyenne dans la population est... » ou « la fréquence dans la population est... ». Cette hypothèse est appelée parfois **hypothèse nulle** ou  $H_0$ . L'hypothèse contraire est  $H_1$ .

Il s'agit de prendre une décision sur la base d'un échantillon pour savoir si on accepte l'hypothèse ou si on la rejette. La règle est énoncée **Avant** de d'effectuer le prélèvement.

- On détermine l'intervalle de fluctuation  $I_F$  sous l'hypothèse  $H_0$  avec un seuil de risque  $\alpha$  donné.
- La règle est :
  - Si la moyenne [ou fréquence] constatée dans l'échantillon  $\in I_F$ , on accepte  $H_0$ .
  - Sinon, on accepte  $H_1$  (on rejette  $H_0$ ).

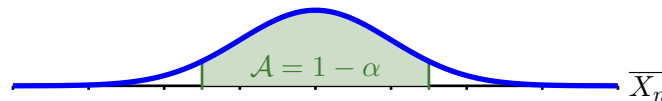
**Cas bilatéral**

On se place ici dans le cas où l'hypothèse porte sur une moyenne. Dans ce cas, on est obligé de connaître la valeur de  $\sigma_{pop}$ . On pourrait faire le même travail pour une fréquence, mais dans ce cas pas besoin d'avoir  $\sigma$  en plus.

$H_0$  : la moyenne dans la population est  $\mu_{pop} = \mu_A$ . j'ai mis  $A$  pour « affirmation »

$H_1$  :  $\mu_{pop} \neq \mu_A$

Pour décider, on prélève un échantillon de taille  $n$  et on va considérer la moyenne  $\bar{x}_e$  obtenu dans cet échantillon. Avant de réaliser l'échantillon, on ne connaît pas  $\bar{x}_e$ . On ne peut faire que des prédictions à l'aide de probabilités, en supposant que  $H_0$  est vraie. La moyenne de l'échantillon est décrite par  $\bar{X}_n$  (loi des moyennes) qui suit une loi normale.



Si l'hypothèse  $H_0$  est vraie, la réalisation de l'expérience devrait nous faire tomber dans la zone centrale. Si on tombe dans une des zones latérales, on considèrera que  $H_0$  est probablement fausse.

Quand on dit «  $I_F$  au risque  $\alpha$  » (par exemple  $\alpha = 5\%$ ),  $\alpha$  représente l'aire totale des deux zones latérales. L'aire centrale est donc  $1 - \alpha$ .

$I_F$  est donc calculé pour correspondre à l'aire  $1 - \alpha$  sachant que  $\bar{X}_n$  suit la loi  $\mathcal{N}$  avec  $\mu = \mu_A$  (selon l'hypothèse) et  $\sigma = \frac{\sigma_{pop}}{\sqrt{n}}$ .

$$I_F = \left[ \mu_A - t \frac{\sigma_{pop}}{\sqrt{n}} ; \mu_A + t \frac{\sigma_{pop}}{\sqrt{n}} \right], \begin{cases} t = 1,645 \text{ pour } I_F \text{ à } 90\% \\ t = 1,96 \text{ pour } I_F \text{ à } 95\% \\ t = 2,575 \text{ pour } I_F \text{ à } 99\% \end{cases}$$

Dans le cas fréquence, on retrouve le  $I_F$  habituel (où  $p$  est donné par  $H_0$ ) :

$$I_F = \left[ p - t \sqrt{\frac{p(1-p)}{n}} ; p + t \sqrt{\frac{p(1-p)}{n}} \right], \begin{cases} t = 1,645 \text{ pour } I_F \text{ à } 90\% \\ t = 1,96 \text{ pour } I_F \text{ à } 95\% \\ t = 2,575 \text{ pour } I_F \text{ à } 99\% \end{cases}$$

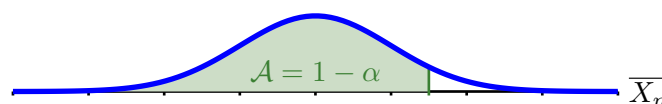
**Cas unilatéral**

Il peut arriver que l'on ne considère qu'un seul risque.

$H_0$  : la moyenne dans la population est  $\mu_{pop} = \mu_A$ .

$H_1$  :  $\mu_{pop} > \mu_A$

Dans ce cas, le calcul de  $I_F$  ne considère comme risque que la bande latérale à droite de la courbe (c'est à dire  $\bar{X}_n > \mu_{pop}$ )



Il n'est plus considéré comme un risque d'être à gauche.

$$I_F = \left] -\infty ; \mu_A + t \frac{\sigma_{pop}}{\sqrt{n}} \right], \begin{cases} t = 1,28 \text{ pour } I_F \text{ à } 90\% \\ t = 1,645 \text{ pour } I_F \text{ à } 95\% \\ t = 2,33 \text{ pour } I_F \text{ à } 99\% \end{cases}$$

**Remarque 8 :** Quand  $H_0$  est validée, cela ne veut pas dire que  $\mu_{pop}$  exactement égal à  $\mu_A$ . Cela veut dire que  $\mu_{pop}$  est suffisamment proche de  $\mu_A$  pour que le test ne permette pas de voir la différence. Quand  $H_1$  est validée au contraire, cela signifie que la différence est assez grande pour que le test la mette en évidence.

#### Risques de première et seconde espèce

Se peut-il que  $\mu_{\text{échantillon}} \notin I_F$  et que pourtant  $H_0$  soit vraie ?

Oui bien sûr. C'est un **faux négatif**. C'est une erreur. Quand on dit « Au risque  $\alpha$  », le risque dont on parle est le risque de faire un faux négatif.

$\alpha$  est donc la probabilité de faire un faux négatif avec le test (décider de rejeter  $H_0$  alors que  $H_0$  est vraie). On parle de **risque de première espèce**.

Le risque opposé existe : le faux positif, c'est à dire accepté  $H_0$  alors que  $H_0$  est faux. La probabilité d'une telle est erreur est  $\beta$ . C'est le **risque de seconde espèce**. Plus difficile à calculer. Il suppose de connaître plus finement les caractéristique de la population. En général, diminuer  $\alpha$  augmente  $\beta$ . Il faut donc trouver un compromis. Prendre  $\alpha = 5\%$  est souvent un bon compromis, c'est pourquoi on prend souvent un  $I_F$  à 95 %.