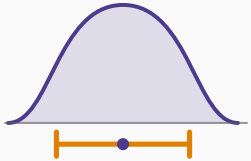
Statistique inférentielle

Chapitre 4 : Proportion et variance



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

La proportion

Une proportion est une moyenne

Le déguisement

Posons $X_i = 1$ si l'individu possède le caractère étudié, 0 sinon. Alors

$$F = \frac{X_1 + \cdots + X_n}{n} = \overline{X}$$

La fréquence observée **est** la moyenne d'une variable indicatrice.

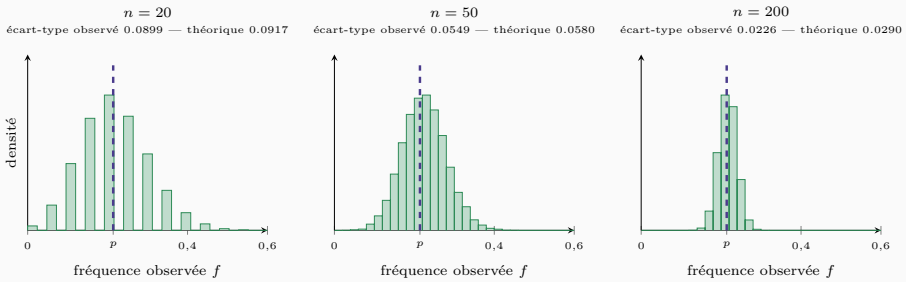
Tout le chapitre 3 s'applique sans un mot de plus

Il suffit de savoir que pour une indicatrice, $E(X_i) = p$ et $\sigma(X_i) = \sqrt{p(1-p)}$ – c'est la loi de Bernoulli.

$$E(F) = p \qquad \sigma(F) = \sqrt{\frac{p(1-p)}{n}}$$

Aucun résultat nouveau : la même formule, avec un σ particulier.

Sa distribution d'échantillonnage



Une proportion n'est qu'une moyenne : celle d'une variable qui vaut 1 ou 0.

Tout ce qui a été dit de $\overline{X}$ s'applique donc mot pour mot, avec $\sigma = \sqrt{p(1-p)}$:

$$E(F) = p \qquad \sigma_F = \sqrt{\frac{p(1-p)}{n}} \qquad \text{et } F \text{ tend vers une normale.}$$

À $n = 20$, la loi est encore visiblement discrète et dissymétrique — l'approximation normale serait abusive. La règle usuelle est $np \geq 5$ et $n(1-p) \geq 5$: ici $20 \times 0{,}214 = 4{,}3$, insuffisant.

Quand peut-on lire une table normale

$$np \geq 5 \qquad \text{et} \qquad n(1-p) \geq 5$$

Ces deux conditions viennent directement de l'approximation normale de la binomiale, chapitre 18 des probabilités.

Sur nos données

Proportion de diplômés Bac+5 dans la population : $p = 0{,}214$.

$n = 20$: $np = 4{,}3$ – **insuffisant**, la figure montre une loi encore visiblement discrète. $n = 50$: $np = 10{,}7$ et $n(1-p) = 39{,}3$ – les deux conditions sont remplies.

Un maximum utile à connaître

Le pire cas est en $p = 0{,}5$

La fonction $p(1-p)$ est maximale en $p = 0{,}5$, où elle vaut $0{,}25$.

$$\sigma(F) \leq \frac{0{,}5}{\sqrt{n}} \qquad \text{quelle que soit la valeur de } p$$

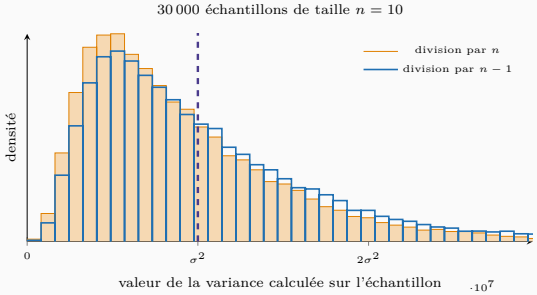
Pourquoi les instituts annoncent toujours la même marge

Ne connaissant pas p avant l'enquête, on dimensionne au cas le plus défavorable.

C'est pourquoi la marge d'erreur publiée est la même quel que soit le résultat : elle est **majorée**, donc valable pour toutes les questions du questionnaire à la fois.

La variance

L'énigme du dénominateur



Le décalage se voit à l'œil.

En divisant par n, la moyenne des 30 000 valeurs vaut 6107414 — soit $0.90\,\sigma^2$: **trop petite**.

En divisant par $n - 1$, elle vaut 6786015, soit σ^2 exactement.

Pourquoi ? Les écarts sont mesurés à $\overline{x}$, qui est *lui-même* calculé sur l'échantillon et se place donc au plus près des points. On sous-estime nécessairement la dispersion autour du vrai m.

Le facteur $\dfrac{n}{n-1}$ corrige exactement ce rapprochement. C'est le *degré de liberté* perdu à estimer m.

Ce que la figure établit

Deux formules, deux comportements

$$s_n^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2 \qquad\qquad s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$$

La première **sous-estime** systématiquement σ^2. La seconde est sans biais : $E(s^2) = \sigma^2$.

La raison, en une phrase

Les écarts sont mesurés à $\overline{x}$, qui est calculé **sur le même échantillon** et se place donc au plus près des points.

La somme des carrés obtenue est ainsi la plus petite possible — plus petite que si l'on avait mesuré les écarts au vrai m, que l'on ignore. On sous-estime nécessairement la dispersion.

Ce que le $n - 1$ compte

Une fois $\overline{x}$ calculé, les n écarts $(x_i - \overline{x})$ ne sont plus libres : leur somme est **nulle** par construction.

Connaissant $n - 1$ d'entre eux, le dernier est déterminé. Il n'y a donc que $n - 1$ écarts réellement indépendants : c'est le **nombre de degrés de liberté**.

Une règle qui reviendra sans cesse

On perd un degré de liberté par paramètre estimé à partir des données.

Ici, un seul : m, remplacé par $\overline{x}$. Au chapitre 20, la régression en estimera deux — pente et ordonnée — et le dénominateur sera $n - 2$. Le principe ne change jamais.

Population et échantillon

	Formule	Valeur
Population ($N = 500$)	division par N	$\sigma = 2\,598{,}10$
Échantillon ($n = 50$)	division par $n - 1$	$s = 2\,250{,}98$

Le point de vocabulaire

Sur la **population**, on divise par N : ce n'est pas une estimation, c'est une définition. Le cours de descriptive faisait cela.

Sur un **échantillon** dont on veut estimer σ, on divise par $n - 1$. Les deux formules sont correctes, chacune dans son rôle — et c'est pourquoi les calculatrices proposent les deux touches.

La loi de s^2

Le résultat, pour une population normale

$$\frac{(n-1)\,s^2}{\sigma^2} \;\sim\; \chi^2_{n-1}$$

La variance d'échantillon, correctement normalisée, suit une loi du khi-deux à $n-1$ degrés de liberté.

Cela n'a rien d'étonnant

s^2 est une **somme de carrés** d'écarts. Or le khi-deux a été défini, au chapitre 19 des probabilités, comme une somme de carrés de normales centrées réduites.

Le degré de liberté $n-1$ est le même que celui du dénominateur : ce n'est pas une coïncidence, c'est la même perte d'information.

Ce qu’il faut retenir

Le résumé du chapitre

Six points

1. Une **proportion est une moyenne** d'indicatrices : tout le chapitre 3 s'y applique.
2. $E(F) = p$ et $\sigma(F) = \sqrt{p(1-p)/n}$.
3. Conditions de normalité : $np \geq 5$ **et** $n(1-p) \geq 5$.
4. $p(1-p)$ est maximal en $0{,}5$: c'est le cas défavorable, et donc celui qu'annoncent les sondeurs.
5. La variance d'échantillon se divise par $n-1$: les écarts sont mesurés à $\overline{x}$, ce qui les rend trop petits.
6. **Un degré de liberté perdu par paramètre estimé** — la règle vaudra jusqu'à la régression.

La suite

Trois distributions d'échantillonnage sont maintenant établies : moyenne, proportion, variance.

Deux d'entre elles font apparaître des lois que le cours de probabilités avait présentées sans dire à quoi elles servaient. Le chapitre 5 boucle cette dette.