L'INDUCTION STATISTIQUE DANS LA MÉTHODE SCIENTIFIQUE HYPOTHÉTICO-DÉDUCTIVE

Théorie
↓
Déduction des implications
↓
Confrontation avec les observations
↓
Révision ou rejet de la théorie si contradiction

Théorie
↓
(formalisation)
↓
Modèle
↓
Hypothèses théoriques
↓
(opérationnalisation, mesure)
↓
Hypothèses opérationnelles
↓
Confrontation avec les observations
↓
Révision ou rejet de la théorie
ou de la formalisation
ou de l'opérationnalisation
si contradiction

VARIABLES ALÉATOIRES CONTINUES : DENSITÉ DE PROBABILITÉ ET ESPÉRANCE MATHÉMATIQUE

Fonction de distribution cumulative d'une variable aléatoire

Nombre de face	Probabilité		Probabilité cumulée
x_i	$f(x_i)$	$+ F(x_{i-1})$	$= F(x_i)$
0	1/16		1/16
1	4/16	+ 1/16	= 5/16
2	6/16	+ 5/16	= 11/16
3	4/16	+ 11/16	= 15/16
4	1/16	+ 15/16	= 16/16

Fonction de densité de probabilité d'une variable aléatoire continue

$F(x)$ = Prob (variable aléatoire $\leq x$) = Fonction de distribution cumulative

Fonction de densité : $f(x) = \frac{d}{dx} F(x)$

Probabilité cumulative = intégrale de la fonction de densité

Prob$(a \leq$ variable aléatoire $\leq b) = F(b)\text{-}F(a)$

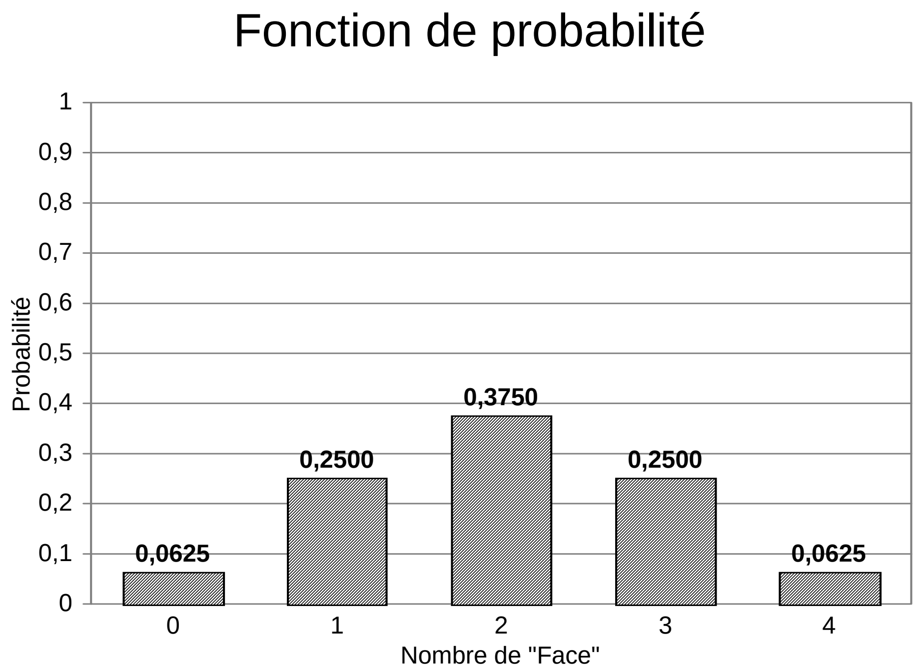
Prob$(a \leq$ variable aléatoire $\leq b) =$ surface sous $f(x)$ entre a et $b = \int_a^b f(x)\,dx$

Espérance mathématique

$$E(x) = \int_{-\infty}^{+\infty} f(x)\, x\, dx$$

$$\sigma_X^2 = E\left\{\left[x - E(x)\right]^2\right\} = \int_{-\infty}^{+\infty} f(x)\left[x - E(x)\right]^2 dx$$

FONCTION DE PROBABILITÉ
ET FONCTION DE DISTRIBUTION CUMULATIVE
(FIGURES)



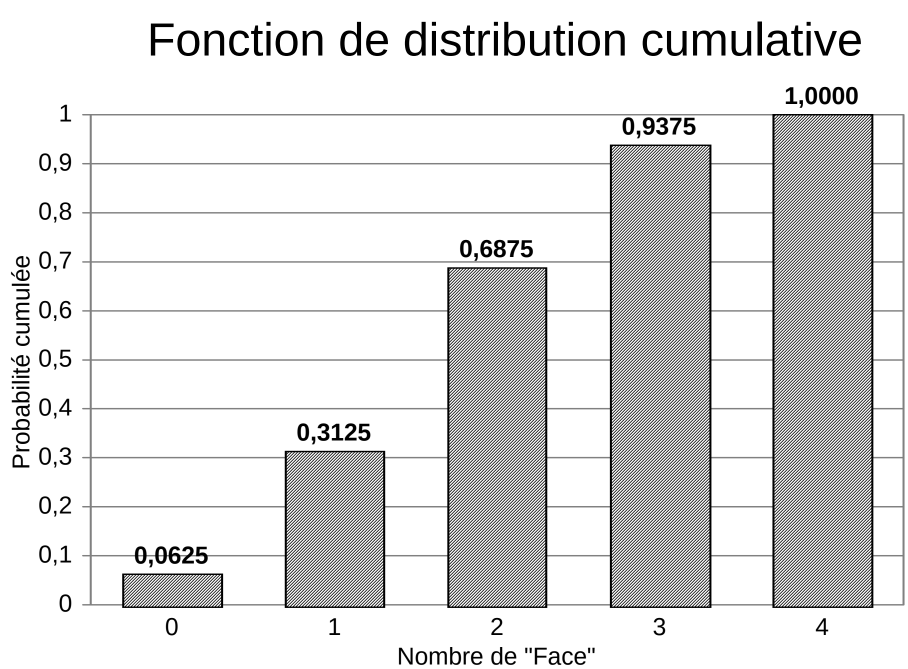


SCHÉMA 1A

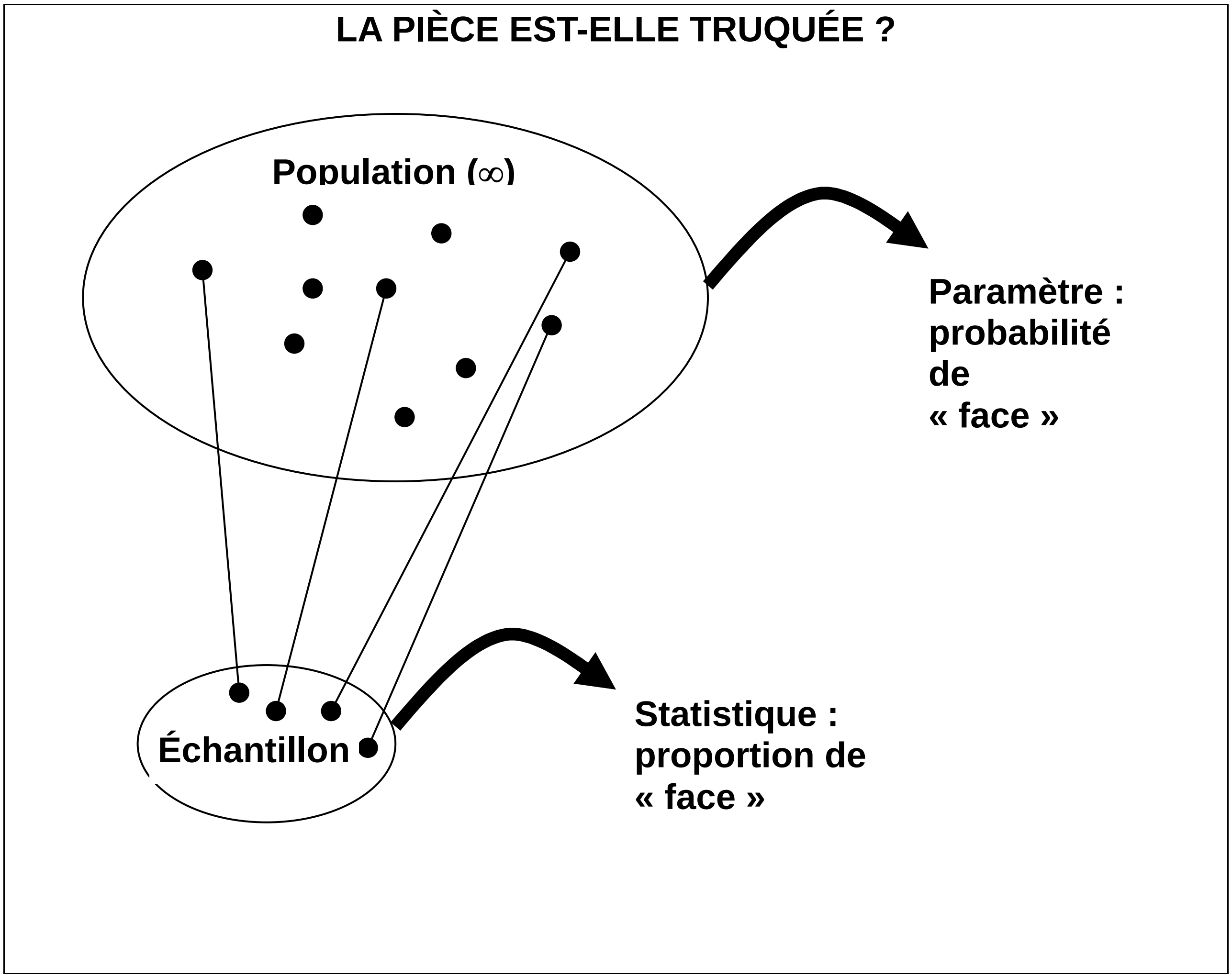

LA PIÈCE EST-ELLE TRUQUÉE ?
Population (∞)
Paramètre :
probabilité
de
« face »
Échantillon
Statistique :
proportion de
« face »

SCHÉMA 1B

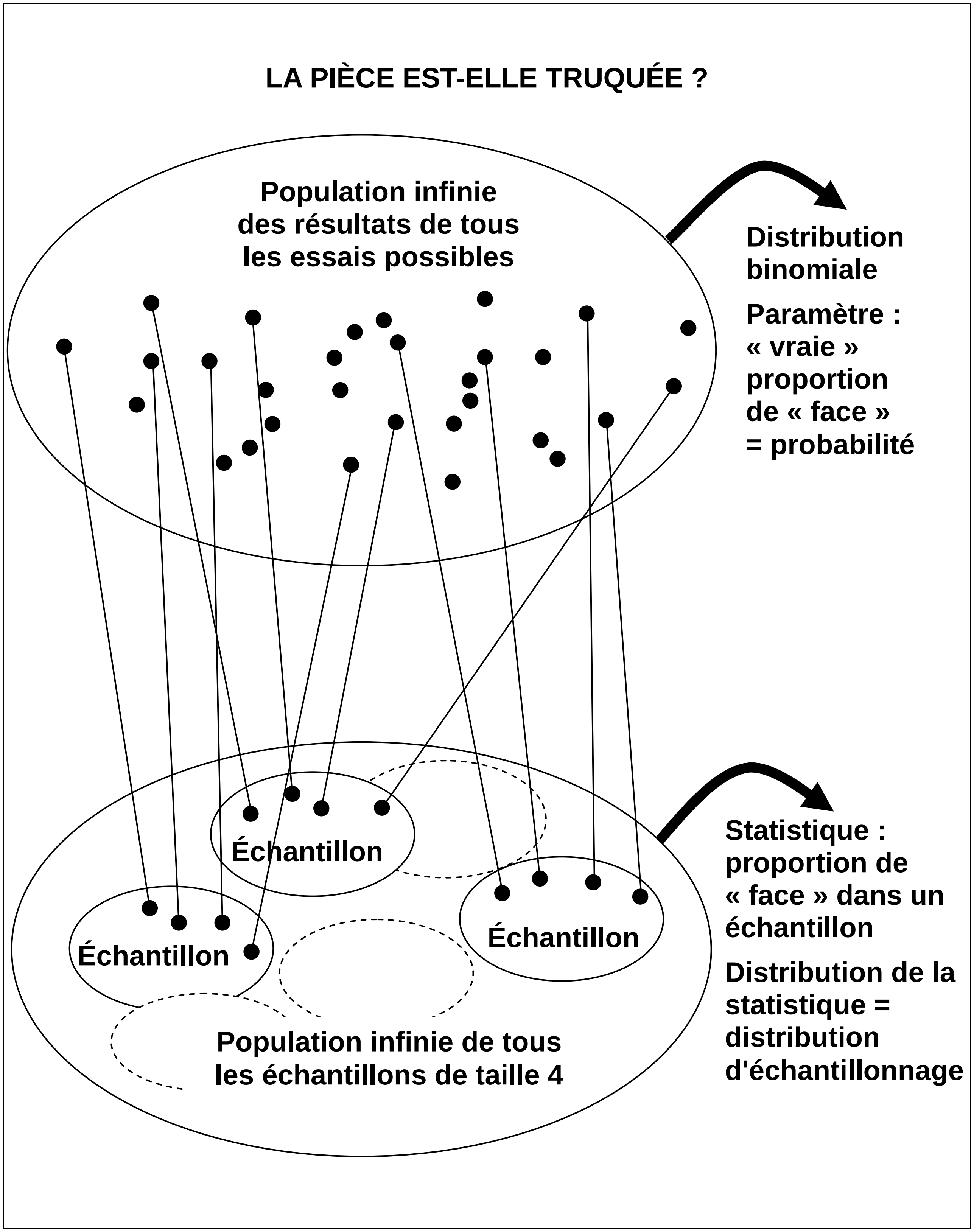
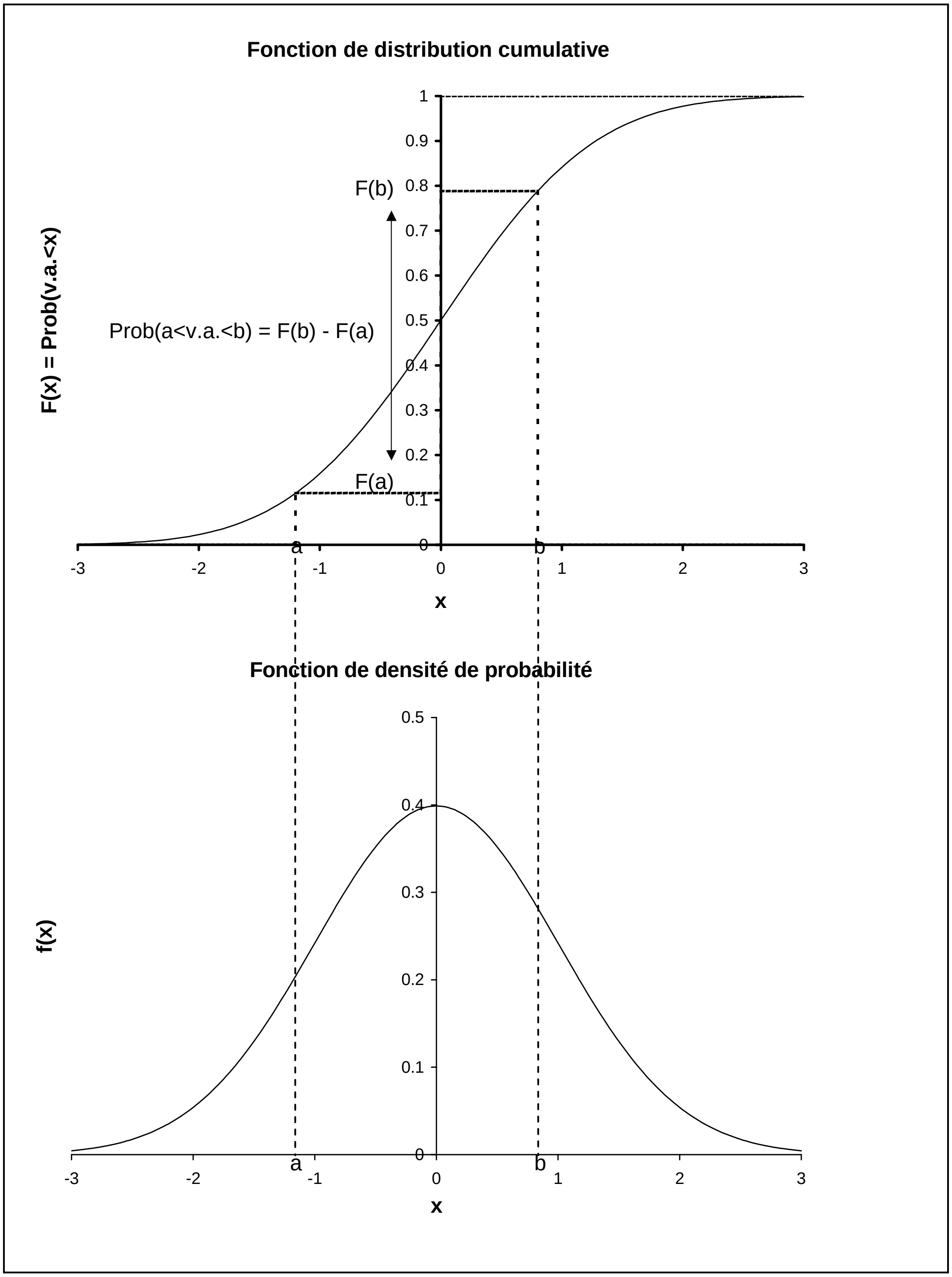

LA PIÈCE EST-ELLE TRUQUÉE ?
Population infinie des résultats de tous les essais possibles
Distribution binomiale
Paramètre : « vraie » proportion de « face » = probabilité
Échantillon
Échantillon
Échantillon
Population infinie de tous les échantillons de taille 4
Statistique : proportion de « face » dans un échantillon
Distribution de la statistique = distribution d'échantillonnage

FIGURE 1 – FONCTION CUMULATIVE ET FONCTION DE DENSITÉ



DISTANCE PARCOURUE ET VITESSE (EXEMPLE NUMÉRIQUE)

Temps écoulé (minutes)	Distance parcourue (km)	Vitesse moyenne entre t–1 et t+1 (km/h)	Vitesse instantanée (km/h)	Différence
t	$d(t)$	$v_M(t)$	$v_i(t)$	$v_M(t) - v_i(t)$
0	0,00		0,03	
1	0,00	0,20	0,15	0,05
2	0,01	0,72	0,57	0,15
3	0,03	2,25	1,90	0,35
4	0,08	6,06	5,37	0,68
5	0,23	13,98	12,96	1,02
6	0,55	27,70	26,62	1,07
7	1,15	47,12	46,60	0,51
8	2,12	68,85	69,53	-0,67
9	3,45	86,44	88,38	-1,94
10	5,00	93,25	95,75	-2,49
11	6,55	86,44	88,38	-1,94
12	7,88	68,85	69,53	-0,67
13	8,85	47,12	46,60	0,51
14	9,45	27,70	26,62	1,07
15	9,77	13,98	12,96	1,02
16	9,92	6,06	5,37	0,68
17	9,97	2,25	1,90	0,35
18	9,99	0,72	0,57	0,15
19	10,00	0,20	0,15	0,05
20	10,00		0,03	

Distance parcourue générée par la fonction [1] : $d(t) = 10 \int_{-\infty}^{t} \frac{1}{2{,}5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-10}{2{,}5}\right)^2}$

Calcul de la vitesse moyenne sur un intervalle : $v_M(t) = \frac{d(t+1) - d(t-1)}{(t+1) - (t-1)} = \frac{d(t+1) - d(t-1)}{2}$

Vitesse instantanée : $v_i(t) = 60 \times 10 \times \frac{1}{2{,}5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-10}{2{,}5}\right)^2}$. C'est la dérivée de *d(t)*.

La vitesse est à la distance parcourue

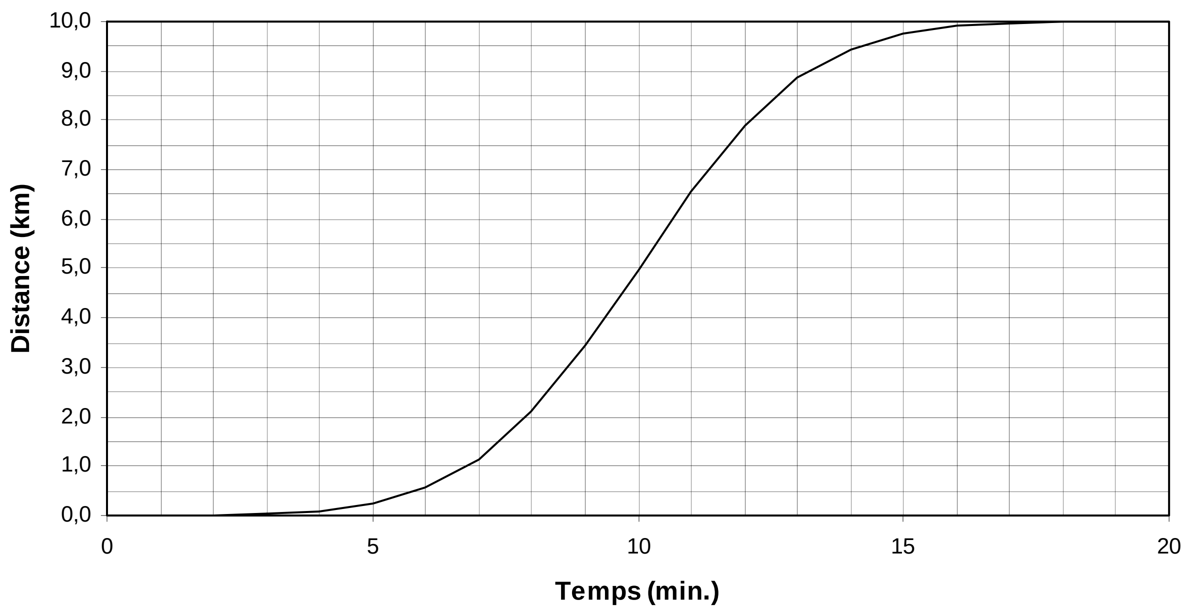
ce que la densité de probabilité est à la probabilité cumulative.

---

[1] C'est 10 fois la probabilité de la normale cumulative d'une variable aléatoire *t* de moyenne 10 et d'écart type 2,5. Les valeurs sont calculées au moyen de la fonction `NORMDIST` d'Excel.

DISTANCE PARCOURUE ET VITESSE (FIGURES)

Distance parcourue (km)



Vitesse moyenne d'intervalle et vitesse instantanée (km/h)

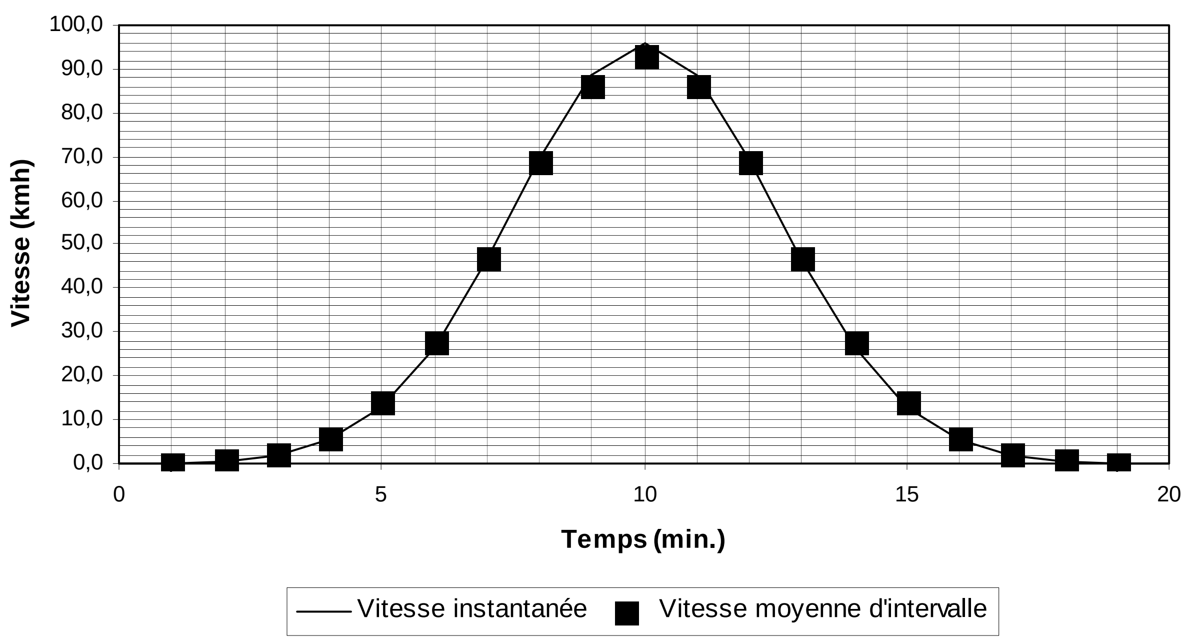
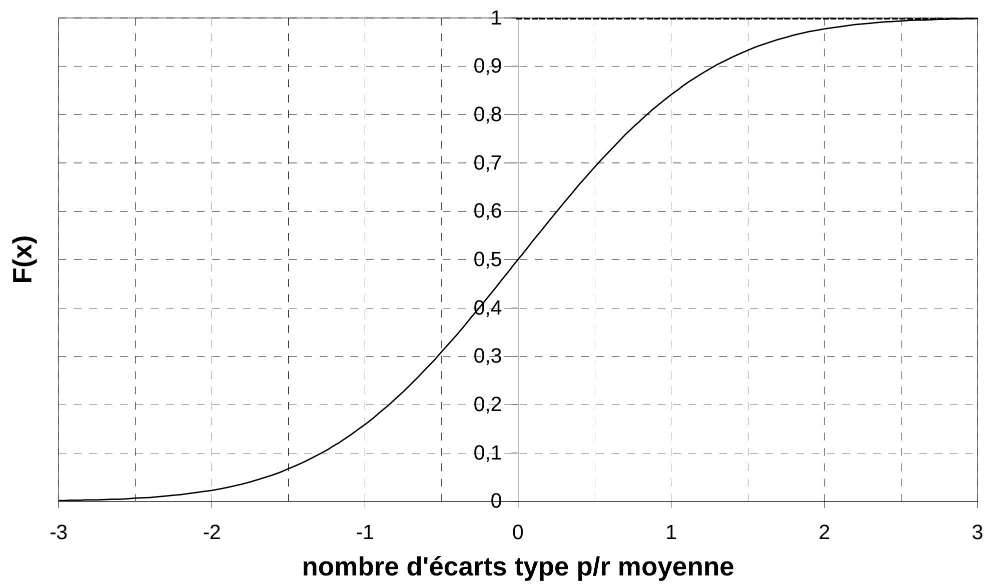
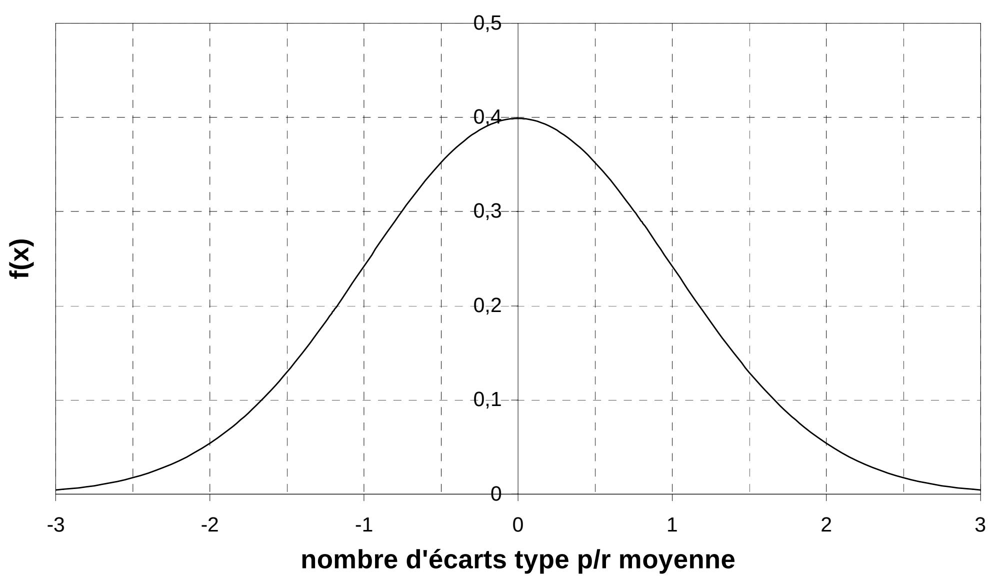


FIGURE 2 – DISTRIBUTION NORMALE

Fonction de distribution normale cumulative



Fonction de densité normale



ÉCHANTILLONNAGE, ESTIMATION
ET TESTS D'HYPOTHÈSES

1. Échantillonnage

(1) Sélection

- échantillon aléatoire simple
- échantillon aléatoire stratifié
- échantillon par grappes
- etc.

(2) Taille

2. Estimation

(1) Méthodes

- Estimation selon l'approche analogique
- Principe des moindres carrés
- Principe du maximum de vraisemblance

(2) Propriétés désirables

- Absence de biais
- Efficacité relative
- Convergence
- Suffisance

3. Tests d'hypothèse

Une logique de rejet/non-rejet

- Si une théorie (ou un modèle, ou une hypothèse) est vraie, alors ses implications doivent être vraies aussi.
- Donc, si les observations contredisent les implications d'une théorie, cette théorie ne peut pas être vraie : elle est fausse.

Si les observations ne contredisent pas les implications d'une théorie, on n'a pas le droit de conclure que cette théorie est vraie !

Une logique non pas déterministe, mais probabiliste

- Dans une logique probabiliste, une observation est *plus ou moins compatible* avec l'hypothèse : plus une observation est improbable lorsqu'on suppose que l'hypothèse est vraie, moins elle est compatible avec cette hypothèse.

QUELQUES IDÉES CLÉS

1. Les méthodes d'induction statistique sont une expression mathématique de principes épistémologiques en vertu desquels, à partir de l'information contenue dans un ensemble de données particulier, on peut arriver à des propositions de portée plus générale (p. 2-1.3).
2. Un échantillon n'est qu'un des échantillons possibles de même taille qu'on aurait pu tirer de la population étudiée. D'où, le lien aléatoire entre l'échantillon et la population (p. 2-1.4).
3. Cependant, la théorie des probabilités nous donne des outils pour évaluer la *probabilité* que l'écart (l'erreur d'estimation) entre la statistique et le paramètre se situe à l'intérieur d'une certaine marge (p. 2-1.5).
4. Pour être confrontés aux observations, les modèles déterministes ont besoin d'être complétés de façon à tenir compte du lien aléatoire entre la population et l'échantillon. Lorsqu'on combine un modèle déterministe et un modèle du lien aléatoire entre l'échantillon et la population, on obtient un modèle aléatoire (probability model). Dans un modèle aléatoire, le modèle du lien entre l'échantillon et la population s'appelle « modèle d'échantillonnage » (sampling model) (p.2-2.3 et 4).
5. Un modèle aléatoire est, comme tout modèle, de caractère hypothétique ; il est constitué entre autres d'*hypothèses sur la structure aléatoire*, sur les lois de probabilité qui régissent le hasard (p. 2-2.4).
6. Chaque exercice de confrontation entre une théorie et les observations repose en fait sur un modèle plus général qui n'est pas remis en question (p. 2-2.5).
7. La plupart du temps, les tests d'hypothèse portent sur des formes particulières d'un modèle théorique général, qui n'est pas remis en question, et s'appuient sur un modèle aléatoire, qui n'est pas remis en question non plus (p. 2-2.5).

QUELQUES IDÉES CLÉS (SUITE)

8. La *distribution d'échantillonnage* d'une statistique est sa distribution de probabilité dans la population des échantillons d'une taille donnée qu'on peut tirer au hasard de la population étudiée (p. 2-2.10).
9. Avec une variable aléatoire continue, le nombre de valeurs possibles est infini. Il s'ensuit que la probabilité que la variable aléatoire prenne *une valeur en particulier* est normalement *infiniment petite* (p. 2-2.13).
10. Avec une variable aléatoire continue, le concept de fonction de probabilité tel que défini pour une variable aléatoire discrète ne s'applique pas, mais la fonction de distribution cumulative existe bel et bien en général. Et c'est à partir de la fonction de distribution cumulative $F(x)$ que l'on définit la fonction de *densité* $f(x)$ (p. 2-2.14).
11. L'ordonnée d'une fonction de densité (sa hauteur) *n'est pas une probabilité* (alors que l'ordonnée d'une fonction de probabilité, elle, est une probabilité). Par contre, la *surface sous la courbe* d'une fonction de densité est une probabilité : techniquement, puisque $f(x)$ est la *dérivée* de $F(x)$, il s'ensuit que $F(x)$ est donnée par l'*intégrale* de $f(x)$ (p. 2-2.14).
12. Logique de rejet/non-rejet : « si les observations ne contredisent pas les implications d'une théorie, on n'a pas le droit de conclure que cette théorie est vraie ! » (p. 2-2.27); elle demeure « acceptable », mais elle n'est pas « acceptée ».
13. Logique probabiliste : « une observation est *plus ou moins compatible* avec l'hypothèse » (p. 2-2.27).