UNIVERSITÉ DE GENÈVE

Introduction à la Statistique

Inférence statistique: estimateurs

Stéphane Guerrier, Mucyo Karemera, Samuel Orso & Lionel Voirol

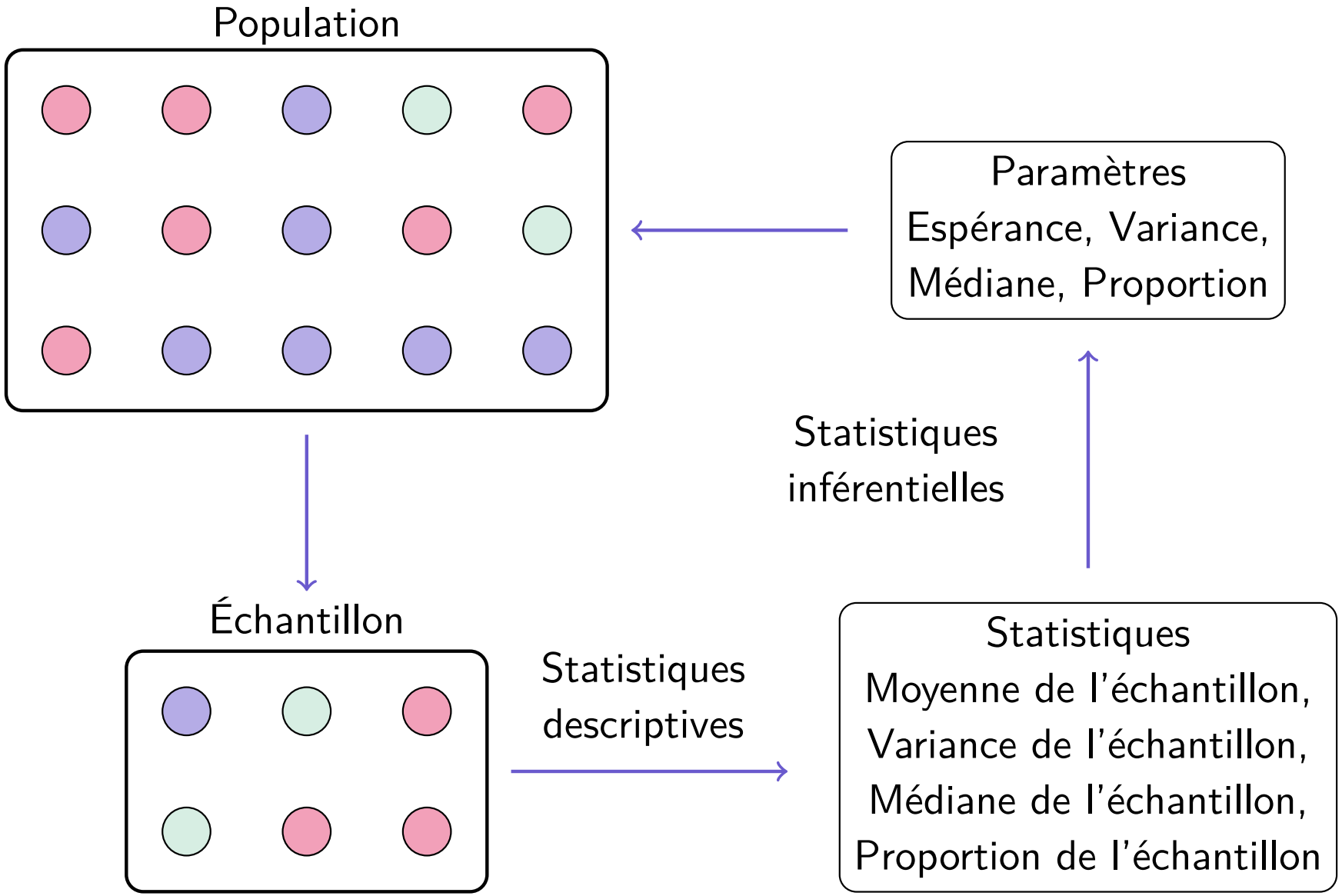
Data Analytics Lab

Statistiques descriptives et inférentielles

Statistiques descriptives : décrire les données que nous avons collectées qui composent l'échantillon

Inférence statistique : faire des généralisations sur un ensemble plus large, la population

Statistiques descriptives et inférentielles



Exemple

Supposons que Stephen Curry ait marqué en moyenne 30.1 points lors de 79 matches de basketball.

Qu'est-ce qui est aléatoire ?

Qu'est-ce qui est inconnu ?

$$X = \text{points marqués dans un seul match}$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Qu'est-ce que μ ? Et σ ?

Statistiques vs. paramètres

Paramètre : une caractéristique de la population. Typiquement inconnue en raison du grand nombre d'individus dans la population et/ou de l'impossibilité de mesurer tous les résultats possibles.

Statistique : une quantité qui est calculée à partir des données que nous avons collectées de la population (échantillon).

Statistiques vs. paramètres

Statistiques : sont calculées à partir de l'échantillon

$\bar{X}$: moyenne de l'échantillon

s : écart-type de l'échantillon

$\hat{p}$: proportion de l'échantillon

Paramètres : inconnus

μ : moyenne de la population

σ : écart-type de la population

p : proportion de la population

Distribution d'échantillonnage

Les données sont des variables aléatoires

Les statistiques sont des fonctions des données. Elles sont donc aussi des variables aléatoires

La distribution des statistiques dépend des paramètres de la distribution des données

Exemple : Moyenne de l'échantillon

Statistique $= \bar{X}$

Paramètre $= \mu$

$\bar{X}$ en tant que variable aléatoire

Données : $X_1, X_2, X_3, \ldots, X_n$

Échantillon Aléatoire Simple (EAS)

Observations Indépendantes et Identiquement Distribuées (iid)

Par exemple, chaque

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \text{ pour } i = 1, \ldots, n$$

Malheureusement, μ et σ sont inconnus.

Espérance de $\bar{X}$

$$\mathbb{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

Pourquoi ?

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \mu$$

Écart-type de $\bar{X}$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

L'erreur dans $\bar{X}$ diminue avec n

Dépend de σ, l'écart-type des données, qui est inconnu

Erreur standard de $\bar{X}$

$$e.s.(\bar{X}) = \frac{s}{\sqrt{n}}$$

Quelle est la différence entre $\sigma_{\bar{X}}$ et $e.s.(\bar{X})$?

σ est généralement inconnu

Nous remplaçons σ par la statistique s

Exemple : 5 observations normales et indépendantes

Jeu de données : $63, 65, 72, 74, 74$

Moyenne de l'échantillon (statistique) :

$$\bar{X} = \frac{63 + 65 + 72 + 74 + 74}{5} = 69.6$$

Écart-type de l'échantillon (correspond à 1 variable aléatoire) :

$$s = \sqrt{\frac{1}{4}\left(\sum_{i=1}^{5} x_i^2 - 5(69.6)^2\right)} = 5.225$$

Erreur standard :

$$e.s.(\bar{X}) = \frac{s}{\sqrt{5}} = \frac{5.225}{\sqrt{5}} = 2.337$$

Distribution de $\bar{X}$

Règle :

Lorsque $X_1, X_2, \ldots, X_n$ sont indépendantes et suivent une distribution normale avec une moyenne μ et un écart-type σ (c'est-à-dire que chaque $X_i \sim \mathcal{N}(\mu, \sigma^2)$), alors

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardisation

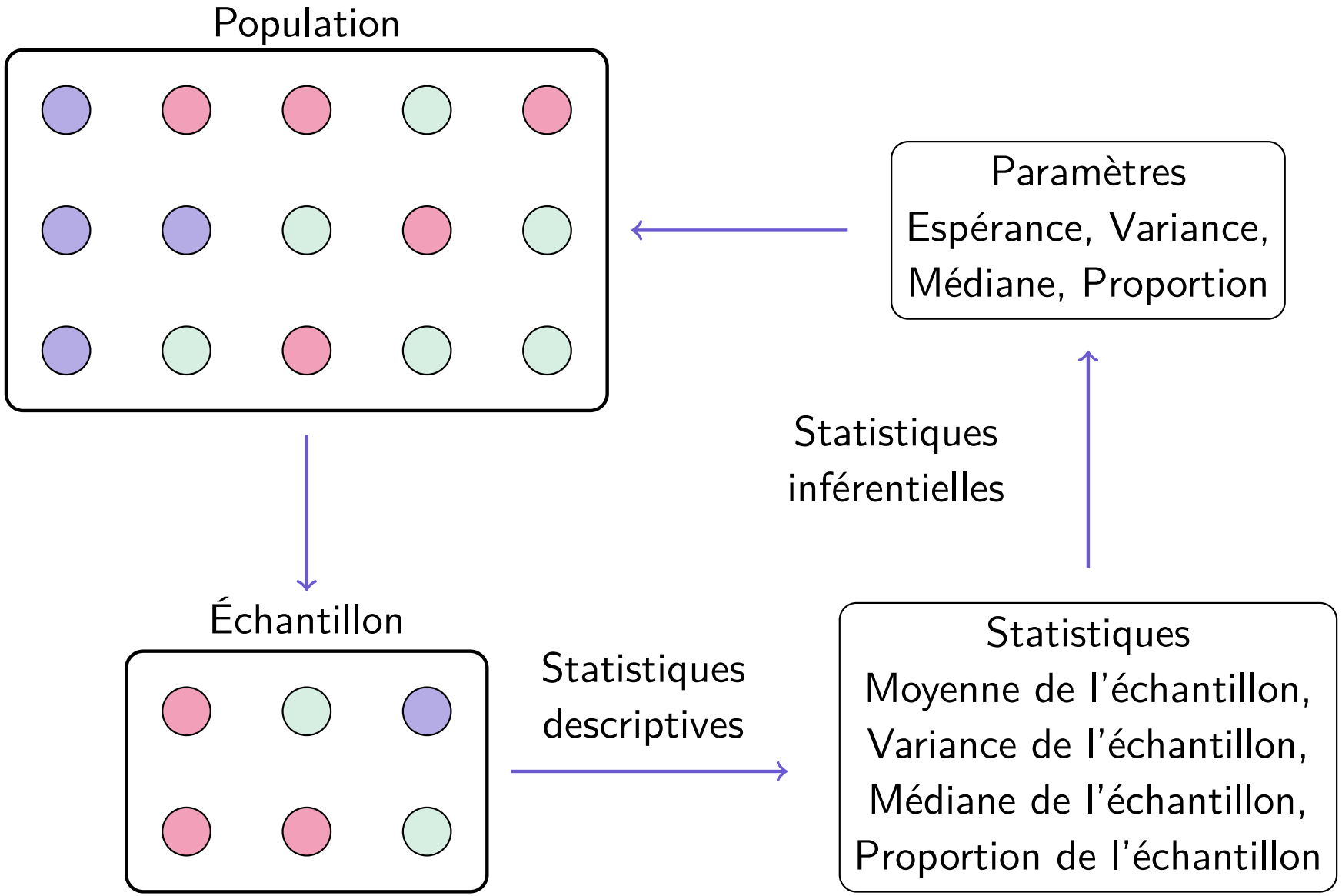
Pour une variable X :

$$Z = \frac{X - \mu}{\sigma}$$

Pour la moyenne de l'échantillon $\bar{X}$ de n variables $X_1, \ldots, X_n$:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma}$$

Population et échantillon



Exemple : Nombre moyen de voitures dans les foyers américains

On considère le nombre de voitures dans chaque foyer aux États-Unis.

Population : Tous les foyers américains

Taille de la population : $N = 324,227,000$

Jeu de données : $x_1, \ldots, x_N$, où x_1 serait le nombre de voitures dans le 1er foyer de la population, etc.

Moyenne de la population :

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Écart-type de la population :

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{\text{pour tous } x} (x - \mu)^2}$$

Échantillon

Jeu de données de l'échantillon : $\{X_1, \ldots, X_n\}$

Taille de l'échantillon : n

Ordre de l'échantillon : $n \ll N$ (c'est-à-dire, n est (généralement) grand mais beaucoup plus petit que N)

Objectif :

À partir des valeurs de l'échantillon, $\{X_1, \ldots, X_n\}$, où n est la taille de l'échantillon, nous essayons de tirer des conclusions sur les paramètres d'intérêt.

L'échantillon idéal doit être représentatif et non biaisé

Choisir l'échantillon de manière aléatoire.

Échantillon aléatoire simple

$\{X_1, \ldots, X_n\}$ est un échantillon aléatoire simple si

- un membre particulier de la population est choisi, cela n'affecte pas les chances qu'un autre membre soit choisi.

- chaque membre de la population a la même probabilité d'être choisi.

En d'autres termes...

$\{X_1, \ldots, X_n\}$ sont indépendants

$\{X_1, \ldots, X_n\}$ sont identiquement distribués (c'est-à-dire qu'ils ont la même fonction de masse ou fonction de densité de probabilité).

Estimation de μ

Un *estimateur* d'un paramètre est une *statistique* dont la valeur dans l'échantillon est utilisée pour estimer ce paramètre.

Un estimateur pour μ est la moyenne de l'échantillon, $\bar{X}$.

Un estimateur pour σ est l'écart-type de l'échantillon, s.

Exemples :

$X_1, \ldots, X_n$ est un échantillon de n foyers américains.

Un estimateur pour μ sera

$$\bar{X} = \frac{x_1 + \ldots + x_n}{n}$$

Un estimateur pour σ sera $s = \sqrt{\frac{1}{n-1} \sum_{\text{pour tous } x} (x - \bar{X})^2}$.

Propriétés de $\bar{X}$

Question

Est-ce que $\bar{X}$ est un “bon" estimateur pour μ ?

Cela dépend de comment l'échantillon $X_1, \ldots, X_n$ est choisi.

Si $X_1, \ldots, X_n$ sont iid, c'est-à-dire si nous avons un *échantillon aléatoire simple* (EAS), alors la moyenne de l'échantillon $\bar{X}$ possède certaines propriétés très intéressantes.

Espérance et écart-type de $\bar{X}$

$$\mathbb{E}(\bar{X}) = \mu$$

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

qui est remplacé par

$$e.s.(\bar{X}) = \frac{s}{\sqrt{n}}$$

puisque nous ne connaissons pas σ.

Remarque

$\bar{X}$ est une variable aléatoire,

Elle a une distribution, que nous appellerons "distribution d'échantillonnage".

Exemple :

Supposons que plusieurs chercheurs collectent des EAS de foyers (tous avec la même taille d'échantillon n).

Ensuite, chaque chercheur va calculer une moyenne d'échantillon $\bar{X}$ différente, c'est-à-dire que nous aurons le tableau suivant :

Chercheur 1	$\bar{X}_1$
Chercheur 2	$\bar{X}_2$
...	...

Exemple avec Stephen Curry

Supposons que le score du joueur dans chaque match de basket soit distribué normalement avec $X \sim \mathcal{N}(28, 8.5^2)$ (il y a 82 matchs dans une saison).

$\mu = 28$ et $\sigma = 8.5$ $\mathbb{E}(\bar{X}) = 28$ et $\sigma_{\bar{X}} = \frac{8.5}{\sqrt{82}} = 0.939$

Quelle est la probabilité que Steph marque en moyenne plus de 28.33 points ?

$$\begin{aligned}\mathbb{P}(\bar{X} \geq 28.33) &= \mathbb{P}\left(Z \geq \frac{28.33 - 28}{8.5/\sqrt{82}}\right) \\ &= \mathbb{P}(Z \geq 0.352) = 0.3632\end{aligned}$$

Mais que faire si mes variables ne sont pas normalement distribuées ?

Théorème Central Limite (TCL)

Si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires indépendantes et identiquement distribuées tirées de n'importe quelle distribution avec une moyenne μ et un écart type σ qui est connu et fini, et que n est suffisamment grand (c'est-à-dire, $n > 30$), alors

$$\mathbb{E}(\bar{X}) = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Remarque : Le TCL s'applique aux sommes de variables aléatoires iid, $\bar{X}$ étant un cas spécifique !

Quand utiliser le TCL ?

Quelle taille devrait avoir n ?

Le théorème central limite ne dit pas que chaque variable aléatoire individuelle suit une distribution normale.

Le théorème central limite s'applique indépendamment de la distribution de X. Il suffit de connaître son espérance et son écart type.

Exemple : scores d'examen

Les résultats des étudiants au dernier examen final du cours *"Introduction à la statistique"* des années précédentes suivent une distribution avec une moyenne 74 et un écart type 14.

Le professeur a donné un examen final cette année dans une classe de 64 étudiants.

Nous nous intéressons à estimer la probabilité que la moyenne des scores de cette année dépasse 80.

Remarque : Dans le problème, il n'est pas mentionné que les scores d'examen suivent une distribution normale !

Exemple : scores d'examen

X_i= score de l'examen du i-ème étudiant dans la classe de 64 étudiants, $i = 1, \ldots, 64$.

Moyenne des scores d'examen : $\bar{X} = \frac{X_1 + \ldots + X_{64}}{64}$

Hypothèses du TCL ? (iid, moyenne et variance finies, $n>30$)

$$P(\bar{X} > 80) = P\left(Z > \frac{80 - 74}{14/\sqrt{64}}\right) = P(Z > 3.429)$$
$$= 1 - P(Z \leq 3.429) = 1 - 0.9997 = 0.0003$$

Distribution d'échantillonnage de $\bar{X}$

Deux configurations pour la distribution d'échantillonnage de $\bar{X}$:
1) Si $X_i \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

exactement.

2) Si $X_i \sim F$ indépendamment pour tous les $i = 1, \ldots, n$, où F est une distribution inconnue telle que $\mathbb{E}(X_i) = \mu < \infty$, $\sqrt{Var(X_i)} = \sigma < \infty$ et (approximativement) $n > 30$, alors

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

approximativement (avec l'approximation qui s'améliore à mesure que n augmente). Cette approximation est basée sur le TCL.

Proportion d'échantillon

$X \sim \ Binomial(n, p)$

p est un paramètre de la population (généralement inconnu)

Si nous avons X mais pas p,

Proportion de l'échantillon

$$\hat{p} = \frac{X}{n}$$

La proportion de l'échantillon $\approx$ la proportion de la population

$\hat{p}$ est un estimateur de p.

Enquête

Échantillon aléatoire simple

55% sont en faveur d'une déclaration

Répéter l'enquête

56% sont en faveur

52% sont en faveur

...

L'échantillon est aléatoire

Exemple : *Référendum*

55% soutiennent le Brexit

56% soutiennent le Brexit

52% soutiennent le Brexit

...

Proportion d'échantillon

Exemple : Référendum

X = nombre de répondants qui soutiennent le Brexit

n = nombre total de répondants dans l'enquête

$\hat{p} = \frac{1}{n}X$ = proportion de répondants qui soutiennent le Brexit

Nous estimons la proportion de répondants dans la population qui soutiennent le Brexit basée sur la proportion de l'échantillon.

Distribution d'échantillonnage de $\hat{p}$

Rappelons que

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} Z_i}{n}$$

où $\mathbb{E}(Z_i) = p$ et $Var(Z_i) = p(1-p)$.

Espérance

$$\mathbb{E}(\hat{p}) = p$$

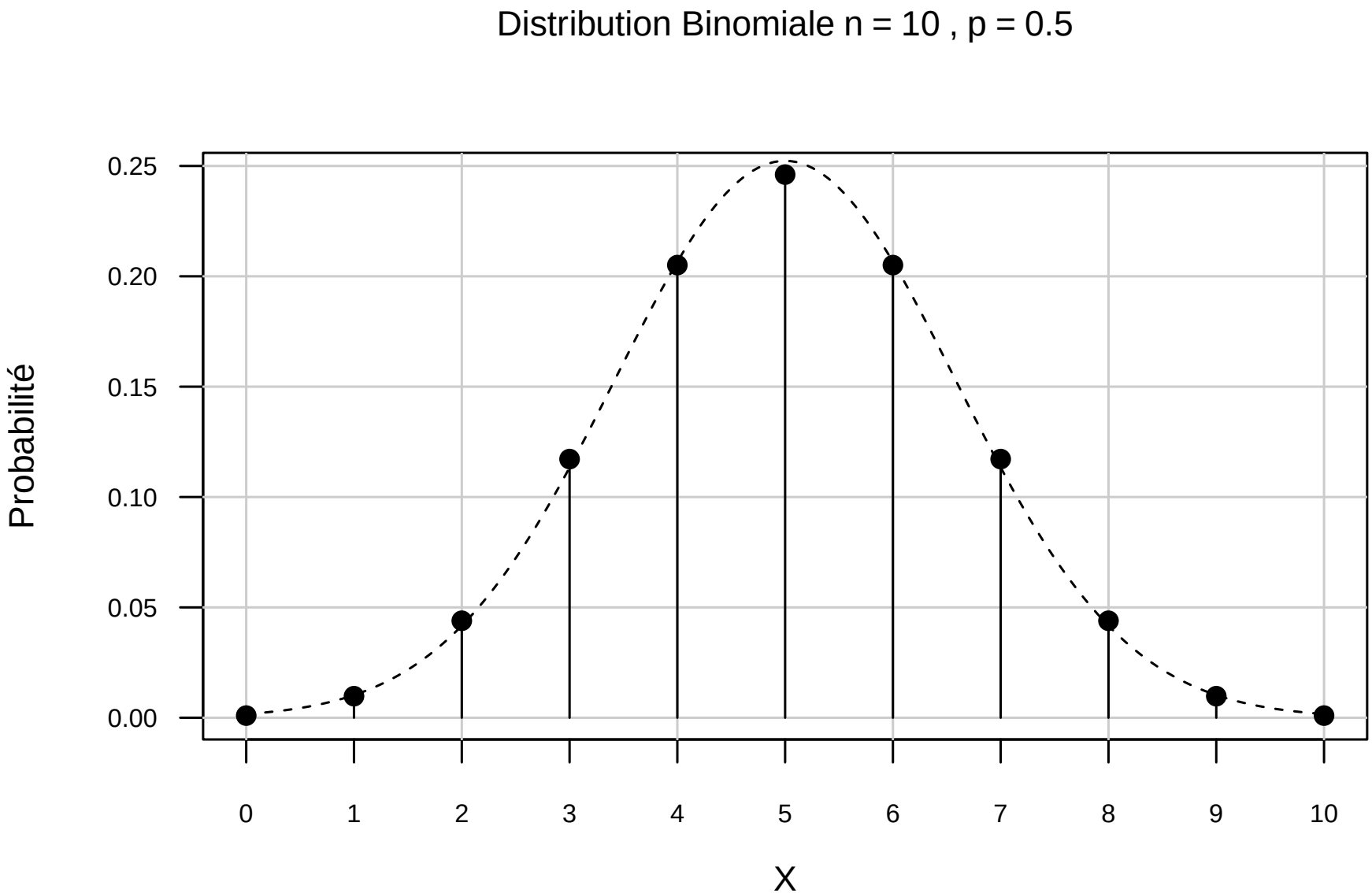
Écart-type

$$\sigma_{\hat{p}} = \sqrt{\frac{p\,(1-p)}{n}}$$

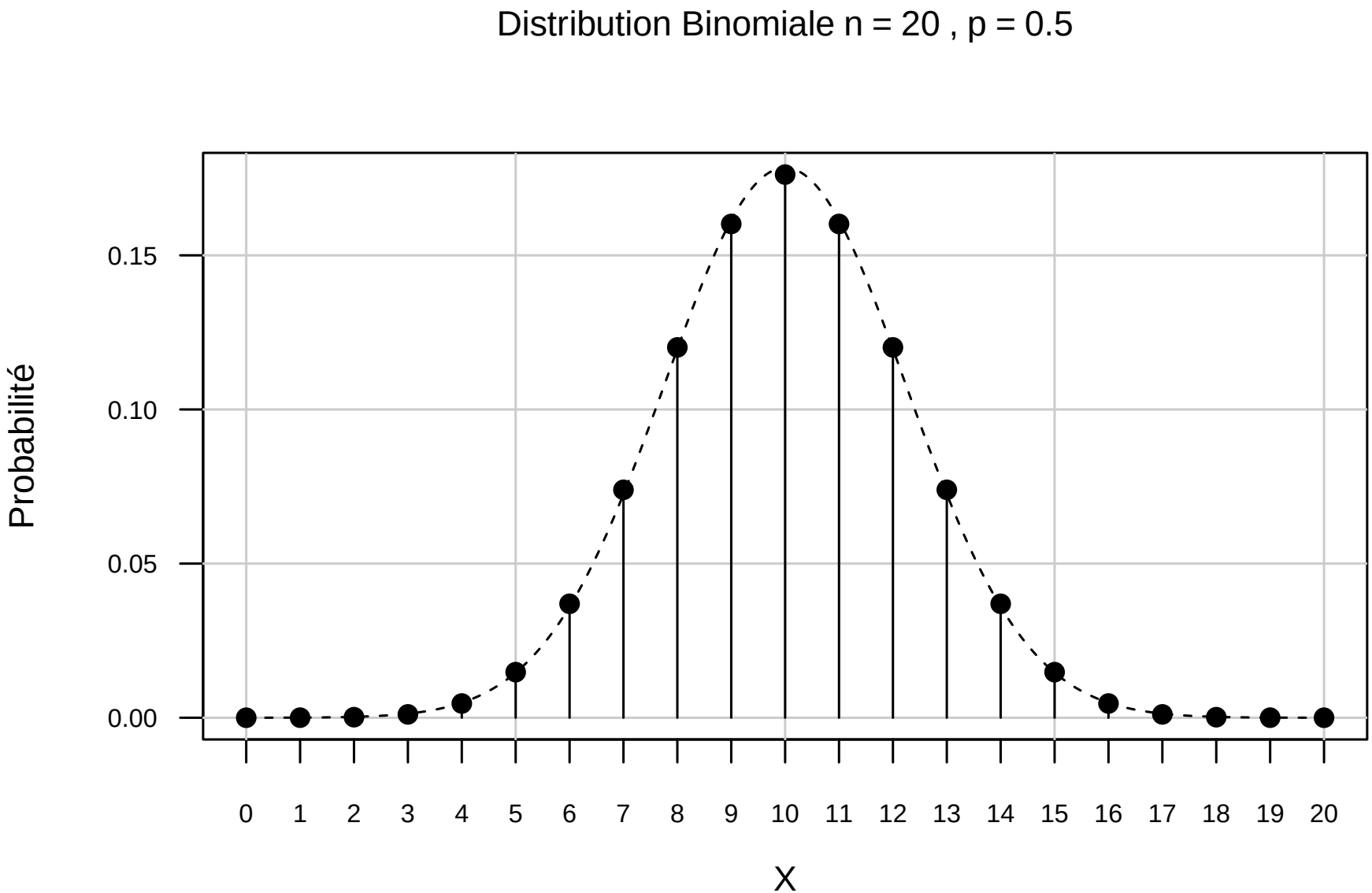
Erreur standard

$$e.s.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

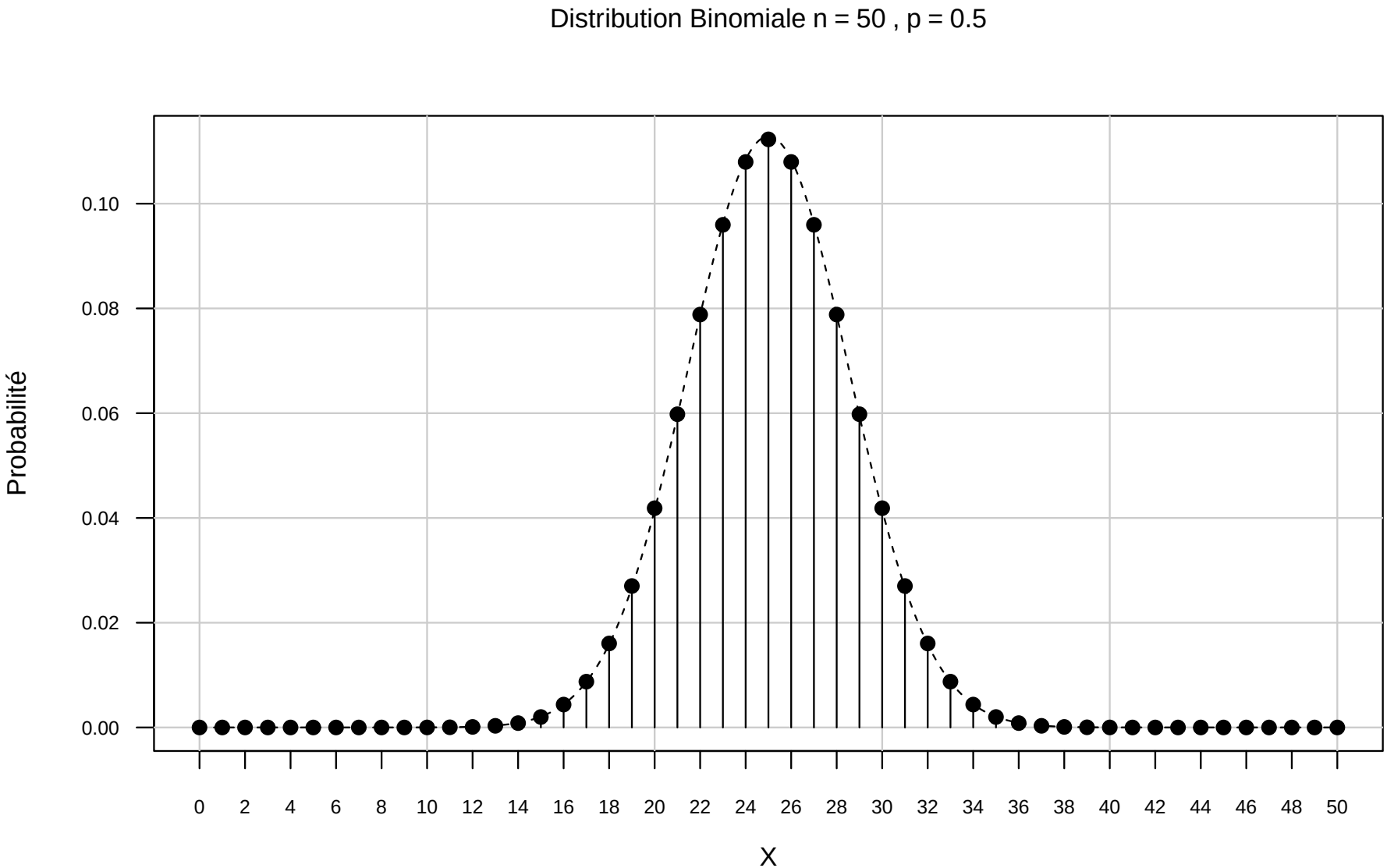
Laissez n augmenter : $n = 10, \quad p = 0.5$



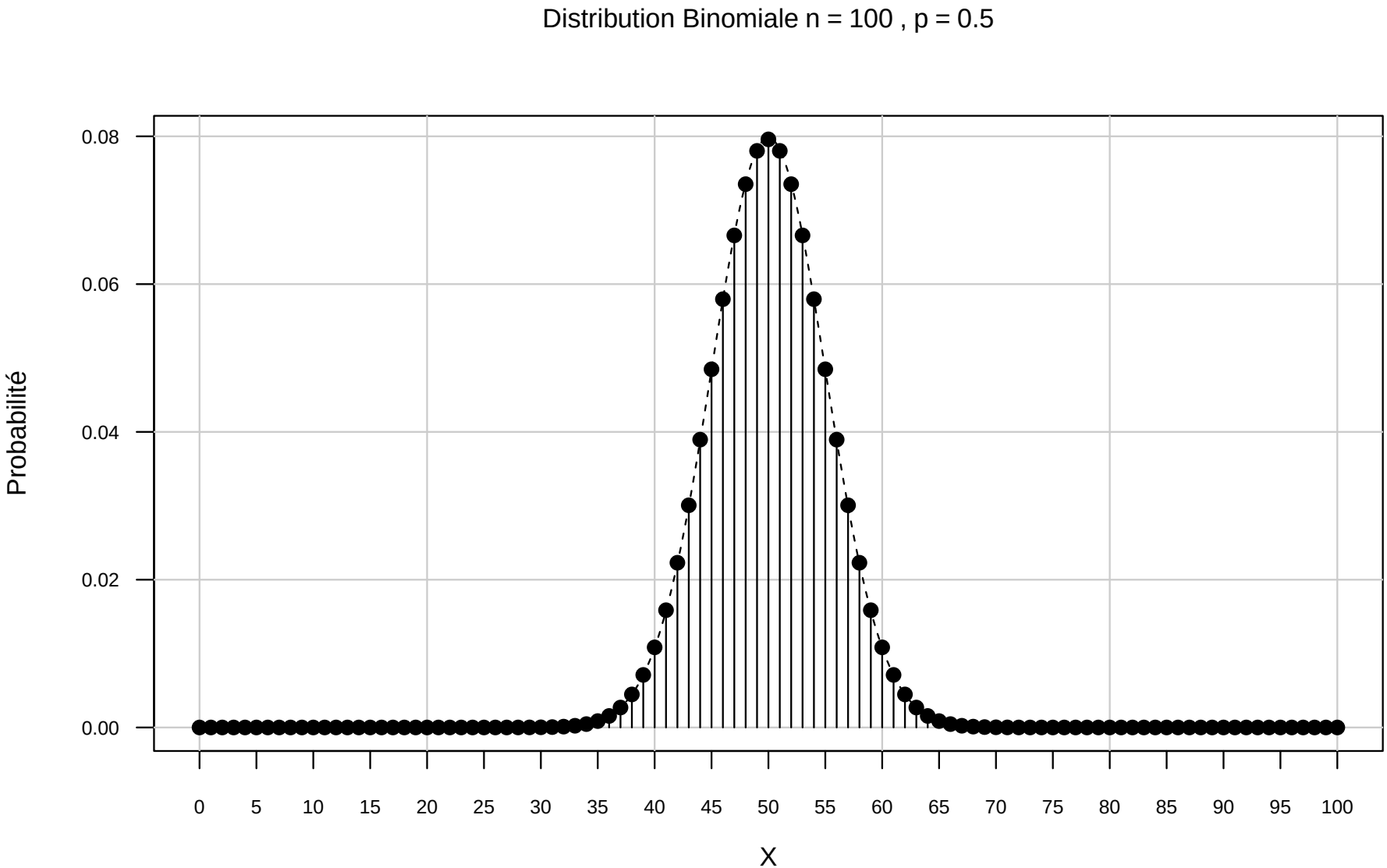
Laissez n augmenter : $n = 20, \quad p = 0.5$



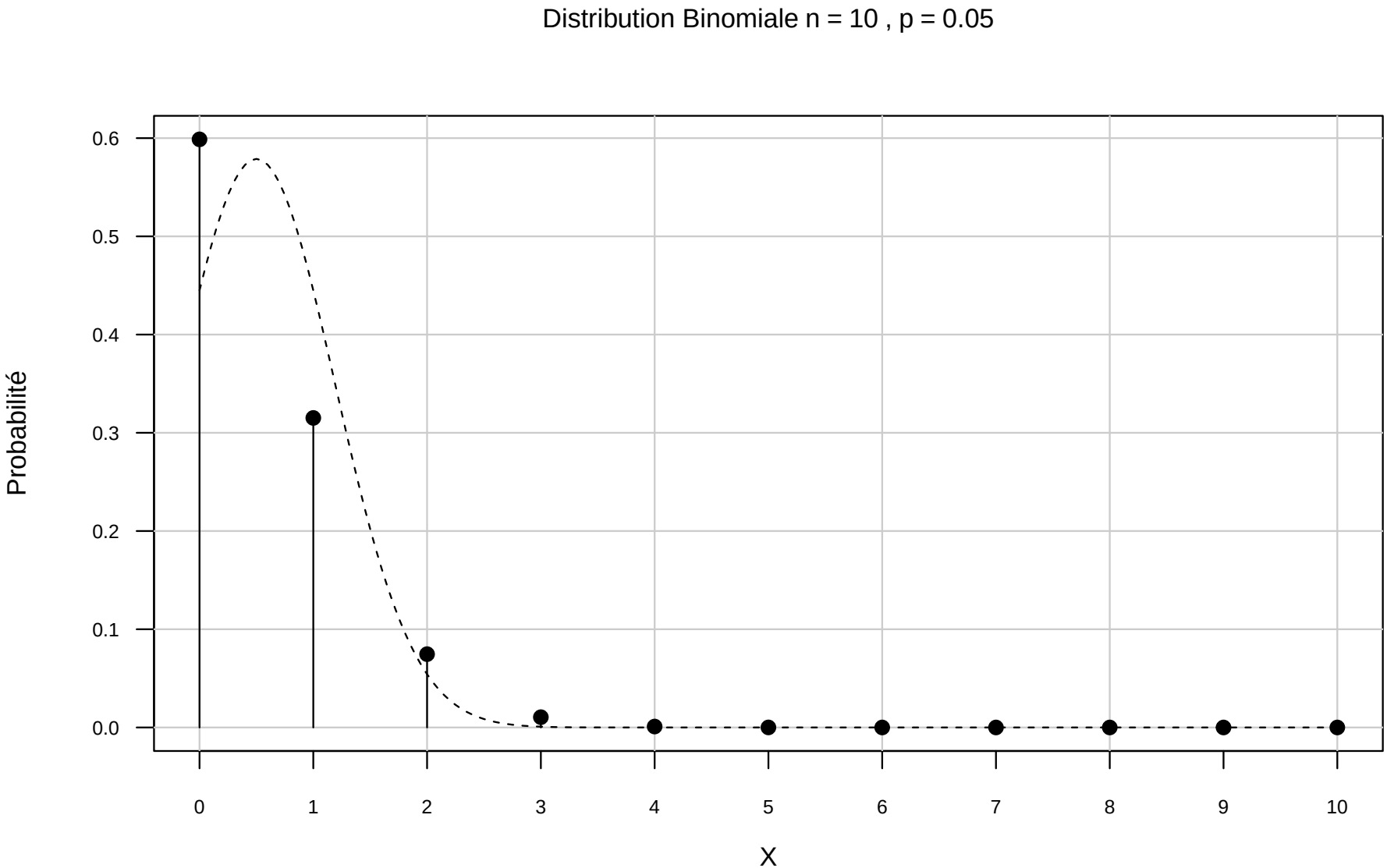
Laissez n augmenter : $n = 50, \quad p = 0.5$



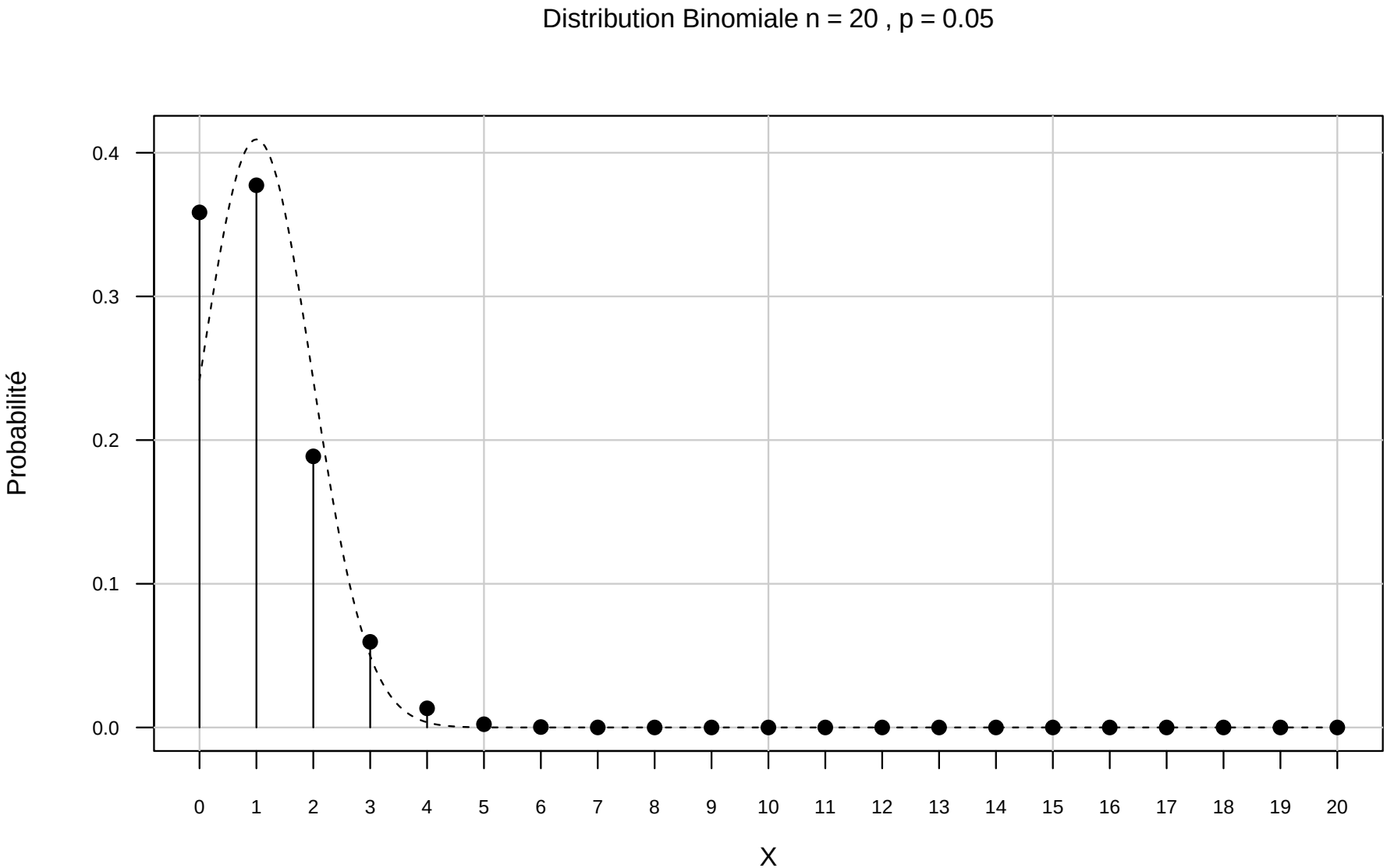
Laissez n augmenter : $n = 100, \quad p = 0.5$



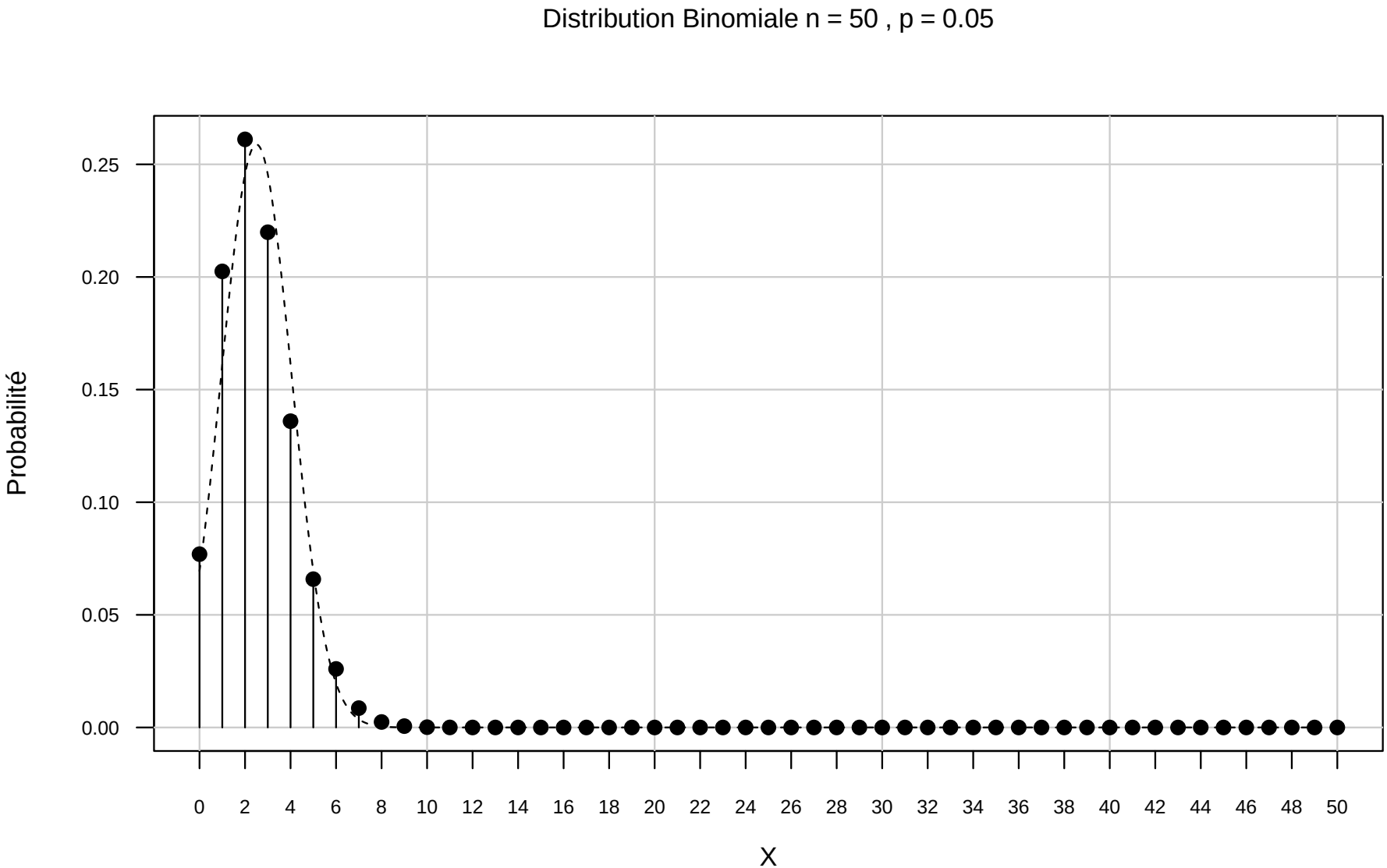
Laissez n augmenter : $n = 10, \quad p = 0.05$



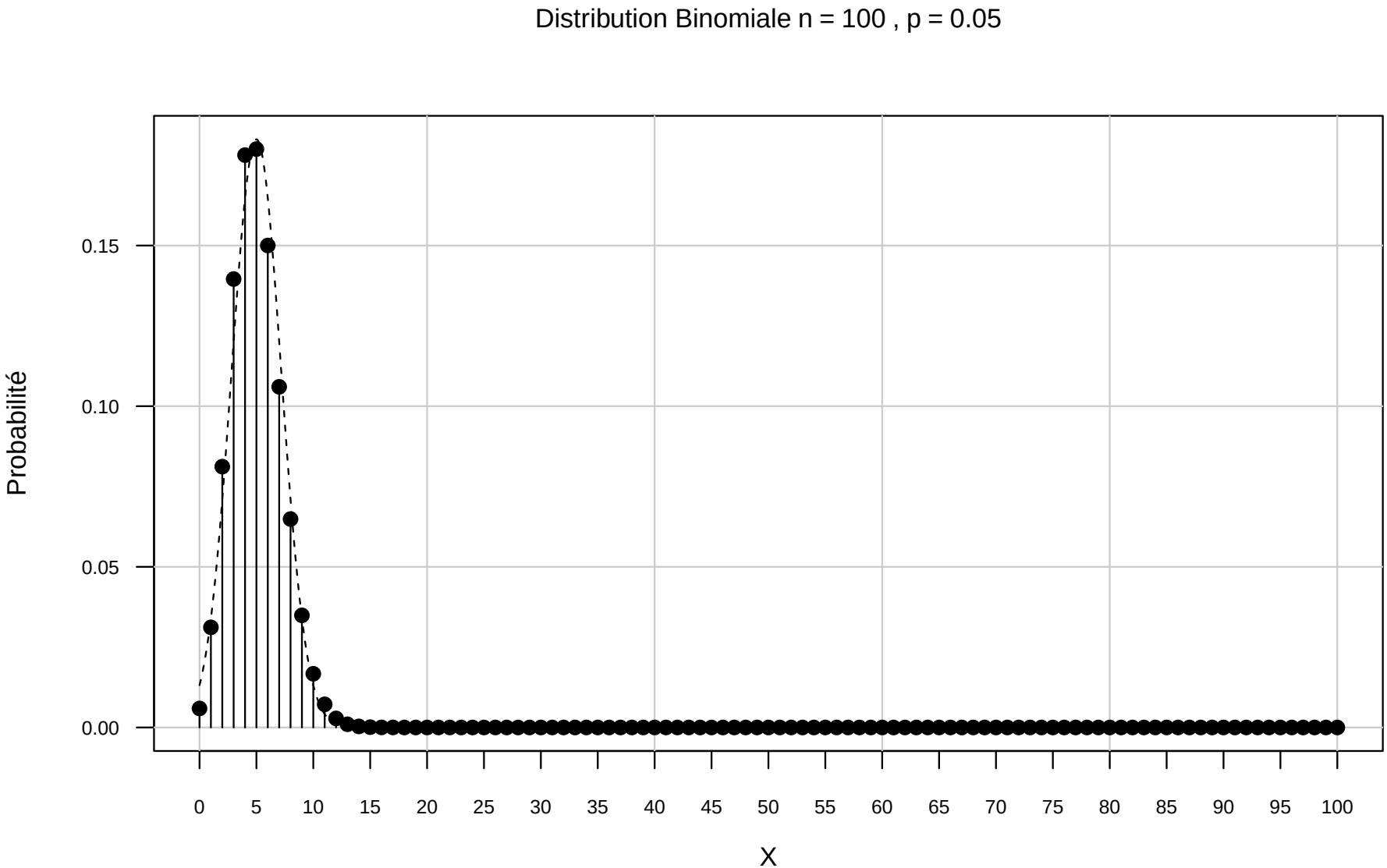
Laissez n augmenter : $n = 20, \quad p = 0.05$



Laissez n augmenter : $n = 50, \quad p = 0.05$



Laissez n augmenter : $n = 100, \quad p = 0.05$



Théorème central limite

Si n est suffisamment grand, la proportion d'échantillon $\hat{p}$ se comporte approximativement comme une variable aléatoire normale avec

Moyenne : $\mu = p$

Écart-type : $\sigma_{\hat{p}} = \sqrt{\frac{p\,(1-p)}{n}}$.

Autrement dit, si n est *suffisamment grand*

$$Z \approx \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p)}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}}$$

$$Z \sim \mathcal{N}(0, 1)$$

Quand pouvons-nous utiliser cette approximation ?

Règle :

Une population avec une proportion fixe

Échantillon aléatoire

Indépendant

Probabilité égale (chance égale)

Taille de l'échantillon suffisamment grande

$np > 10$

$n(1 - p) > 10$

Exemple : lancer une pièce n fois

Supposons que nous lançons une pièce 50 fois (chaque lancer est supposé indépendant des autres) avec $\mathbb{P}(\text{“}\textit{Face}\text{”}) = 0.25$.

Quelle est la distribution de la proportion d'échantillon ?

Quelle est la probabilité d'obtenir plus de 50% de Faces parmi les 50 lancers ?

Exemple : lancer une pièce n fois

$X =$ # de Faces dans les 50 lancers

Nous avons un nombre fixe de lancers ($n = 50$).

Chaque lancer est indépendant des autres.

Il y a deux issues possibles (Face, Pile).

$\mathbb{P}(\text{“}\textit{Succès}\text{”}) = \mathbb{P}(\text{“}\textit{Face}\text{”}) = 0.25$

Il s'agit d'une expérience binomiale !

Vérifiez aussi

$$n\,p = 50 \cdot 0.25 = 12.5 > 10$$
$$n\,(1-p) = 50 \cdot 0.75 = 37.5 > 10$$

(Cela garantit que la taille de l'échantillon est suffisamment grande.)

Exemple : lancer une pièce n fois

$$\mu = p = 0.25$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0.25\,(1-0.25)}{50}} = 0.06$$

$$\hat{p} \sim \mathcal{N}(0.25,\ 0.06^2)$$

$$\begin{aligned} \mathbb{P}(\hat{p} \geq 0.5) &= \mathbb{P}\left(Z \geq \frac{0.5 - 0.25}{0.06}\right) \\ &= \mathbb{P}(Z \geq 4.17) \approx 0 \end{aligned}$$