Théorèmes « limite » en probabilités. Statistiques inférentielles

15 mars 2020

Nous avons déjà vu (*c.f.* TP1 d'info) comment, étant donnée une liste finie de nombres $(x_n)_{n\in\{1,\dots,\mathcal{N}\}}$, issus de mesures sur une population d'individus, on en calcule moyenne $\overline{x}$, écart-type σ_x et variance, médiane et même distribution et histogramme.
Etant donnée une liste finie de couples numériques, $(x_n, y_n)_{n\in\{1,\dots,\mathcal{N}\}}$, on peut de plus calculer covariance, droite de régression d'une variable (en statistiques, on dit aussi un *caractère*) sur l'autre, histogramme du couple, etc..
Ce faisant, on calcule quelques paramètres statistiques de ces séries de données numériques (respectivement *univariée et bivariée*).
La question qui se pose est la suivante : après avoir calculé ces nombres, on fait quoi ? on en conclut quoi ?

Tout tient dans l'interprétation que l'on fait de la population observée : les deux interprétations extrêmes sont les suivantes

1. La population observée est exhaustive.
2. La population observée est un tirage au sort d'individus, tirés au sort *avec remise* dans une population.

Aucune de ces interprétations n'est vraiment réaliste, des interprétations intermédiaires sont possibles : typiquement le tirage au sort sans remise d'une certaine quantité d'individus dans une population finie : ça s'étudie, c'est beaucoup plus compliqué et relève de la théorie des sondages.
Après avoir (brièvement) présenté la première option extrême, nous nous concentrons sur la 2e, qui présente l'avantage de la « simplicité ».

On a réussi à mesurer le *caractère* poids (animal tout juste mort) X de tous les dodos ayant existé. Ceux-ci forment une population *finie* de $\mathscr{N}$ individus.

$$\mathscr{P} = \{D_1, \ldots, D_{\mathscr{N}}\} = \{D_n, n \in \{1, \ldots, \mathscr{N}\}\}$$

Le tableau de données sur le poids de cette population est donné par la suite finie $x = (x_n)_{n \in \{1,\ldots,\mathscr{N}\}}$ où x_n est le poids à sa mort du dodo D_n.

L'interprétation de moyenne $\overline{x}$, écart-type σ_x, histogramme de x d'un point de vue probabiliste sont les suivantes :

1. Je tire au sort un individu, en suivant une loi uniforme sur $\{1, \ldots, \mathcal{N}\}$: celui-ci porte le numéro N. $N \sim \mathcal{U}_{\{1,\ldots,\mathcal{N}\}}$.
2. J'appelle X la variable aléatoire réelle, fonction de N définie par $X = x_N$.
3. On a alors $\mathbb{E}(X) = \overline{x}$, $\sqrt{\mathbb{V}(X)} = \sigma_x$ et
4. l'histogramme de x est une représentation graphique naturelle de la loi de la variable discrète, à nombre fini de valeurs, X.

Hormis les dodos, les pandas et les notes des élèves dans une classe, les données exhaustives étaient plutôt rares jusqu'à présent. La situation évolue avec l'arrivée du « big data ».

L'autre extrême de l'interprétation d'une population est la suivante

- On dispose d'une population « idéale » $\mathscr{P}_\infty$, comportant possiblement une infinité d'individus (que celle-ci soit dénombrable ou encore plus nombreuse).
- Il existe une quantité (un *caractère*) X, que l'on peut mesurer pour chaque individu de la population. L'expérience de tirage au sort d'un individu et la mesure du caractère d'intérêt de cet individu donne la variable aléatoire[1] X.
- X peut-être une v.a prenant un nombre fini de valeurs, une v.a discrète, une v.a à densité, un couplage de telles v.a.,...
- Notre population finie $\mathscr{P}$ de N individus est issue d'un tirage au sort *avec remise* dans la population, x_1 est la valeur de la quantité X pour le 1^{er} individu, x_2, la valeur de X pour le 2^e, etc...

- En conséquence, $(x_1, \ldots, x_N)$ est la suite des valeurs prises sur un tirage au sort d'une suite de v.a $(X_1, \ldots, X_N)$ où $X_1, \ldots, X_N$ sont indépendantes, ayant toutes même distribution que X. (*famille i.i.d.*)
- Une telle famille de v.a. est appelée un (N)-*échantillon* de la variable X. La famille de valeurs $(x_1, \ldots, x_N)$ est une *réalisation* d'un N-échantillon de X.
- Une famille infinie $(X_n)_{n \in \mathbb{N}^*}$ de v.a. i.i.d, de même loi que X est aussi appelée un *échantillon* de la variable aléatoire X. Clairement une sous-famille d'un tel échantillon ne comportant que N membres est un N-échantillon de X.

Définition

La moyenne calculée $\overline{x}$ est la valeur sur un tirage au sort de la variable aléatoire « moyenne empirique » de $X_1, \ldots, X_N$, notée (deux notations) et définie par

$$M_N = \overline{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

Définition

Le carré de l'écart-type calculé σ_x^2 est la valeur sur un tirage au sort de la variable aléatoire « variance empirique » de $X_1, \ldots, X_N$, notée (deux notations) et définie par

$$\Sigma_N^2 = S_N^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - M_N)^2 = \left(\frac{1}{N}\sum_{n=1}^{N}X_n^2\right) - M_N^2 = \overline{X^2}_N - \overline{X}_N^2$$

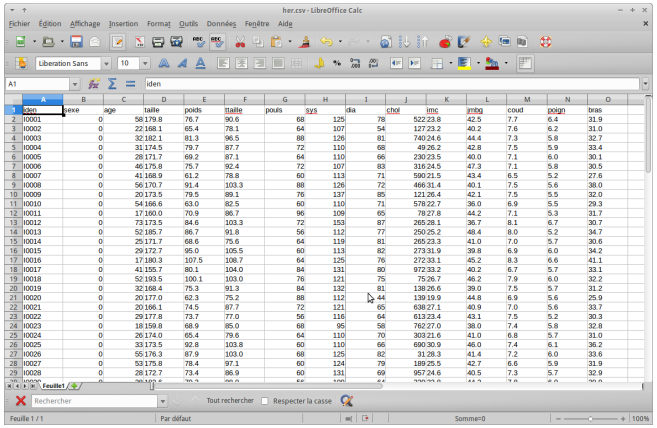

Figure – Les données de fichier `her.csv` : chaque individu est sur une ligne. Une colonne donne les valeurs d'un caractère.

Utiliser le script `python/ExhaustifVSEchantillonnage.py`. Télécharger aussi `her.csv` du même répertoire. Ce fichier contient les données présentées dans la figure 1.

On considère ici que la population représentée dans le fichier `her.csv` est exhaustive et que l'on a donc espérance, écart-type, histogramme de la taille de cette population entière (figure 2). On effectue ensuite divers échantillonnages et on compare moyenne empirique, écart-type empirique, histogramme empirique de ces échantillons avec leurs alter-ego théoriques en figures 3 et 4.

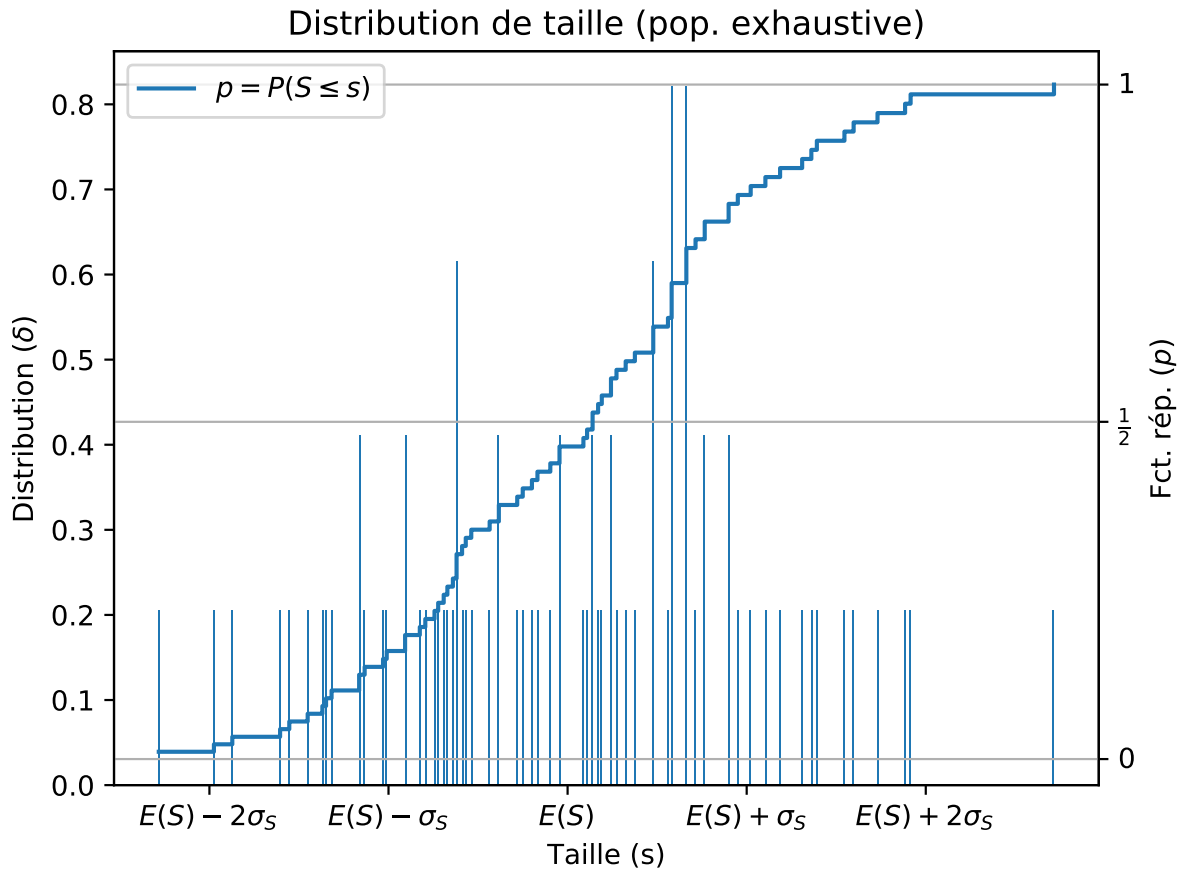

Figure – Une représentation de la distribution exhaustive des tailles.

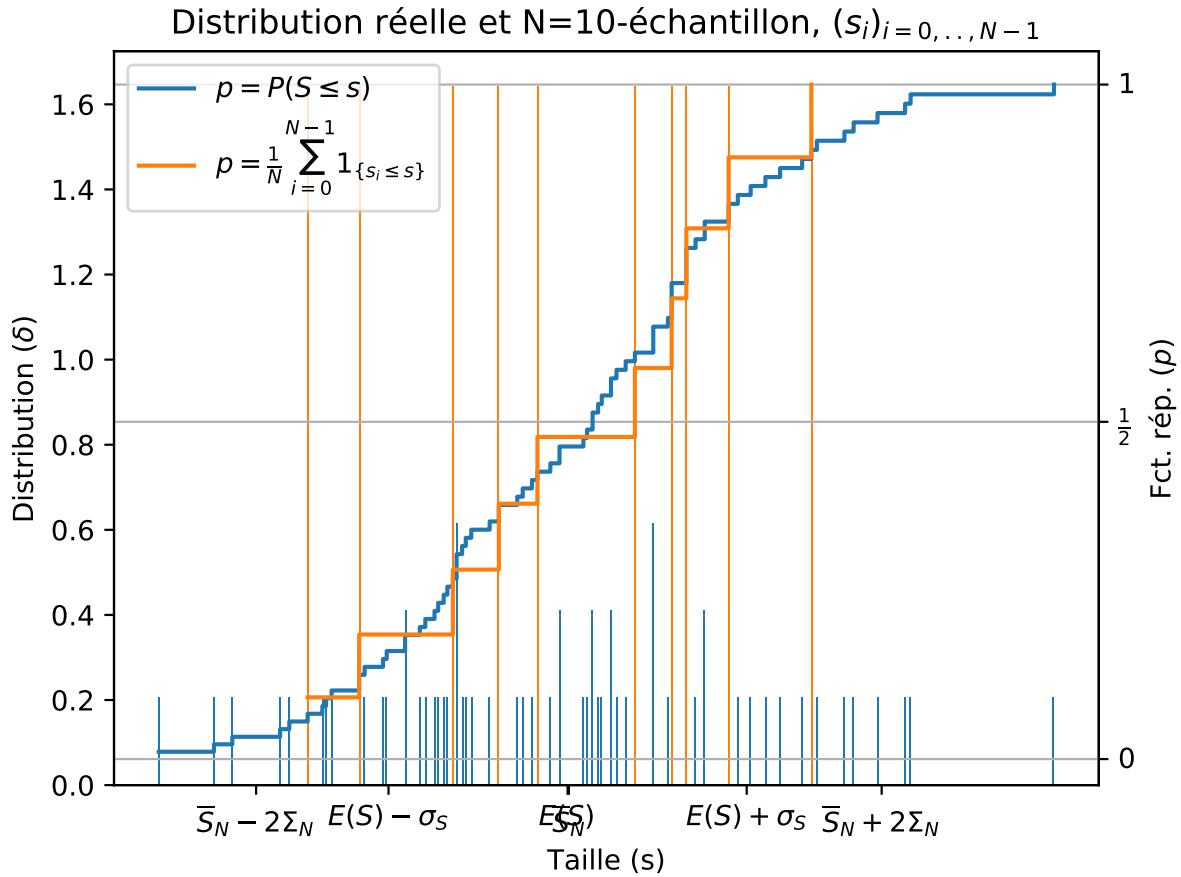

Figure – Comparaison avec une réalisation d'un 10-échantillon.

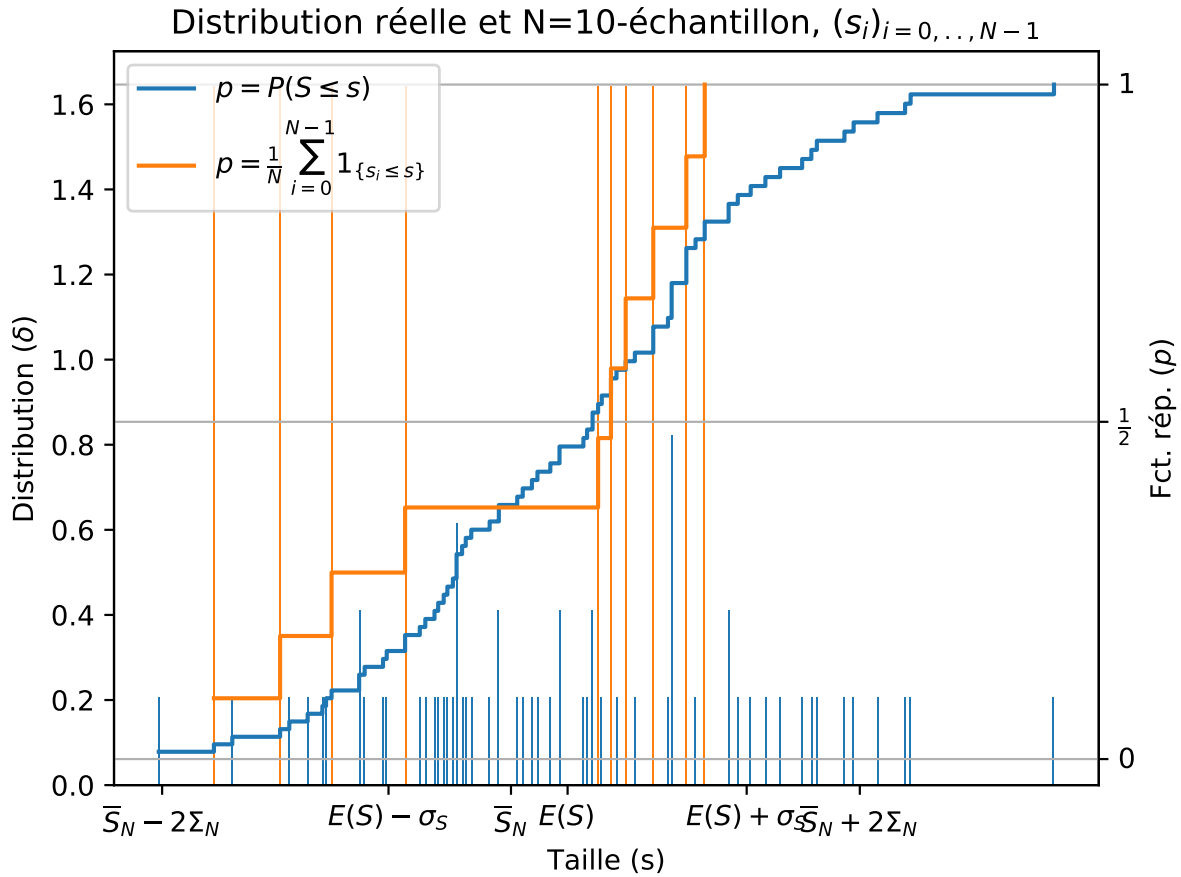

Figure – Autre réalisation d'un 10-échantillon.

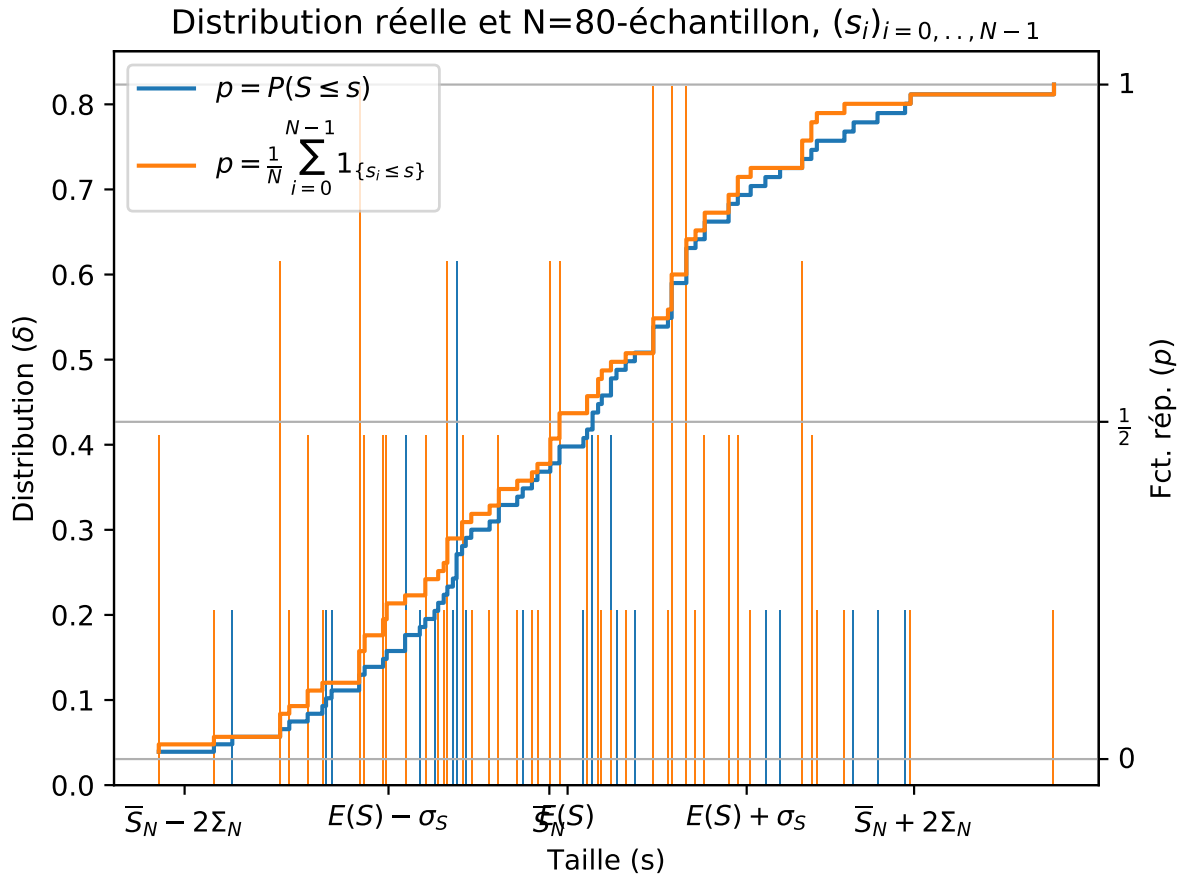

Figure – Comparaison avec une réalisation d'un 80-échantillon.

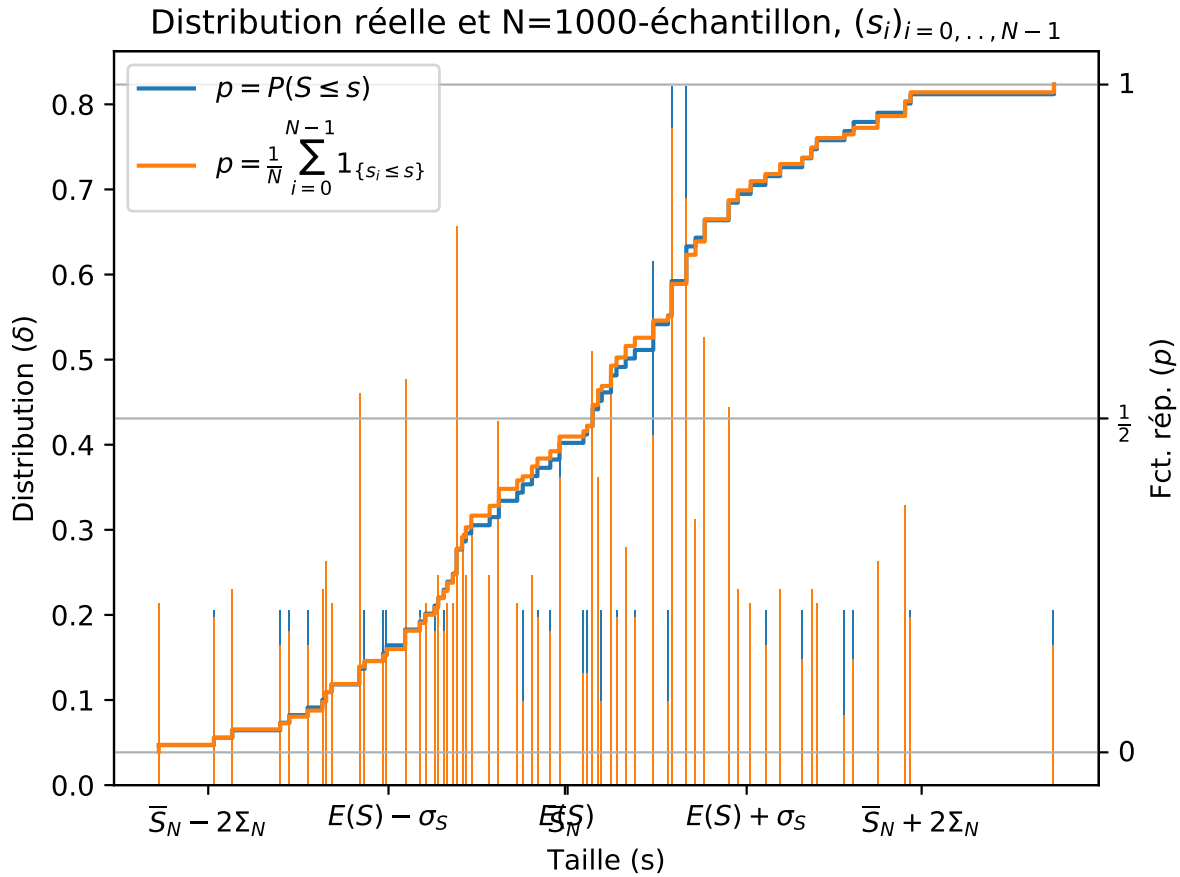

Figure – Comparaison avec une réalisation d'un 1000-échantillon.

Dans une situation réelle, espérance, écart-type, histogramme de la taille d'une population entière sont *inconnus*. Après échantillonnage, on espère que moyenne empirique, écart-type empirique, histogramme empirique de l'échantillons approximent leurs alter-ego théoriques inconnus.

Le but ultime

Spoiler : La théorie dit que *grosso modo* la connaissance d'*une* réalisation d'un échantillon *infini* de X permet de connaitre complètement la loi de X.
Notre but, inaccessible, est de connaitre complètement la loi de X à partir d'une observation d'un échantillon *fini* de X.
Le passage d'infini à fini implique que nous ne pouvons espérer qu'une connaissance approximative de la loi de X et/ou de ses diverses caractéristiques. La signification exacte du sens de ces approximations est à éclaircir. Cela va se traduire par des jeux d'inégalités.
Par modestie, on se limitera à essayer de cerner certaines caractéristiques de cette loi, à commencer par espérance et variance.

Les questions de base sont les suivantes :

1. Quel lien entre M_N et $\mathbb{E}(X)$: peut-on, pourvu que l'échantillon soit assez grand, estimer $\mathbb{E}(X)$, la moyenne « idéale » de X, par les moyennes observées ?
2. Idem pour la variance : Quel lien entre S_N^2 et $\mathbb{V}(X)$: peut-on, pourvu que l'échantillon soit assez grand, estimer $\mathbb{V}(X)$, la variance « idéale » de X, par les variances observées ?

On va voir que

1. M_N est un estimateur[2] de $\mathbb{E}(X)$. L'*erreur d'estimation* est $M_N - \mathbb{E}(X)$ et le *biais* est $\mathbb{E}(M_N - \mathbb{E}(X))$. Dans ce cas précis, le biais est nul.
2. S_N^2 est un estimateur de $\mathbb{V}(X)$. L'*erreur d'estimation* est $S_N^2 - \mathbb{V}(X)$ et le *biais* est $\mathbb{E}(S_N^2 - \mathbb{V}(X))$. Dans ce cas précis le biais n'est pas nul, il vaut $-\frac{\mathbb{V}(X)}{N}$ et tend vers 0 lorsque $N \to +\infty$.

On apportera plusieurs types de réponses à la question vague :

« Quelle est la probabilité que l'erreur soit grande ? »

La statistique *inférentielle* est basée sur le test d'hypothèses : on *infère* une hypothèse et on en déduit *théoriquement* des conséquences satisfaites par la plupart des échantillons. On teste ensuite si notre échantillon satisfait ou pas ces conséquences, ce qui conduit à l'acceptation ou au rejet de l'hypothèse.
Plus précisémment, la démarche est schématiquement la suivante

- On dispose de valeurs observées $(x_1, \ldots, x_N)$ d'un N-échantillon $(X_1, \ldots, X_N)$ de la variable X.
- On énonce clairement les propriétés de base du modèle pour pouvoir travailler : par exemple, « on suppose X de carré intégrable, d'espérance μ ».
- On énonce l'*hypothèse nulle* H_0 qui est l'hypothèse qui va être soumise au test

- On élabore un test numérique, *i.e.* une variable numérique T, fonction de $(X_1, \ldots, X_n)$, un intervalle[3] $I \subset \mathbb{R}$, tels que *théoriquement*, *si* on suppose l'hypothèse nulle *vraie* alors

$$\mathbb{P}(T \in I) \geq 95\%$$

- Sur nos données, on calcule la quantité t, obtenue à partir de $x_1, \ldots, x_N$ de la même manière que T est obtenue à partir de $X_1, \ldots, X_N$. Deux possibilités existent
  1. Soit $t \notin I$: on rejette l'hypothèse H_0
  2. Soit $t \in I$: on « accepte » H_0

La logique de la statistique est la logique des tests : si on passe le test, on peut continuer mais rien ne garantit que c'est *vraiment* bon. Si par contre, on ne passe pas le test, on met le candidat à la poubelle. C'est un peu comme le permis de conduire ou les admissions en école d'ingé...

Reprenons notre exemple concret des tailles d'individus et mettons en oeuvre le *test de conformité de la moyenne*.
On suppose que les données du fichier `her.csv` sont la réalisation d'un 80-échantillon d'une population nettement plus importante. Utiliser le script `TestH0.py`.
Dans ce cas

1. On suppose que S est une variable de carré intégrable, d'espérance μ et que N, le nombre de données est suffisamment important.
2. l'hypothèse nulle est $(H_0) : \mu = 170$
3. On démontre (95% du reste du cours est consacré à cette question) que sous cette hypothèse,

$$\mathbb{P}\left(\left|\frac{M_N - 170}{\frac{S_N}{\sqrt{N}}}\right| > 2\right) \leq \mathbb{P}\left(\left|\frac{M_N - 170}{\frac{S_N}{\sqrt{N}}}\right| > 1.960\right) \leq 0.05$$

Théorème (Inégalité de Markov)

Si X est une v.a. positive, on a, pour tout $\lambda > 0$,

$$0 \leq \mathbb{P}(X > \lambda) \leq \frac{\mathbb{E}(X)}{\lambda}$$

Mesurer la différence entre deux v.a. réelles

Pour mesurer l'écart entre deux v.a réelles (de carré intégrable) X et Y, on peut considérer la quantité $\mathbb{E}((X-Y)^2)$. La racine carrée de cette quantité s'appelle l'*écart quadratique moyen* de X et Y.
Pourquoi ? D'abord, le cas de nullité général est intéressant : Si $\mathbb{E}((X-Y)^2)=0$ alors, p.s. $X=Y$.

Proposition (rappel)

Soit X une v.a réelle de carré intégrable. La quantité $\mathbb{E}((X-a)^2)$, dépendant du réel a est minimale pour (et seulement pour) $a = \mathbb{E}(X)$. La valeur minimale est $\mathbb{V}(X) = \mathbb{E}((X-\mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Proposition (Inégalité de Bienaymé–Tchebycheff(rappel))

Soit X une v.a réelle de carré intégrable, $\lambda > 0$, on a

$$\mathbb{P}(|X - \mathbb{E}(X)| > \lambda) \leq \frac{\mathbb{V}(X)}{\lambda^2}$$

On s'intéresse maintenant au problème de l'approximation d'une distribution de probabilités par une suite de telles distributions.
Cette question intervient à tous étages de la théorie

- Un moyen de dire qu'une v.a est proche d'une constante est d'affirmer la proximité de sa loi avec la loi de la variable constante (Distribution « piquée » en un point.)
- L'importance de la distribution Gaussienne tient à ce qu'elle est « limite » en ce sens de beaucoup de distributions.
- D'un point de vue pratique, cela permet des simplifications de calculs.

Rappel : la loi hypergéométrique

On s'intéresse au problème suivant qui se rencontre lorsque l'on fait un sondage : on dispose de N objets dont $p.N$ d'un certain type A et $q.N$ d'un autre type B. On choisit au hasard *sans remise* n objets et on cherche la loi du nombre A_s d'objets de type A tirés.
On cherche à comparer cette expérience avec la même expérience *avec remise*, où on cherche la loi du nombre A_r d'objets de type A tirés.
Il est intuitivement assez clair que si $\frac{n}{N}$ est petit les deux lois obtenues doivent être assez proches : le fait de remettre ou pas l'objet juste tiré ne peut avoir une grande influence si le nombre d'opérations de tirage n est petit devant la quantité globale d'objets N.

Rappel : la loi hypergéométrique

1. La loi de A_s s'appelle la loi hypergéométrique de paramètres N, n et p. Elle est donnée par

$$\forall \nu \in \{0, \ldots, n\}, \mathbb{P}(A_s = \nu) = \frac{\binom{p.N}{\nu} . \binom{q.N}{n-\nu}}{\binom{N}{n}}$$

2. La loi de A_r est la loi binomiale de paramètres n et p, elle est donnée par

$$\forall \nu \in \{0, \ldots, n\}, \mathbb{P}(A_r = \nu) = \binom{n}{\nu} p^{\nu} q^{n-\nu}$$

Rappel : la loi hypergéométrique

1. $\mathbb{E}(A_s) = n.p$, $\mathbb{V}(A_s) = \frac{N-n}{N-1} n.p.q$
2. $\mathbb{E}(A_r) = n.p$, $\mathbb{V}(A_r) = n.p.q$

Les deux lois ont même espérance mais des variances différentes

On va comparer ces deux distributions lorsque n est fixé et $N \to +\infty$. On a, pour $\nu \{0, \ldots, n\}$,

$$\begin{aligned}
\frac{\mathbb{P}(A_s = \nu)}{\mathbb{P}(A_r = \nu)} &= \frac{(p.N)!(q.N)!(N-n)!}{(p.N-\nu)!(q.N-n+\nu)!N!} p^{-\nu} q^{-n+\nu} \\
&= \frac{(p.N)!}{(p.N)^{\nu}(p.N-\nu)!} \cdot \frac{(q.N)!}{(q.N)^{n-\nu}(q.N-(n-\nu))!} \cdot \cdot \\
&\quad \frac{(N-n)!N^n}{N!} \\
&\to 1
\end{aligned}$$

car chacun des termes de ce produit tend vers 1 lorsque $N \to +\infty$.

On a donc,

$$\forall \nu \in \{0, \ldots, n\}, \mathbb{P}(A_s = \nu) \overset{N \to +\infty}{\to} \mathbb{P}(A_r = \nu)$$

Ceci implique que pour toute fonction $h : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(h(A_s)) \overset{N \to +\infty}{\to} \mathbb{E}(h(A_r))$$

On trouve dans certains ouvrages ou sur Wikipedia, que l'on peut considérer que l'approximation est « bonne » pour $n/N < 0,1$. Une telle affirmation n'a pas grand sens : Quel est la précision attendue ? Que calcule-t-on ? Dans le cas où $n/N \simeq 0,1$, l'erreur relative sur la variance est de 10%, ce qui n'est pas à proprement parlé négligeable.

Un exercice[4] pour les $\frac{5}{2}$ ou à reprendre en fin d'année :

Exercice 2.— On considère n coureurs numérotés de 1 à n tirant dans une urne un numéro de dossart. Les tirages se font avec remise. Une série correspond à un tirage des n joueurs. Dès qu'un joueur tire son numéro, on s'arrête. On note X_n le nombre de séries que l'on fait.

1. Donner la loi de X_n et son espérance.

2. Soit k un entier naturel non nul, montrer que la suite $(\mathbb{P}(X_n = k))_{n\geq 1}$ converge et donner sa limite, notée p_k.

3. Montrer que $(p_k)_{k\geq 1}$ est la loi de probabilité d'une v.a Y à valeurs dans $\mathbb{N}^*$.

4. Comparer $\mathbb{E}(Y)$ et la limite possible de $\mathbb{E}(X_n)$.

Démo graphique

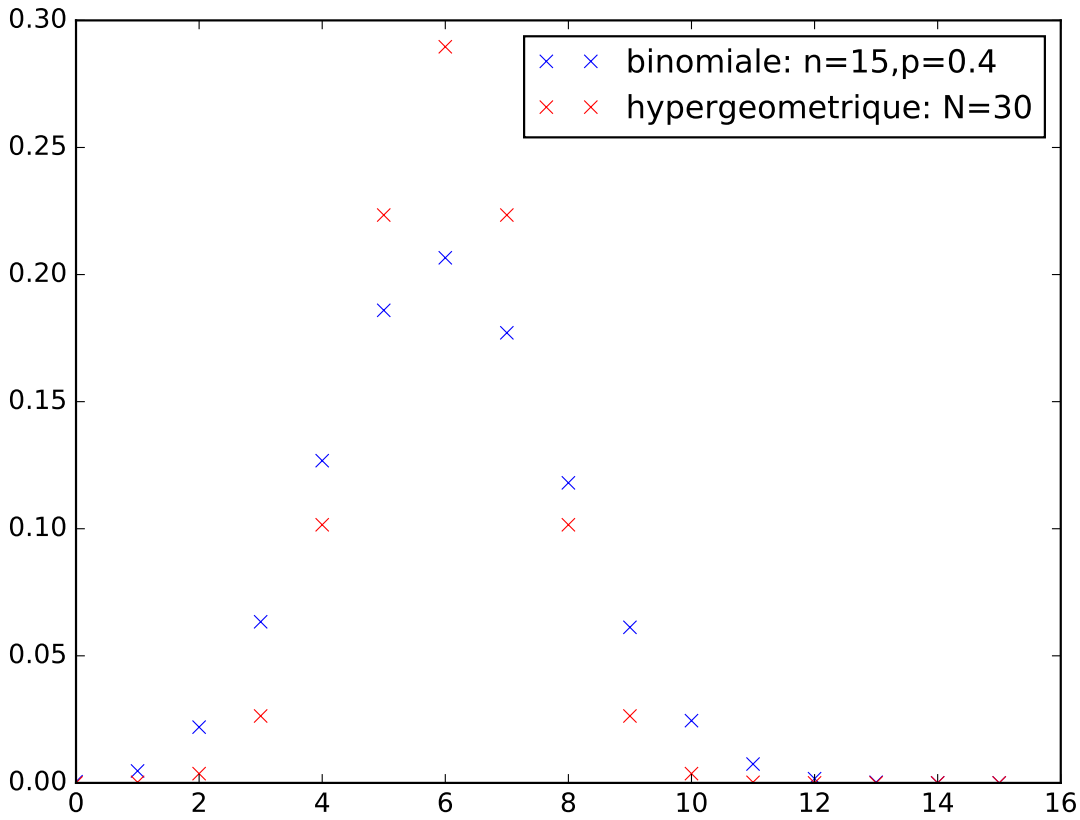

Figure – Comparaison Hypergéométrique/Binomiale

Démo graphique

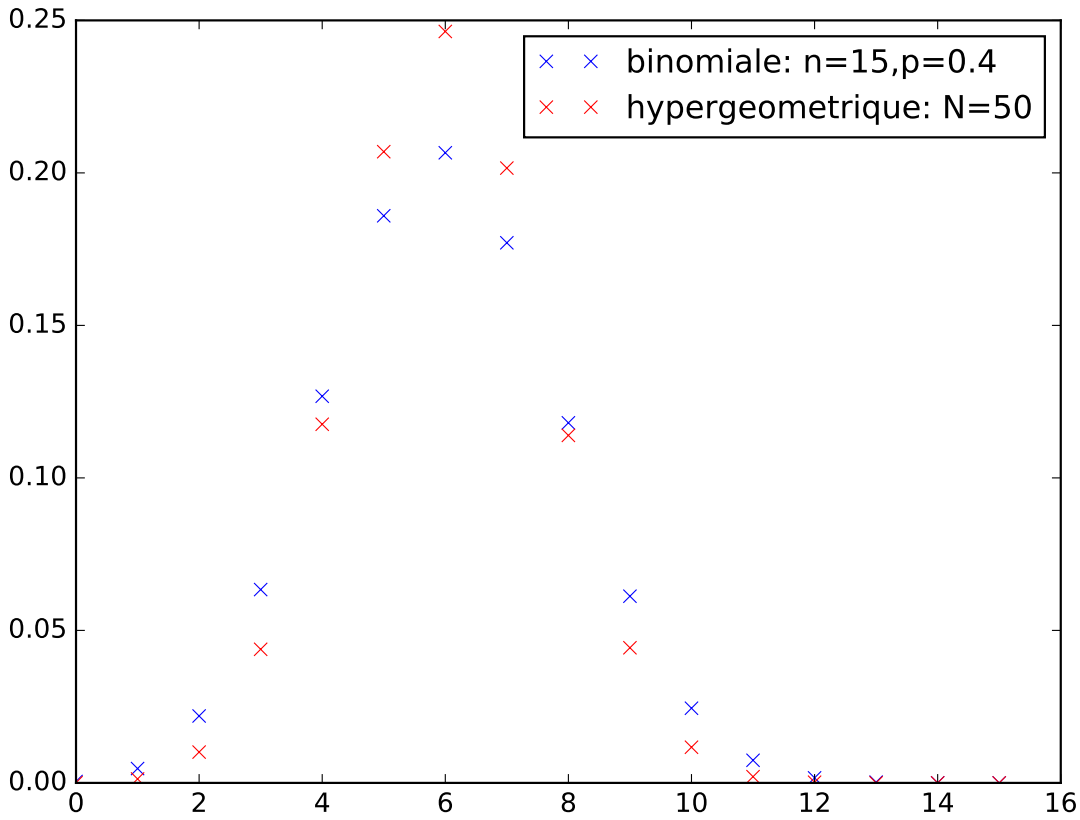

Figure – Comparaison Hypergéométrique/Binomiale

Démo graphique

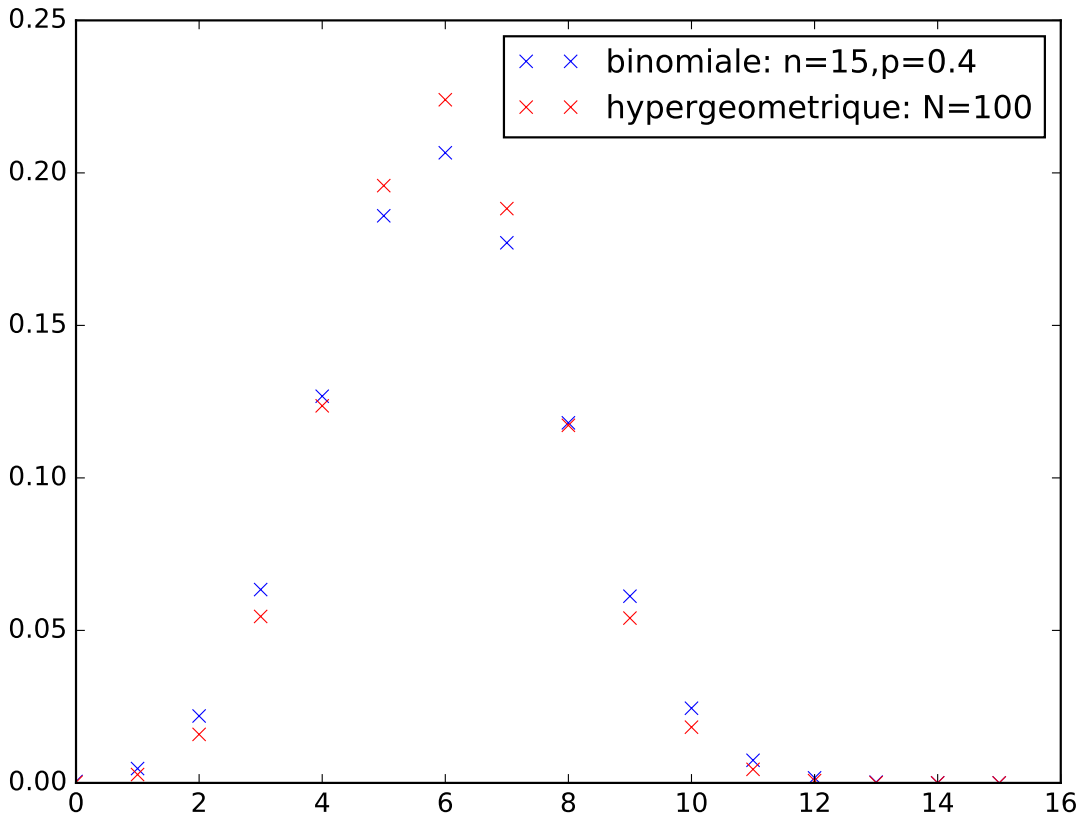

Figure – Comparaison Hypergéométrique/Binomiale

Démo graphique

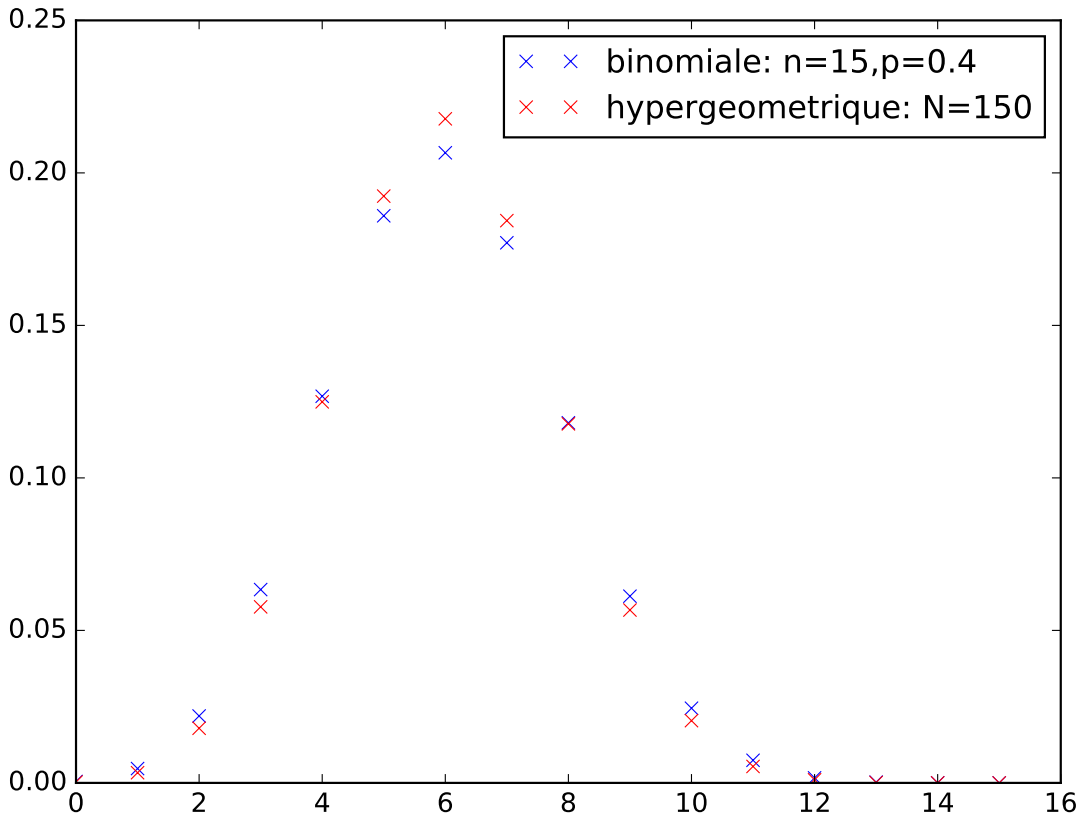

Figure – Comparaison Hypergéométrique/Binomiale

Démo graphique

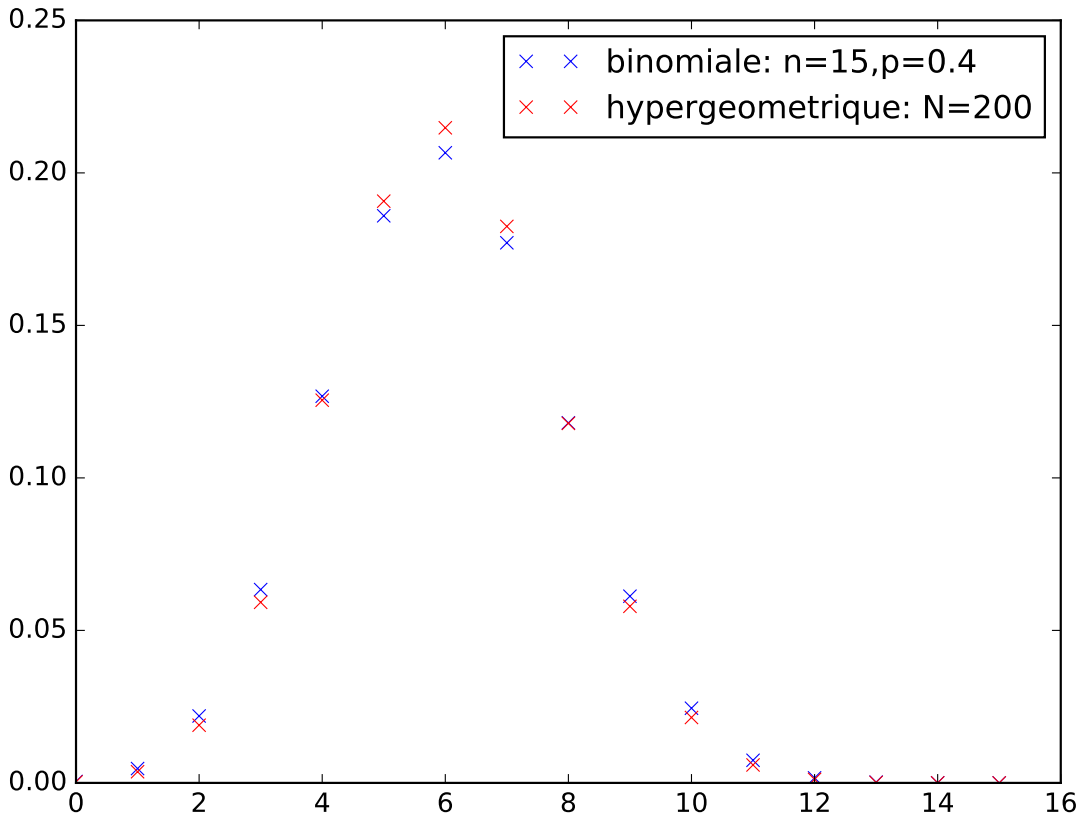

Figure – Comparaison Hypergéométrique/Binomiale

Démo graphique

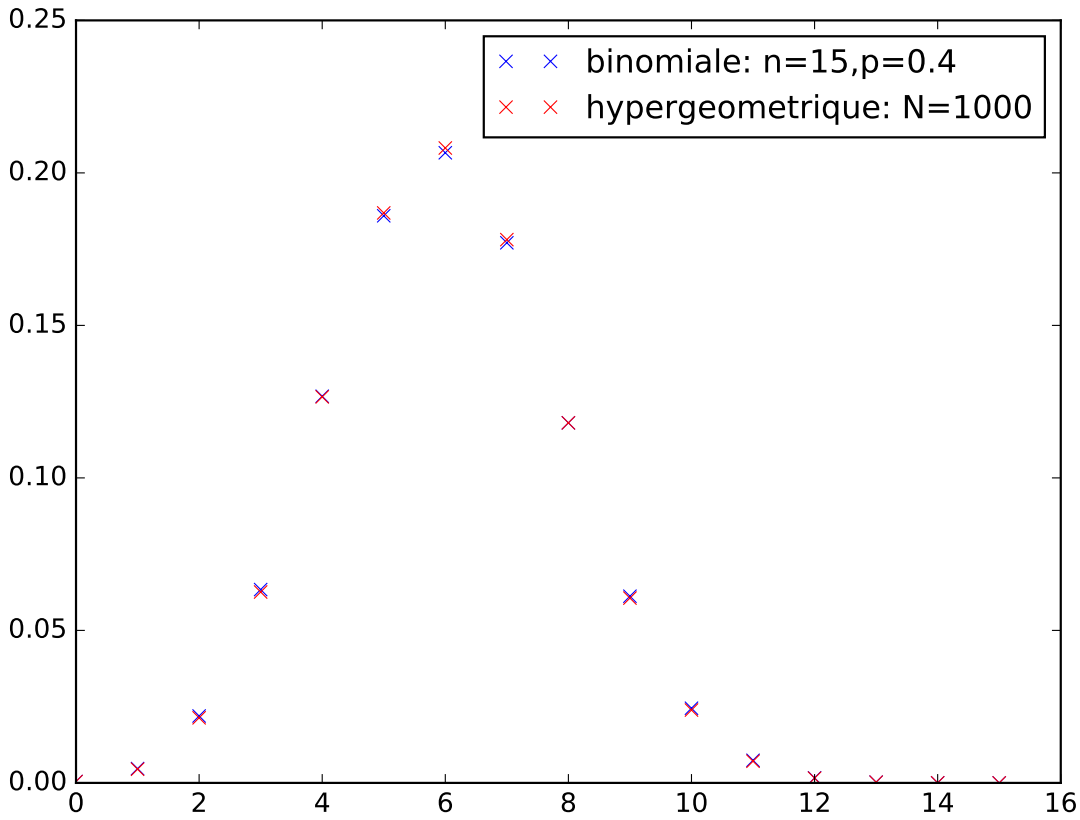

Figure – Comparaison Hypergéométrique/Binomiale

Approximation binomiale de la loi de Poisson

Soit $X \sim \mathscr{B}(n,p)$. On verra dans le chapitre sur les lois discrètes que si $n >> 1$, $n.p \sim \lambda$ alors, pratiquement, X suit approximativement une loi de Poisson.
Mathématiquement,

Proposition

Soit $\lambda > 0$, $p_n \in]0,1[$, $p_n \overset{n\to+\infty}{\to} 0^+$, *tels que* $n.p_n \to \lambda$. *Pour tout* $k \in \mathbb{N}$,

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

Démonstration sur le poly.

Lois des extrêmes

Exercice 2.— Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a. indépendantes, uniformément distribuées sur $]0,1[$. On pose, pour $N \in \mathbb{N}^*$,

$$U_N = \min_{1\leq n\leq N} X_n \text{ et } V_N = \max_{1\leq n\leq N} X_n$$

alors, asymptotiquement, lorsque $N \to +\infty$, U_N est proche de la constante 0 et V_N est proche de la constante 1.

Exercice 3.—

1. Donner espérance et variance de V_N (définie dans l'exercice 2) ainsi que des équivalents simples v_N et s_N^2 de ces quantités lorsque $N \to +\infty$.

2. Soit [5] $V_N^* = \frac{1}{s_N}(V_N - v_n)$. Déterminer sa fonction de répartition F_N^* déterminer, pour $v \in \mathbb{R}$, $F^*(v)$ la limite de $F_V^*(v)$ lorsque $N \to +\infty$ et montrer que F^* est fonction de répartition d'une variable à densité.

3. Interpréter graphiquement ?

Exercice 5.— On suppose que X est une v.a.r de loi $\mathscr{E}(1)$ et que $(X_n)_{n\in\mathbb{N}^*}$ est un échantillon de X.
On pose, pour $n \in \mathbb{N}^*$,

$$M_n = \max(X_1, \ldots, X_n)$$

1. Donner la fonction de répartition de M_n et montrer qu'une densité m_n de M_n est donnée par la formule

$$\forall x \in \mathbb{R},\ m_n(x) = ne^{-x}(1-e^{-x})^{n-1}\mathbb{1}_{\{x\geq 0\}}$$

2. Quel est le maximum α_n de m_n ?
3. On pose $Y_n = M_n - \alpha_n$. Calculer F_n, la fonction de répartition de Y_n ainsi que la limite $F(x)$ de $F_n(x)$ lorsque $n \to +\infty$ pour tout $x \in \mathbb{R}$.
4. Vérifier que F est une fonction de répartition. On se donne Y_∞, v.a.r répartie suivant F. Donner une densité de Y_∞. Tracer son graphe.

La forme des coefficients binomiaux

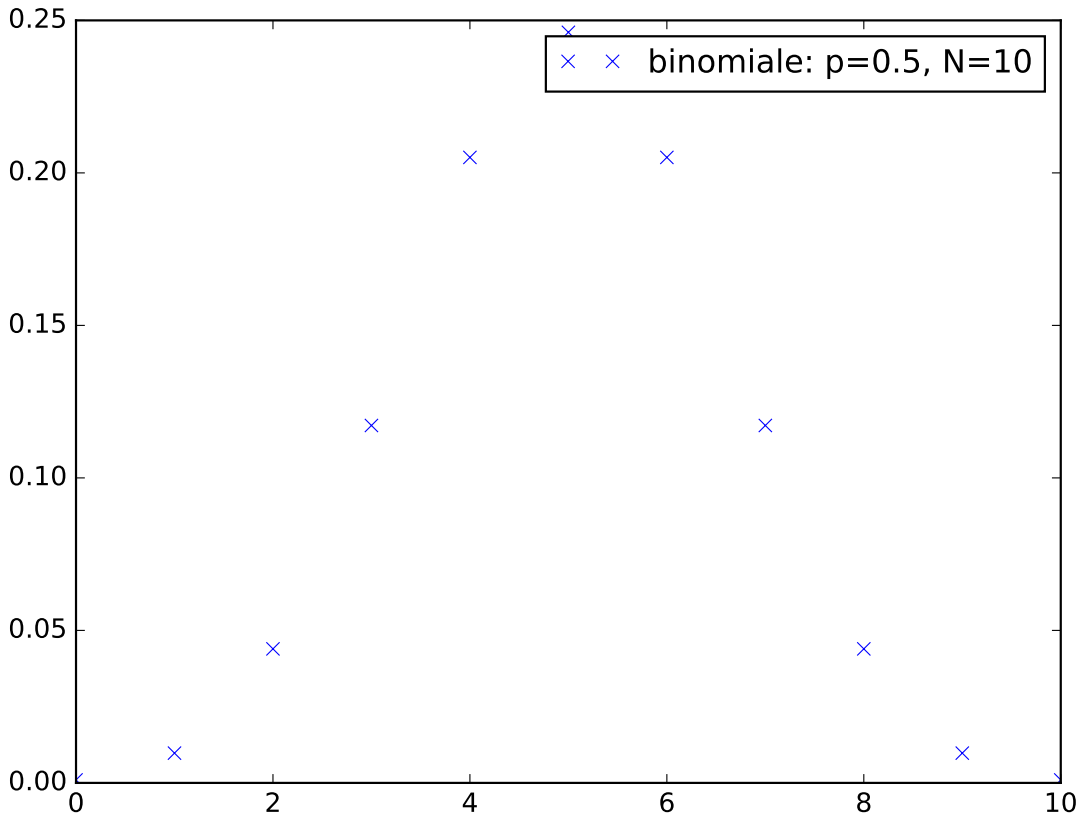

Figure – Binomiale

La forme des coefficients binomiaux

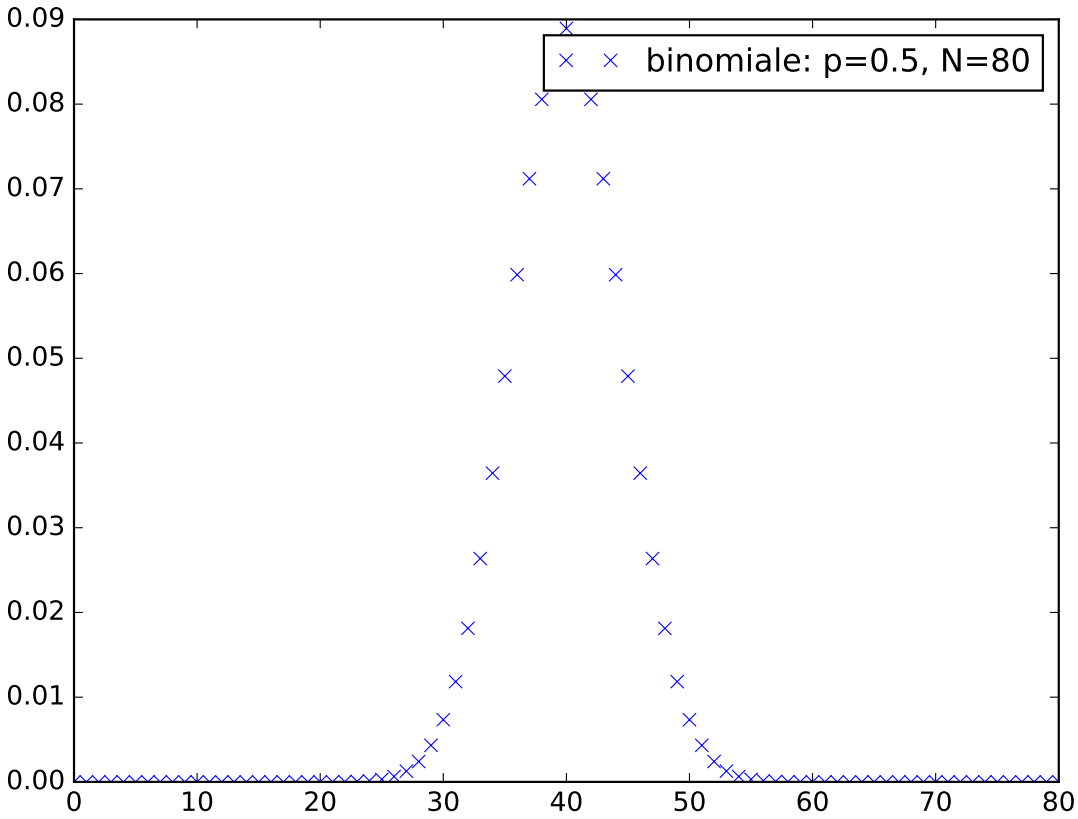

Figure – Binomiale

La forme des coefficients binomiaux

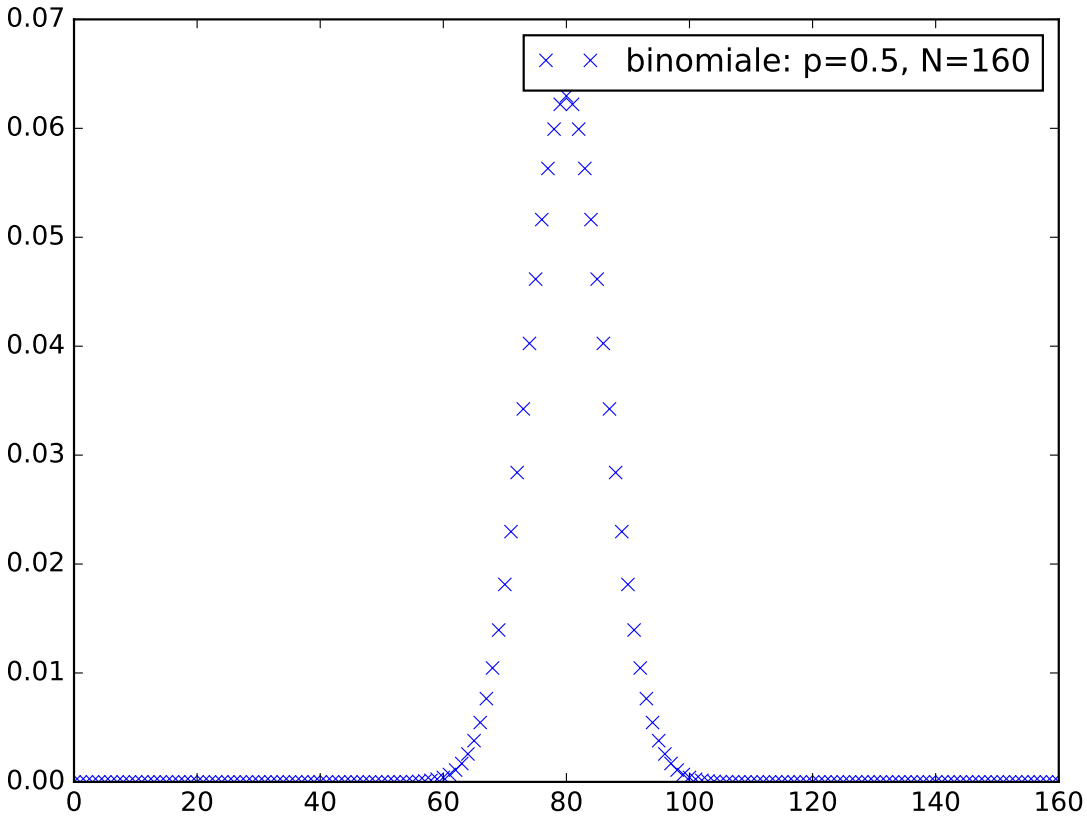

Figure – Binomiale

La forme des coefficients binomiaux

Si $S \sim \mathscr{B}(n,p)$, $0<p<1$, $n>>1$, alors $X = \frac{S-\mathbb{E}(S)}{\mathbb{V}(S)^{\frac{1}{2}}}$ vérifie $\mathbb{E}(X)=0$, $\mathbb{V}(X)=1$ et est distribuée approximativement comme une $\mathscr{N}(0,1)$.

1. Noter la renormalisation : X vérifie $\mathbb{E}(X)=0$, $\mathbb{V}(X)=1$.
2. Le sens de « distribué approximativement comme » s'agissant d'une part de v.a discrète et d'autre part de v.a. à densité est à préciser. Cela signifie dans le cas présent

$$\mathbb{P}(a \leq X \leq b) \simeq \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

Une série de graphiques

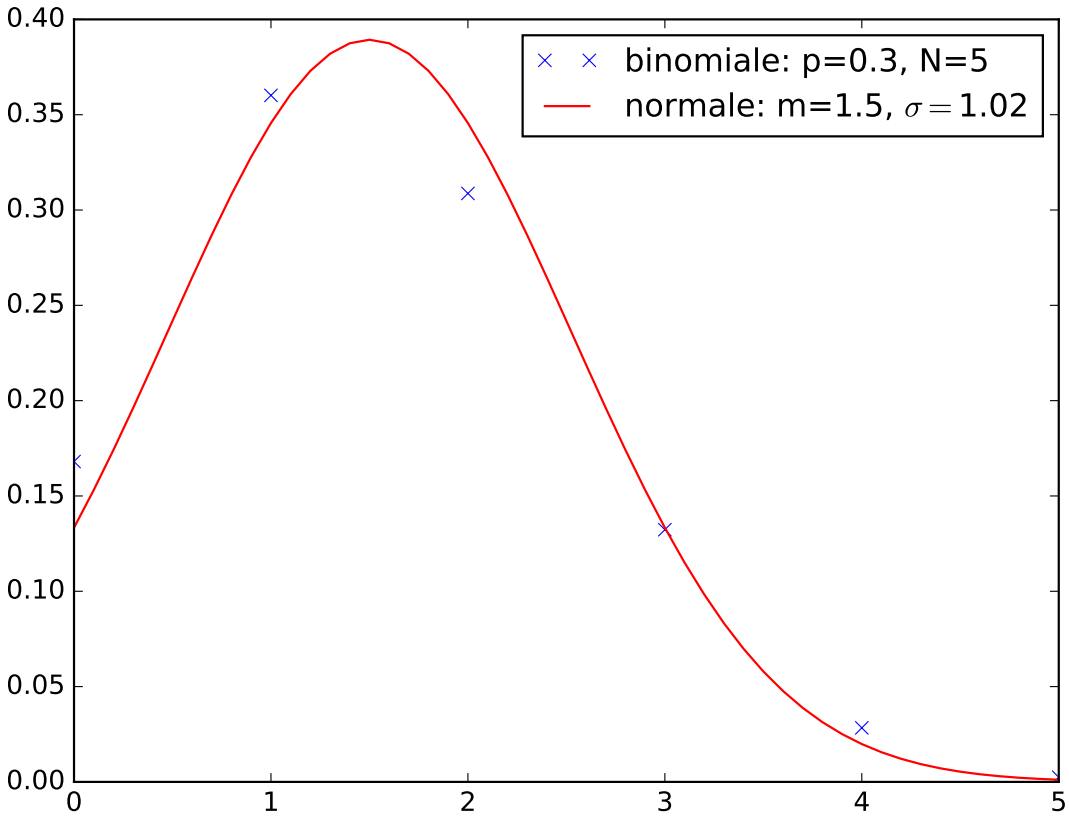

Figure – Comparaison Binomiale/Normale

Une série de graphiques

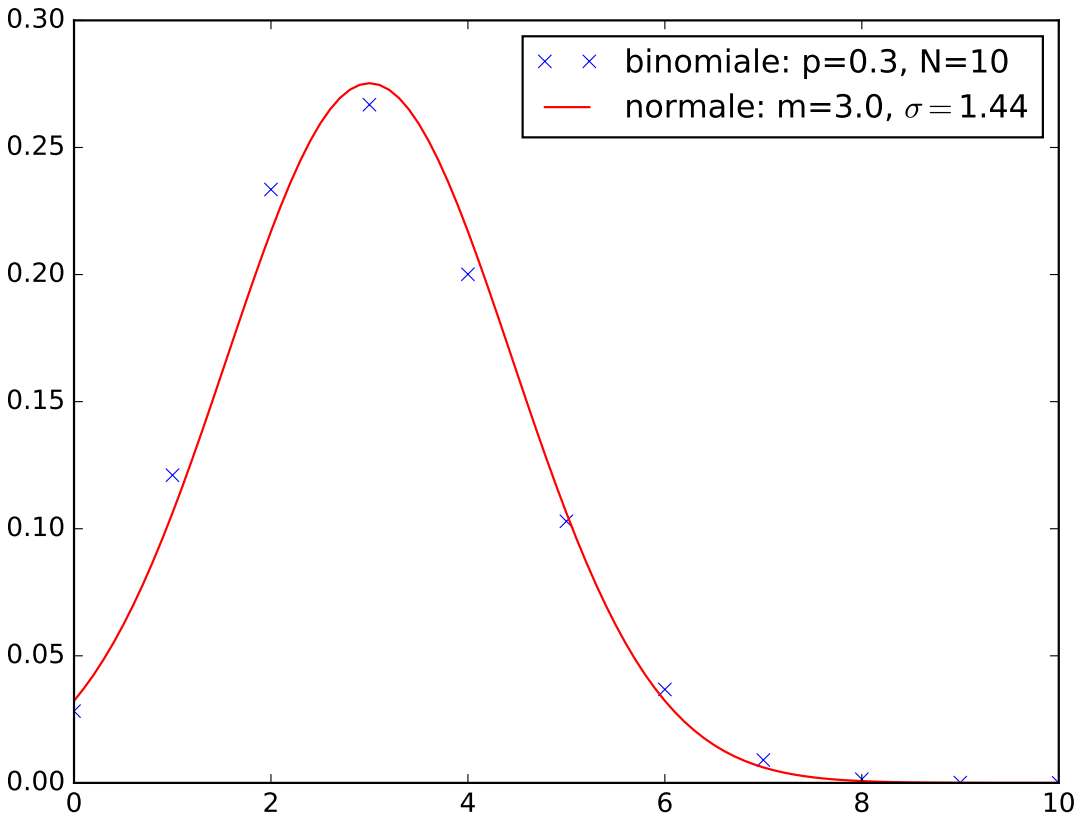

Figure – Comparaison Binomiale/Normale

Une série de graphiques

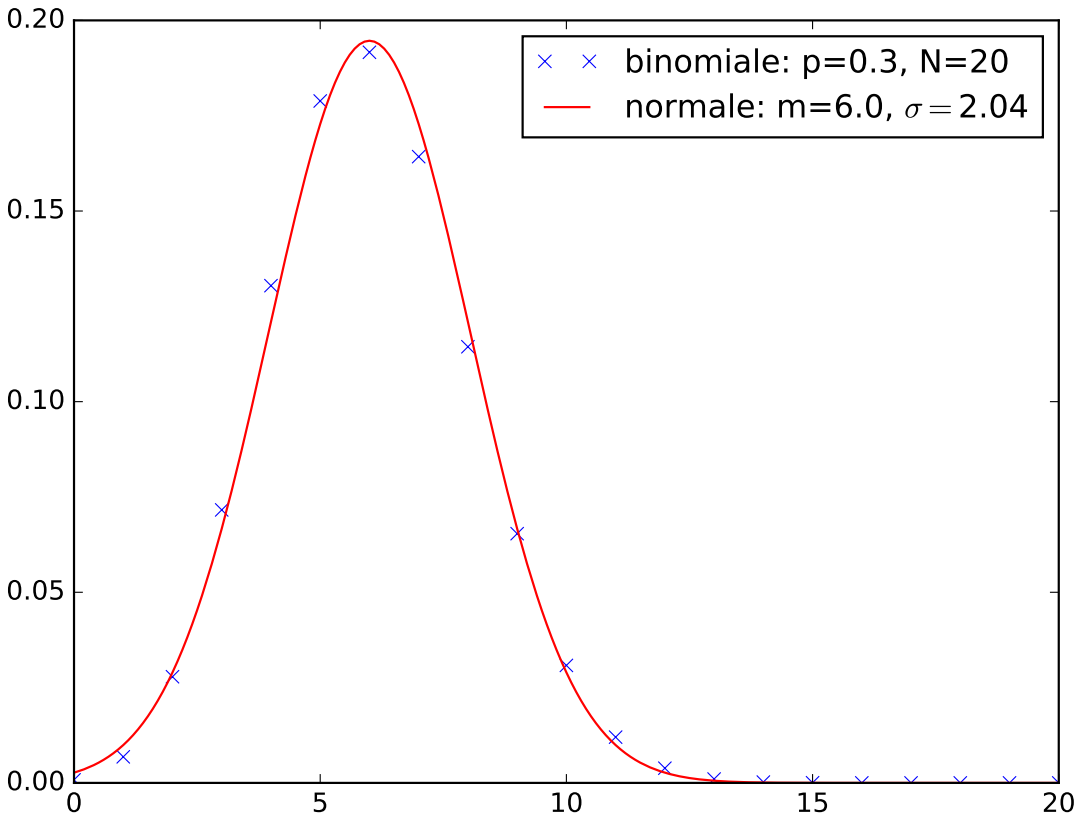

Figure – Comparaison Binomiale/Normale

Une série de graphiques

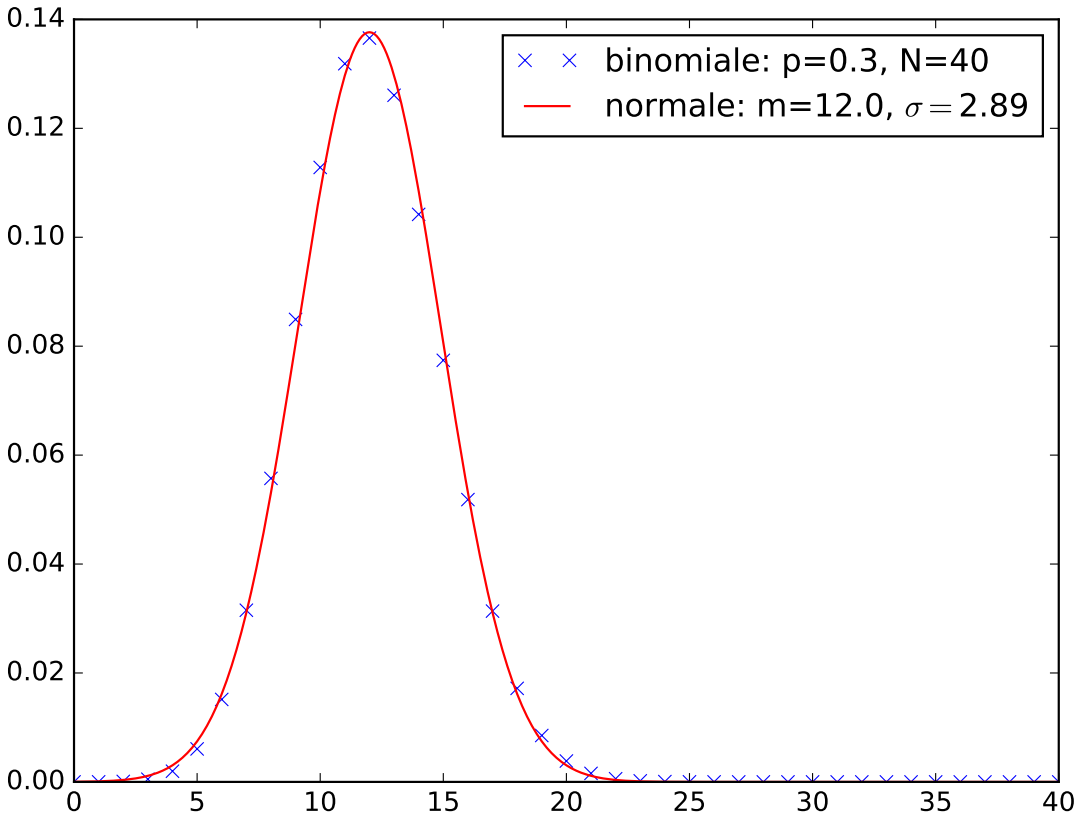

Figure – Comparaison Binomiale/Normale

Une série de graphiques

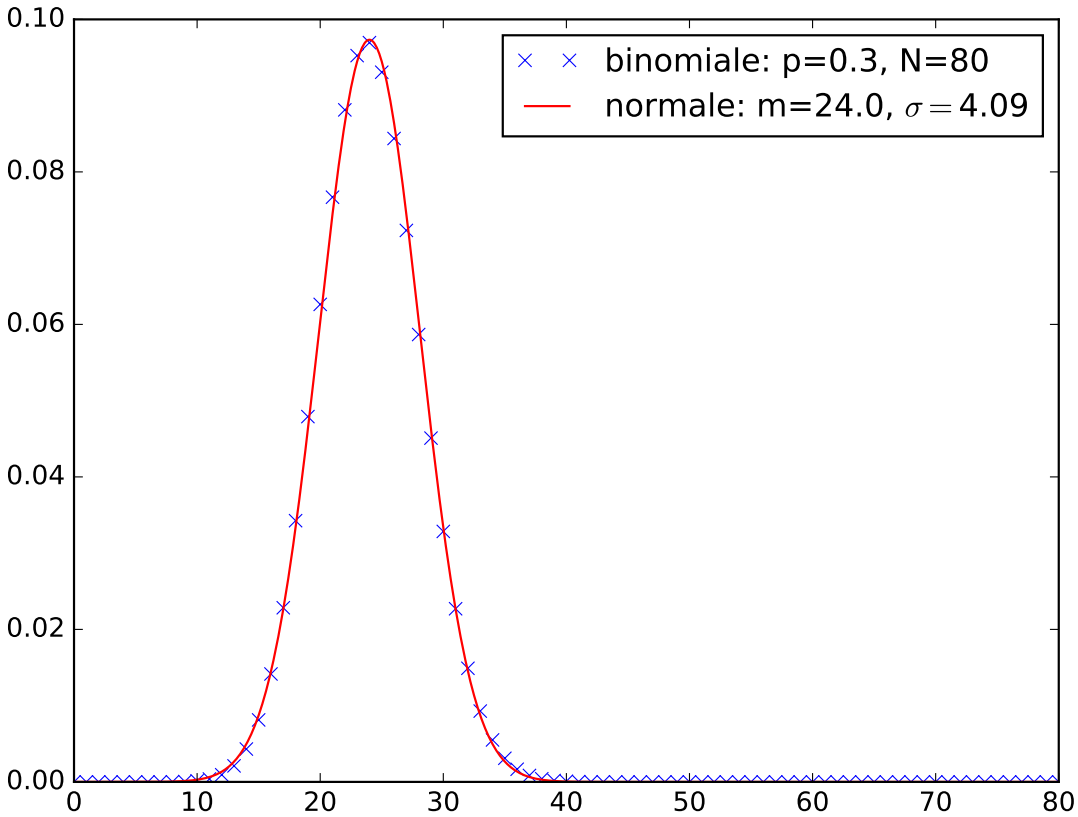

Figure – Comparaison Binomiale/Normale

Une série de graphiques

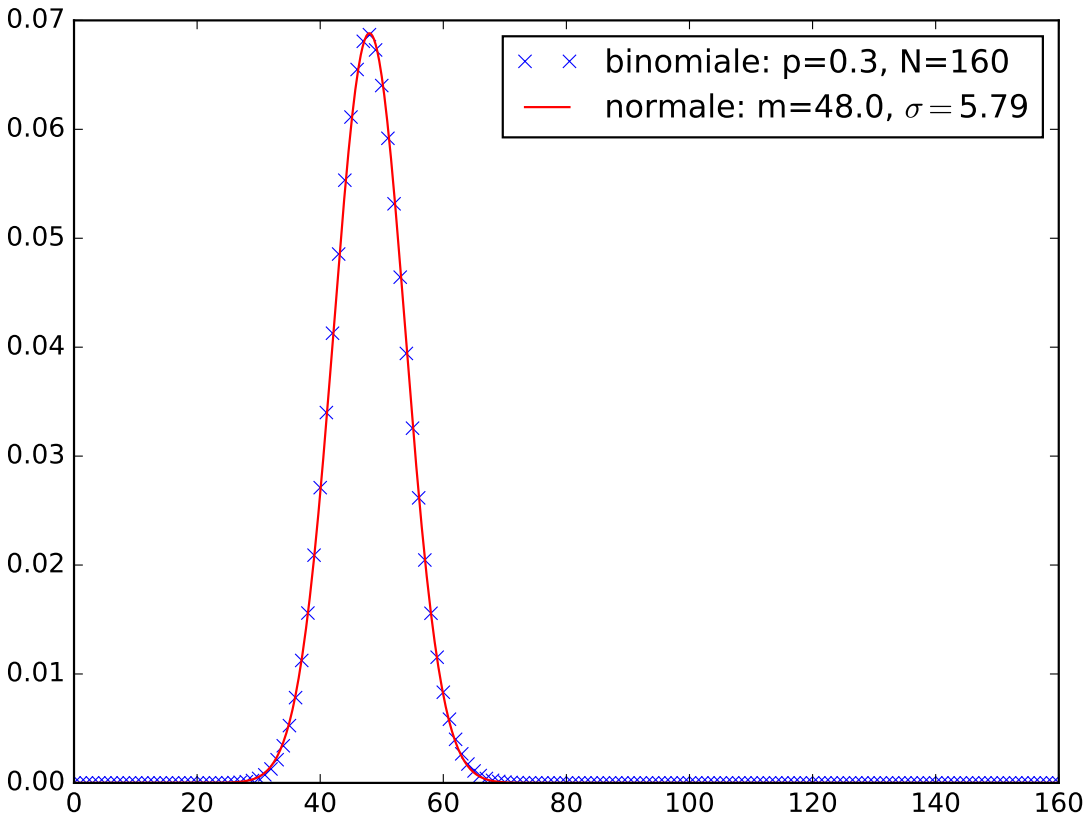

Figure – Comparaison Binomiale/Normale

Plaçons nous dans le cadre de la statistique d'échantillonnage. Supposons que X, la variable observée dans le série statistique, admette une espérance, $\mathbb{E}(X) = \mu$.
Supposons que $(X_n)_{n\in\mathbb{N}^*}$ soit un échantillon de X,
La N-ième moyenne empirique,

$$\overline{X}_N = M_N := \frac{1}{N}\sum_{n=1}^{N} X_n$$

est un estimateur non-biaisé de μ : $\mathbb{E}(M_N) = \mu$
La question qui se pose est la proximité des valeurs de M_N avec μ. (On rappelle qu'on calcule UNE valeur de M_N lorsque l'on fait la moyenne des données statistiques). On cherche à évaluer la *proportion* de valeurs possibles pour M_N qui s'écartent « un peu trop » de μ.

Théorème ((Admis) Loi faible des grands nombres/LFGN)

Soit X une v.a. réelle admettant une espérance μ. $(X_n)_{n\geq 1}$ un échantillon de X. Pour tout $\delta > 0$, lorsque $N \to +\infty$,

$$\mathbb{P}(|\overline{X}_N - \mu| > \delta) \to 0$$

Remarque : la signification de ceci est, que lorsque N est grand, $\overline{X}_N$ est approximativement distribuée comme la variable aléatoire constante μ. Plus précisémment, cet énoncé se reformule en termes de formule de transfert générique par

Proposition ((Hors programme))

Pour toute fonction h **continue et bornée** *sur* $\mathbb{R}$,

$$\mathbb{E}(h(\overline{X}_N)) \overset{N\to+\infty}{\to} \mathbb{E}(h(\mu)) = h(\mu)$$

Applications en simulation

Une application évidente de la LFGN est l'estimation d'une espérance d'une v.a X par simulation

- On écrit une fonction `X()` simulant la variable X avec comme convention que chaque appel à `X` est indépendant des autres
- On peut évaluer l'espérance de X en effectuant une moyenne des valeurs obtenues par `NS=1000` appels à `X()`

On a appliqué ce principe assez souvent.

Applications en simulation

Le même principe sert à l'estimation d'une probabilité d'un événement concernant X. Imaginons que nous voulions estimer $\mathbb{P}(X \leq \frac{1}{2})$.

- On applique la LFGN à $Y_{\frac{1}{2}} = \mathbb{1}_{\{X \leq \frac{1}{2}\}}$ pour évaluer, par simulation

$$\mathbb{E}(Y_{\frac{1}{2}}) = \mathbb{P}(X \leq \frac{1}{2})$$

- On effectue donc `NS` simulations de $Y_{\frac{1}{2}}$ dont on moyenne les valeurs
- Cela revient à effectuer `NS` appels à `X()` et à évaluer la proportion de valeurs retournées $\leq \frac{1}{2}$

Applications en simulation

Cela explique pourquoi, si l'on effectue `NS` appels à `X()` et que l'on trace la fonction de répartition des valeurs obtenues, on obtient une approximation de la véritable fonction de répartition de X. Il s'agit d'une application de la LFGN à $Y_x := \mathbb{1}_{\{X \leq x\}}$

Applications en simulation

Il en est de même pour les histogrammes. Là encore, on a utilisé ce principe dès les premières simulations de v.a.
Supposons maintenant que X soit une v.a. prenant les valeurs distinctes $x_1, \dots x_K$ avec probabilité $p_k = \mathbb{P}(X = x_k)$, simulée par la fonction `X()`.
Considérons, pour chaque $k \in \{1, \dots, K\}$, la v.a. $Y_k = \mathbb{1}_{\{X = x_k\}}$. Y_k est une v.a de Bernoulli de paramètre de succès p_k.

Applications en simulation

Si l'on effectue `NS` appels à `X()` et que l'on trace l'histogramme des valeurs obtenues, on place, au dessus de chaque x_k la proportion d'apparition de x_k dans la liste des valeurs. *i.e.* la valeur tirée au sort de

$$\overline{(Y_k)}_{\mathtt{NS}} = \frac{1}{\mathtt{NS}} \sum_{s=1}^{\mathtt{NS}} 1\!\!1_{\{X_s = x_k\}}$$

Par la LFGN, celle-ci est probablement proche de $\mathbb{E}(Y_k) = p_k$ et ce, d'autant plus que `NS` est grand.
Le graphe obtenu est donc probablement proche du graphe de $x_k \mapsto p_k$, *i.e.* l'histogramme théorique de la loi de X.

Le cas d'un échantillon Gaussien

Faisons comme hypothèse que $X \sim \mathcal{N}\left(m, \sigma^2\right)$ et rappelons le résultat suivant

Théorème

Si X et Y sont indépendantes, $X \sim \mathcal{N}\left(m_x, \sigma_x^2\right)$, $Y \sim \mathcal{N}\left(m_y, \sigma_y^2\right)$ et $Z = X + Y$. On a alors

$$Z \sim \mathcal{N}\left(m_z, \sigma_z^2\right)$$

où

$$m_z = m_x + m_y \text{ et } \sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

Le cas d'un échantillon Gaussien

On en déduit

Proposition

Soit $(X_1, \ldots, X_N, \ldots,)$ *un échantillon de* X*, on a alors, pour* $N \in \mathbb{N}^*$*,*

$$\overline{X}_N \sim \mathcal{N}\left(m, \sigma^2/N\right)$$

et donc, pour $\delta > 0$, par l'inégalité de Bienaymé–Tchebycheff

$$\mathbb{P}(|\overline{X}_N - m| > \delta) \leq \frac{\sigma^2}{N.\delta^2} \overset{N \to +\infty}{\to} 0$$

Le cas d'un échantillon Gaussien

On a donc démontré la loi faible des grands nombres dans le cas d'une échantillon Gaussien par un calcul effectif de la loi de la moyenne empirique. Ce qui est puissant dans la LFGN générale, c'est que celle-ci s'applique, quelle que soit la loi de X, pourvu que X soit intégrable. L'utilisation de l'inégalité de Bienaymé–Tchebycheff pour ce cas complètement explicite est un peu grossière et ne donne pas une bonne estimation de la vitesse de convergence vers 0.

Preuve d'un cas particulier simple et typique

On va démontrer la loi faible des grands nombres dans le cas où X est de carré intégrable.

Démonstration.

Soit $N \in \mathbb{N}^*$. On a

$$\overline{X}_N - \mu = \overline{X}_N - \mathbb{E}(\overline{X}_N) = \frac{1}{n}\sum_{n=1}^{N}(X_n - \mathbb{E}(X_n))$$

et donc, par indépendance deux à deux des X_k,

$$\mathbb{V}(\overline{X}_N) = \frac{N}{N^2}\mathbb{V}(X) = \frac{\mathbb{V}(X)}{N}$$

Preuve d'un cas particulier simple et typique

On va démontrer la loi faible des grands nombres dans le cas où X est de carré intégrable.

Démonstration.

En appliquant l'inégalité de Bienaymé–Tchebycheff, on obtient alors, pour $\delta > 0$ fixé, lorsque $N \to +\infty$, que

$$0 \leq \mathbb{P}(|\overline{X}_N - \mu| > \delta) \leq \frac{\mathbb{V}(\overline{X}_N)}{\delta^2} \leq \frac{\mathbb{V}(X)}{N\delta^2} \to 0$$

Exercice 6.— Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes suivant toutes la même loi de Bernoulli de paramètre $p \in]0,1[$. Pour tout $n \geq 1$, on note

$$Y_n = X_n.X_{n+1}$$

1. Pour $n \geq 1$, donner la loi de Y_n, $\mathbb{E}(Y_n)$, $\mathbb{V}(Y_n)$.
2. Pour tout $(i,j) \in \mathbb{N}^*$, calculer $\mathbb{C}ov(Y_i, Y_j)$.
3. Soit $S_n = \frac{Y_1+\cdots+Y_n}{n}$ pour $n \geq 1$. Calculer $\mathbb{E}(S_n)$ et $\mathbb{V}(S_n)$. Montrer que pour tout $\varepsilon > 0$,

$$\mathbb{P}(|S_n - p^2| > \varepsilon) \overset{n\to+\infty}{\to} 0$$

Exercice 7.— Soit $n \in \mathbb{N}^*$, $p = \frac{1}{6} \in]0,1[$. On lance n fois un dé équilibré et on considère S_n le nombre de fois où 6 est sorti durant les n premiers tirages. On pose $Y_n = \exp(\frac{S_n}{n})$.

1. Calculer l'espérance et la variance de Y_n.

2. Prouver que lorsque $n \to +\infty$, $\mathbb{E}(Y_n) \to e^p$ et $\mathbb{V}(Y_n) \to 0$. En déduire que pour tout $\varepsilon > 0$, lorsque $n \to +\infty$,

$$\mathbb{P}(|Y_n - e^p| < \varepsilon) \to 1$$

Une autre quantité communément calculée sur un échantillon statistique $(X_n)_{n\geq 1}$ est la *variance empirique*

$$S_N^2 := \frac{1}{N}\left(\sum_{n=1}^{N}(X_n - \overline{X}_N)^2\right)$$

Cette quantité se présente elle aussi sous forme d'une moyenne. Supposons X de carré intégrable, d'espérance μ, de variance σ^2. On a

$$\mathbb{E}(S_N^2) = \frac{N-1}{N}\sigma^2$$

Une loi faible des grands nombres pour S_N^2.

Proposition (Admise)

Si X est de carré intégrable, alors pour tout $\delta > 0$, lorsque $N \to +\infty$,

$$\mathbb{P}(|S_N^2 - \sigma^2| > \delta) \to 0$$

A remarquer :

1. S_N^2 n'est pas écrite a priori comme somme de v.a. indépendantes. La LFGN ne s'applique pas directement.
2. Le cas où X^4 est intégrable peut se règler avec Bienaymé–Tchebycheff en calculant la variance de S_N^2, *c.f.* exercice 8.

Une distribution stable

Comme on l'a rappelé lors de la preuve de la loi faible des grands nombres dans le cas d'un échantillon Gaussien :
Si $(X_n)_{n\geq 1}$ sont indépendantes, *normales*, d'espérance μ et de variance $\sigma^2 > 0$ alors, pour $N \in \mathbb{N}^*$, la moyenne empirique, centrée et réduite

$$M_N^* = \frac{M_N - \mu}{\frac{\sigma}{\sqrt{N}}} = \sqrt{N}\frac{\overline{X}_N - \mu}{\sigma}$$

est $\mathcal{N}(0,1)$, *i.e.* normale, centrée et réduite. La distribution de M_N^* est *indépendante* de N.

Soit X une v.a.r de carré intégrable, d'espérance μ, de variance σ^2. Si $(X_n)_{n\geq 1}$ est échantillon de la variable X, on définit, pour $N \in \mathbb{N}^*$, sa N-ième moyenne empirique centrée et réduite par

$$M_N^* = \frac{M_N - \mu}{\frac{\sigma}{\sqrt{N}}}$$

C'est une variable centrée et réduite. En général, calculer la loi de M_N^* en fonction de la loi de X est mission impossible. Cependant...

Théorème ((Admis) de la limite centrale/TCL, Paul Levy, 1935)

Soit X une v.a.r de carré intégrable, d'espérance μ, de variance σ^2. Si $(X_n)_{n\geq 1}$ est échantillon de la variable X alors, pour tous $a, b \in \mathbb{R}$, $a < b$, lorsque $N \to +\infty$,

$$\mathbb{P}(a \leq M_N^* \leq b) \to \int_a^b \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, dx$$

Exercice 7.-⋆- Soit $(X_k)_{k\in\mathbb{N}^*}$ une suite de variables aléatoires mutuellement indépendantes suivant chacune la loi uniforme sur $[0,1]$.
On pose, pour $n\in\mathbb{N}^*$, $S_n=\sum_{k=1}^n X_k$.
Quelle est la limite (si elle existe) quand $n\to+\infty$ de $\mathbb{P}(n/2-\sqrt{n}<S_n<n/2+\sqrt{n})$?

Exercice 11.— Soit X une v.a réelle de carré intégrable, d'espérance nulle, de variance $\sigma^2 > 0$. On suppose que si X_1 et X_2 sont indépendantes, de même loi que X alors $\frac{X_1+X_2}{\sqrt{2}}$ est distribuée comme X.
Montrer en utilisant le TCL que X suit une loi normale $\mathcal{N}\left(0, \sigma^2\right)$.

Théorème

Soit X une v.a.r de carré intégrable, d'espérance μ. Si $(X_n)_{n\geq 1}$ est un échantillon de X, pour tous $a, b \in \mathbb{R}$, $a < b$, lorsque $N \to +\infty$,

$$\mathbb{P}(a \leq \frac{M_N - \mu}{\frac{S_N}{\sqrt{N}}} \leq b) \to \int_a^b \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, dx$$

où

$$S_N = \left(\frac{1}{N} \sum_{n=1}^{N} (X_n - \overline{X}_N)^2 \right)^{\frac{1}{2}}$$

est l'écart-type empirique.

Fonction de répartition de la Gaussienne normalisée

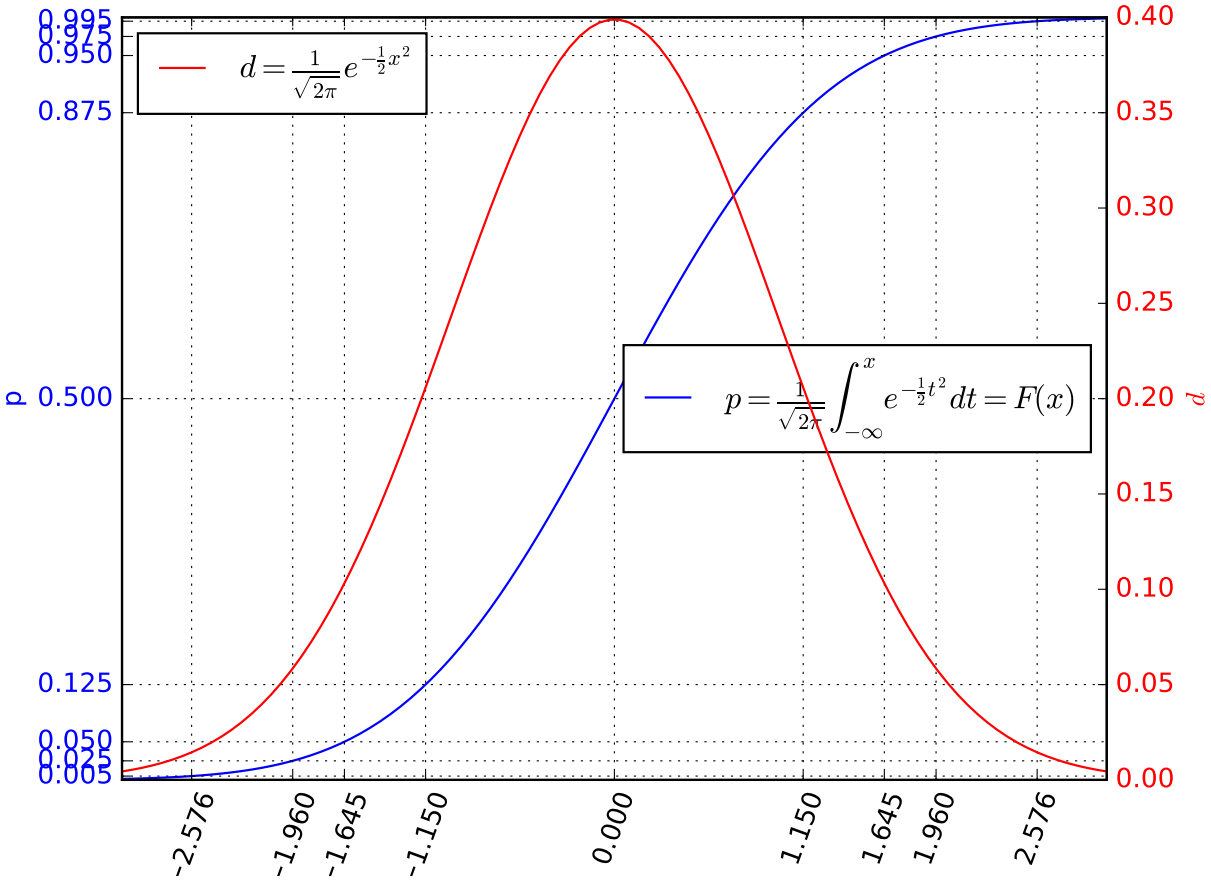

Figure – La densité normale $\mathcal{N}(0,1)$ sa fonction de répartition

Le théorème 10 a pour conséquence (on conserve les notations)

Théorème

Soit X une v.a.r de carré intégrable, d'espérance μ. Si $(X_n)_{n\geq 1}$ est un échantillon de X, pour $0<\alpha<1$, lorsque $N\to+\infty$,

$$\mathbb{P}\left(\sqrt{N}.\left|\frac{M_N-\mu}{S_N}\right|<u_{1-\frac{\alpha}{2}}\right)\to 1-\alpha$$

et

$$\mathbb{P}\left(\sqrt{N}.\left|\frac{M_N-\mu}{S_N}\right|>u_{1-\frac{\alpha}{2}}\right)\to \alpha$$

On a déjà abordé la question du test de conformité d'une moyenne, qui est basée sur l'utilisation du théorème 11.
Voici la recette complète :

1. On dispose de valeurs observées $(x_1, \ldots, x_N)$ d'un échantillon $(X_1, \ldots, X_N)$ d'une v.a.r X supposée de carré intégrable, d'espérance μ inconnue.
2. On en calcule la moyenne $\overline{x}$, valeur observée de la variable moyenne empirique $\overline{X}_N = M_N$ et l'écart type σ_x, valeur observée de la variable S_N, racine carrée de la variable moyenne empirique.
3. Pour μ_0 fixé. On fait l'hypothèse nulle $(H_0) : \mu = \mu_0$.
4. Sous l'hypothèse nulle, on a

$$\mathbb{P}\left(\sqrt{N}.\left|\frac{M_N - \mu}{S_N}\right| < u_{1-\frac{\alpha}{2}}\right) \simeq 1 - \alpha$$

En conséquence, au niveau de confiance $1-\alpha$:

1. Si $\left|\sqrt{N}.\frac{\overline{x}-\mu}{\sigma_x}\right| > u_{1-\frac{\alpha}{2}}$: on rejette (H_0) au niveau de confiance $1-\alpha$: notre population n'est pas conforme à cette hypothèse à ce niveau de confiance.
2. Si $\left|\sqrt{N}.\frac{\overline{x}-\mu}{\sigma_x}\right| < u_{1-\frac{\alpha}{2}}$: on retient (H_0) au niveau de confiance $1-\alpha$: notre population est conforme à cette hypothèse à ce niveau de confiance.

On peut retourner la problématique précédente sous la forme suivante
Etant donnés, un ensemble de valeurs observées $(x_1, \ldots, x_N)$ d'un échantillon $(X_1, \ldots, X_N)$ d'une v.a.r X supposée de carré intégrable, d'espérance μ inconnue, dont on calculé moyenne $\overline{x}$ et écart type σ_x, *pourvu que N soit suffisamment grand*, quelles sont les valeurs de μ_0 pour lesquelles on retient l'hypothèse (H_0) au niveau de confiance $1 - \alpha$?

Un moment de réflexion montre que ceci équivaut à

$$\mu_0 \in I_\alpha = \left[\overline{x} - u_{1-\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{N}}, \overline{x} + u_{1-\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{N}}\right]$$

Cet intervalle s'appelle l'intervalle de confiance pour μ au niveau de confiance $1-\alpha$.
Autrement dit, à ce niveau de confiance,

1. pour tout $\mu_0 \in I_\alpha$, notre population est conforme à l'hypothèse $(H_0) : \mu = \mu_0$.
2. pour tout $\mu_0 \notin I_\alpha$, notre population n'est pas conforme à l'hypothèse $(H_0) : \mu = \mu_0$.

Exercice 15.— On effectue des pesées avec une balance. On sait, pour l'avoir testée, que cette balance donne, pour un objet donné, des résultats qui suivent une loi normale dont la moyenne est la masse de l'objet pesé, et dont l'écart-type est de $\sigma = 1g$.

1. On a effectué 25 mesures d'un certain objet, et la somme des résultats est $30,25g$. Donner un intervalle de confiance à 95% pour la masse de cet objet.

2. Reprendre ce qui précède pour 400 mesures dont la somme donne le résultat $484g$.

3. Peut-on déterminer un nombre de mesures minimum n tel que la l'estimation de la masse soit à $5.10^{-2}.g$ près à un niveau de confiance de 95% ?