Notes de cours 05

Théorèmes « limite » en probabilités. Statistiques inférentielles

Table des matières

1 Statistique inférentielle : Kezako ?

Nous avons déjà vu (*cf.* TP 1 d'info) comment, étant donnée une liste finie de nombres $(x_n)_{n\in\{1,\dots,\mathscr{N}\}}$, issus de mesures sur une population d'individus [1], on en calcule moyenne $\bar{x}$, écart-type σ_x et variance, médiane et même distribution et histogramme.

Etant donnée une liste finie de couples numériques (par exemple lue d'un fichier de données expérimentales, *cf.* TP 2 d'info), $(x_n,y_n)_{n\in\{1,\dots,\mathscr{N}\}}$, 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 ».

1.1 Statistique exhaustive

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.

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

Le tableau de données sur le poids de cette population est donné par la suite finie $x = (x_n)_{n\in\{1,\dots,\mathscr{N}\}}$ où x_n est le poids à sa mort du dodo D_n. L'interprétation de moyenne $\bar{x}$, 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 $\{1,\dots,\mathscr{N}\}$: celui-ci porte le numéro N. $N\sim\mathscr{U}_{\{1,\dots,\mathscr{N}\}}$.
2. J'appelle X la variable aléatoire réelle, fonction de N définie par $X = x_N$.
3. On a alors $\mathbb{E}(X) = \bar{x}$, $\sqrt{\mathbb{V}(X)} = \sigma_x$ 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 ».

1. individus est à prendre en un sens très large. Il peut s'agir d'une population d'animaux, de plantes, de cailloux ou d'atomes, mais aussi, plus abstraitement, d'expériences, de pays, etc... L'important est que ces individus partagent une « nature commune » et qu'étant donné un individu, caractérisé par son numéro, on puisse mesurer une quantité/qualité X : la valeur de X pour l'individu n est x_n.

1.2 Statistique d'échantillonnage

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 [2] 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 [3] 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, $(x_1,\dots,x_N)$ est la suite des valeurs prises sur un tirage au sort d'une suite de v.a $(X_1,\dots,X_N)$ où $X_1,\dots,X_N$ 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 $(x_1,\dots,x_N)$ 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 1. *La moyenne calculée $\overline{x}$ est la valeur sur un tirage au sort de la variable aléatoire «* moyenne empirique *» de $X_1,\dots,X_N$, notée (deux notations) et définie par*

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

Définition 2. *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 $X_1,\dots,X_N$, notée (deux notations) et définie par*

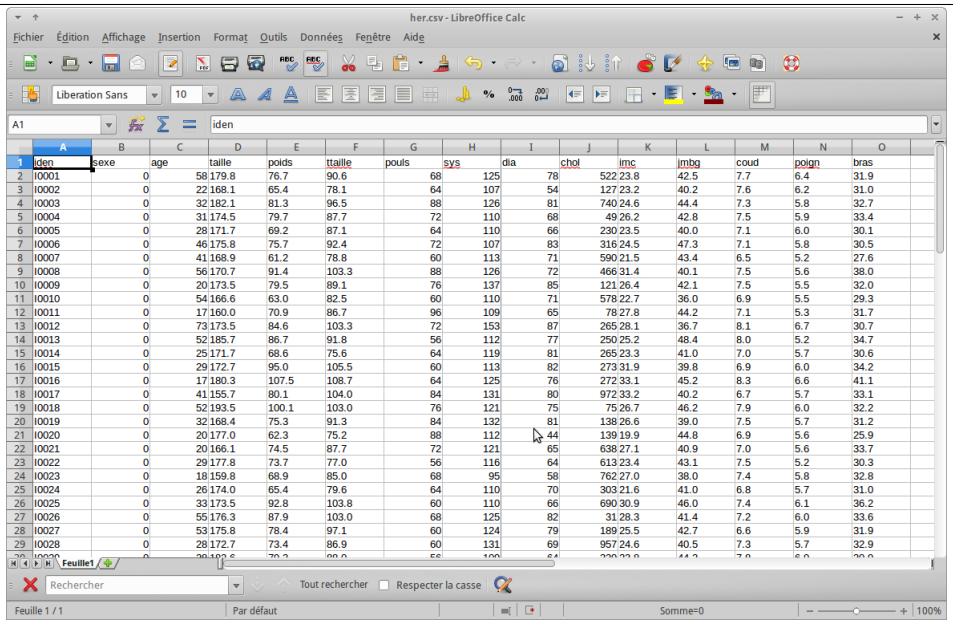
$$\Sigma_N^2 = S_N^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - M_N)^2 = \left(\frac{1}{N}\sum_{n=1}^{N} X_n^2\right) - M_N^2 = \overline{X^2}_N - \overline{X}_N^2$$

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.

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.

2. On désigne le caractère et la v.a. d'une même lettre malgré une différence de nature entre les deux.
3. à partir de maintenant N remplace $\mathscr{N}$



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	iden	sexe	age	taille	poids	ttaille	pouls	sys	dia	chol	imc	imbg	coud	poign	bras
2	I0001	0	58	179.8	76.7	90.6	68	125	78	522	23.8	42.5	7.7	6.4	31.9
3	I0002	0	22	168.1	65.4	78.1	64	107	54	127	23.2	40.2	7.6	6.2	31.0
4	I0003	0	32	182.1	81.3	96.5	88	126	81	740	24.6	44.4	7.3	5.8	32.7
5	I0004	0	31	174.5	79.7	87.7	72	110	68	49	26.2	42.8	7.5	5.9	33.4
6	I0005	0	28	171.7	69.2	87.1	64	110	66	230	23.5	40.0	7.1	6.0	30.1
7	I0006	0	46	175.8	75.7	92.4	72	107	83	316	24.5	47.3	7.1	5.8	30.5
8	I0007	0	41	168.9	61.2	78.8	60	113	71	590	21.5	43.4	6.5	5.2	27.6
9	I0008	0	56	170.7	91.4	103.3	88	126	72	466	31.4	40.1	7.5	5.6	38.0
10	I0009	0	20	173.5	79.5	89.1	76	137	85	121	26.4	42.1	7.5	5.5	32.0
11	I0010	0	54	166.6	63.0	82.5	60	110	71	578	22.7	36.0	6.9	5.5	29.3
12	I0011	0	17	160.0	70.9	86.7	96	109	65	78	27.8	44.2	7.1	5.3	31.7
13	I0012	0	73	173.5	84.6	103.3	72	153	87	265	28.1	36.7	8.1	6.7	30.7
14	I0013	0	52	185.7	86.7	91.8	56	112	77	250	25.2	48.4	8.0	5.2	34.7
15	I0014	0	25	171.7	68.6	75.6	64	119	81	265	23.3	41.0	7.0	5.7	30.6
16	I0015	0	29	172.7	95.0	105.5	60	113	82	273	31.9	39.8	6.9	6.0	34.2
17	I0016	0	17	180.3	107.5	108.7	64	125	76	272	33.1	45.2	8.3	6.6	41.1
18	I0017	0	41	155.7	80.1	104.0	84	131	80	972	33.2	40.2	6.7	5.7	33.1
19	I0018	0	52	193.5	100.1	103.0	76	121	75	75	26.7	46.2	7.9	6.0	32.2
20	I0019	0	32	168.4	75.3	91.3	84	132	81	138	26.6	39.0	7.5	5.7	31.2
21	I0020	0	20	177.0	62.3	75.2	88	112	44	139	19.9	44.8	6.9	5.6	25.9
22	I0021	0	20	166.1	74.5	87.7	72	121	65	638	27.1	40.9	7.0	5.6	33.7
23	I0022	0	29	177.8	73.7	77.0	56	116	64	613	23.4	43.1	7.5	5.2	30.3
24	I0023	0	18	159.8	68.9	85.0	68	95	58	762	27.0	38.0	7.4	5.8	32.8
25	I0024	0	26	174.0	65.4	79.6	64	110	70	303	21.6	41.0	6.8	5.7	31.0
26	I0025	0	33	173.5	92.8	103.8	60	110	66	690	30.9	46.0	7.4	6.1	36.2
27	I0026	0	55	176.3	87.9	103.0	68	125	82	31	28.3	41.4	7.2	6.0	33.6
28	I0027	0	53	175.8	78.4	97.1	60	124	79	189	25.5	42.7	6.6	5.9	31.9
29	I0028	0	28	172.7	73.4	86.9	60	131	69	957	24.6	40.5	7.3	5.7	32.9

FIGURE 1 – Les données de fichier `her.csv` : chaque individu est sur une ligne. Une colonne donne les valeurs d'un caractère.

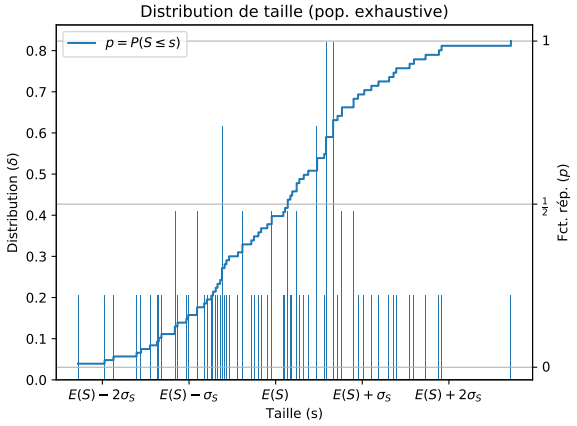


FIGURE 2 – Une représentation de la distribution exhaustive des tailles.

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/réalisation 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 :

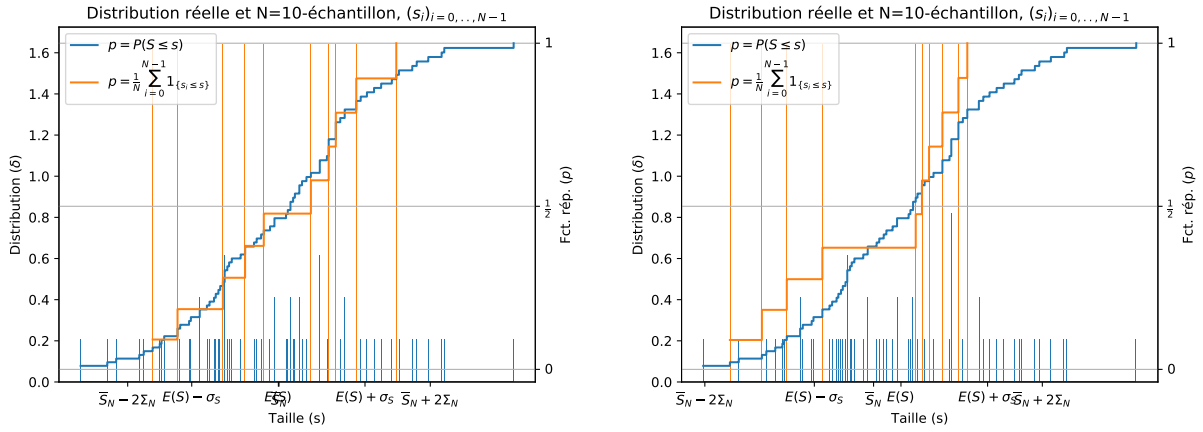


FIGURE 3 – Comparaison avec deux réalisations d'un 10-échantillons.

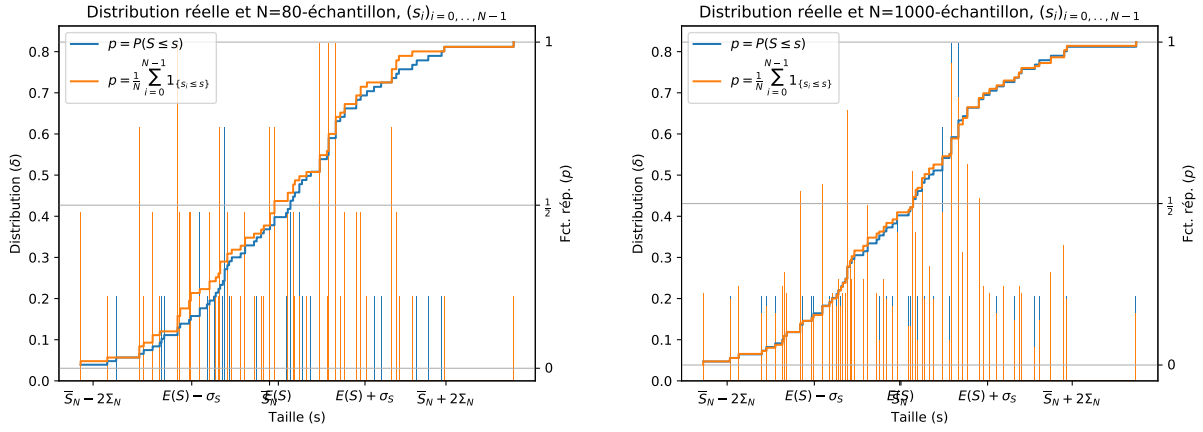


FIGURE 4 – Comparaisons avec des réalisations de 80-échantillon et 1000-échantillon.

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[4] 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 ? »

4. Par ceci, on entend que l'information obtenue par le calcul de M_N sur notre série de données donne des « renseignements » sur $\mathbb{E}(X)$

Il y a évidemment de nombreuses questions plus avancées peut-être déja rencontrées par certains d'entre vous en TIPE. Pour chacune d'entre elles, de nombreuses possibilités d'estimations et de tests statistiques existent.

— Et si on fait une médiane empirique ?

— Et carrément : qu'en est-il de la distribution empirique (*i.e.* l'histogramme des valeurs observées) et de la distribution de X ?

— Si $x_1,\dots,x_M$ et $y_1,\dots,y_N$ sont deux séries de valeurs d'échantillons (indépendants) de variables X et Y, peut-on, par un test, conclure que X et Y n'ont pas même distribution ? Est-ce que le fait que les barres d'erreur dans Exxxel soient disjointes est un bon critère ?

— Si $(x_1,y_1),\dots,(x_N,y_N)$ est une série de valeurs d'un échantillon (bivarié, donc) d'une v.a. (X,Y), Peut-on, par un test, conclure que X et Y ne sont pas indépendantes, ne sont pas corrélées ? A quel point la covariance empirique doit-elle être non nulle pour conclure à la non-corrélation de X et Y ?

— ....

1.3 Inférence et hypothèse nulle H_0

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 $(x_1,\dots,x_N)$ d'un N-échantillon $(X_1,\dots,X_N)$ 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 $(X_1,\dots,X_n)$, un intervalle [5] $I\subset\mathbb{R}$, tels que [6] *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 $x_1,\dots,x_N$ de la même manière que T est obtenue à partir de $X_1,\dots,X_N$. Deux possibilités existent

1. Soit $t\notin I$: on rejette [7] l'hypothèse H_0
2. Soit $t\in I$: on « accepte [8] » 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...

5. ou une partie

6. Afin de ne pas multiplier les lettres grecques, on suppose que l'on fait de la stat. inf. au *niveau de confiance* $1-\alpha=95\%$, on laisse le lecteur modifier le texte pour gérer le niveau de confiance 99%, voir un niveau de confiance générique $1-\alpha$

7. La signification de ceci doit être claire : On sait que pour au moins 95% des tirages au sort de N-échantillons, T est dans l'intervalle I. Si $t\notin I$ et H_0 est vraie, cela signifie que notre tirage au sort se situe dans les 5% de tirages exceptionnels. Comme ceci a peu de chances d'arriver–on ne croise pas un mouton à 5 pattes tous les jours–, on préfère rejeter H_0. Le risque de *première espèce* est la probabilité de rejeter H_0 alors que celle-ci est vraie. Il est ici de $\alpha=5\%$

8. Plutôt, on « non rejette », ce n'est pas parce qu'on a passé un test qu'on est OK, avant d'accepter raisonnablement une hypothèse, il faudrait lui faire passer une batterie de tests numériques indépendants. Le *risque de deuxième espèce* est la probabilité β d'accepter H_0 alors que celle-ci est fausse. La *puissance* du test est la probabilité d'accepter H_0 à raison, ceci vaut $1-\beta$

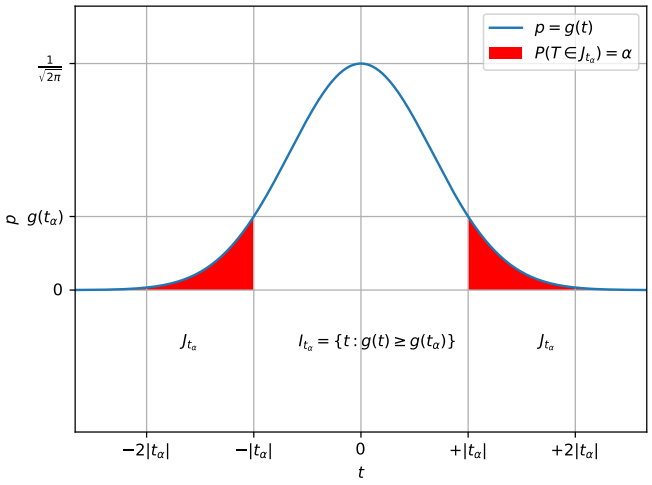


FIGURE 5 – Les intervalles I_{t_α} et J_{t_α} dans le cas Gaussien.

1.4 Construction de l'intervalle I.

La construction de l'intervalle I peut parfois sembler arbitraire. Une méthode est basée sur la notion (hors programme) de p-valeur. Supposons que, sous l'hypothèse H_0, T soit une v.a. réelle à densité continue δ_T. A la valeur t de T calculée sur l'échantillon, on peut associer les parties J_t et I_t de $\mathbb{R}$ définies par

$$J_t = \{\tau \in \mathbb{R},\, \delta_T(\tau) \leq \delta_T(t)\} \text{ et } I_t = \{\tau \in \mathbb{R},\, \delta_T(\tau) > \delta_T(t)\} = \mathbb{R} \setminus J_t$$

En un sens, J_t est l'ensemble des valeurs possibles de T *moins susceptibles* d'être tirées au sort que t. La p-valeur de t est

$$p_t = \mathbb{P}(T \in J_t)$$

— Si p_t est inférieur à $\alpha = 5\%$, cela signifie, sous l'hypothèse H_0, que la probabilité d'observer une valeur de T moins susceptible d'apparaitre que t est inférieure $\alpha = 5\%$: c'est improbable et on *rejette* l'hypothèse H_0 au niveau de confiance $1-\alpha = 95\%$;

— Si p_t est supérieur à $\alpha = 5\%$, cela signifie que t fait partie des $1-\alpha = 95\%$ des valeurs de T les plus susceptibles d'être observées : on *accepte* l'hypothèse H_0 au niveau de confiance $1-\alpha = 95\%$;

En retournant logiquement ce raisonnement, si on construit $t_\alpha \in \mathbb{R}$ de sorte que

$$\mathbb{P}(T \in I_{t_\alpha}) = 1-\alpha \text{ et } \mathbb{P}(T \in J_{t_\alpha}) = \alpha$$

On aura alors l'alternative :

— Si $t \notin I_{t_\alpha}$ alors la p-valeur p_t est $\leq \alpha$: on rejette l'hypothèse H_0 au seuil de confiance $1-\alpha$;

— Si $t \in I_{t_\alpha}$ alors la p-valeur p_t est $\geq \alpha$: on accepte l'hypothèse H_0 au seuil de confiance $1-\alpha$;

Dans le cas (le seul au coeur de notre programme) où T suit une loi normale $\mathcal{N}(0,1)$, les ensembles I_t sont des intervalles centrés en 0 et pour $\alpha = 5\%$, d'après la table de valeurs de la fonction de répartition Gaussienne,

$$I_{t_\alpha} = [-1.96, 1.96]$$

1.5 Test de conformité sur la moyenne

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,

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

1. Quelle quantité [9] T, quel intervalle I prendre pour réaliser le test de (H_0) avec un niveau de confiance de 95% ?
2. Avec nos données : rejette-t-on ou accepte-t-on (H_0) ? Analyser la partie du script `TestH0.py` consacrée à cette question.

Exercice 1.—On suppose qu'une naissance (d'être humain) est une expérience aléatoire dont la caractéristique X première est le fait d'obtenir un garçon ou une fille. On obtient une fille (succès !) avec probabilité p et un garçon (échec !) avec probabilité $q = 1 - p$.

$$\mathbb{P}(X = 1) = p \text{ et } \mathbb{P}(X = 0) = q = 1 - p$$

Le bon sens (et une vue un peu rapide de la biologie sous-jacente) nous dit que, vu que le sexe est déterminé par le chromosome sexuel X ou Y porté par le spermatozoïde fécondant l'ovule, vu le mode de production des spermatozoïdes par meïose produisant une quantité égale de spermatozoïdes X et de spermatozoïdes Y, on a $p = \frac{1}{2}$.

Est-ce correct ? On suppose que l'ensemble [10] des $N = 746977$ naissances, dont de 364 392 femmes et 382 585 hommes, en france en 2016 constitue une réalisation d'un échantillon $X_1, \ldots, X_N$ de X.

En considérant que $p = \mathbb{E}(X)$ et que $T = \frac{\sqrt{N}}{\sqrt{p(1-p)}}(M_N - p)$ suit de très près une loi $\mathcal{N}(0,1)$, l'hypothèse $(H_0) : p = \frac{1}{2}$ est-elle conforme au niveau de confiance 95% avec l'observation ?

2 Paramètres de localisation et de dispersion

Les questions de limite de moyennes empiriques, de variances empiriques, lorsque le nombre d'individus N croit vers $+\infty$, posent le problème de l'approximation d'une constante par une suite de variables aléatoires.

Il faut commencer par s'équiper d'outils qui vont nous permette de mesurer les écarts entre v.a. et constantes, ou entre distributions de v.a. L'inégalité fondamentale est l'inégalité [11] de MARKOV.

9. On prend $T = \frac{M_N - 170}{\frac{S_N}{\sqrt{N}}}$ et $I = [-1.96, +1.96]$
10. Source :INSEE
11. Penser aussi aux variantes du type BIENAYMÉ–TCHEBYCHEFF

2.1 Rappel : Inégalité de MARKOV

Théorème 3 (Inégalité de MARKOV)**.** *Si X est une v.a.* positive*, on a, pour tout $\lambda > 0$,*

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

Démonstration. On a $1\!\!1_{\{X>\lambda\}} \leq \frac{X}{\lambda}$, en utilisant la croissance et la linéarité de l'espérance, on a donc

$$\mathbb{P}(X > \lambda) = \mathbb{E}(1\!\!1_{\{X>\lambda\}}) \leq \mathbb{E}\left(\frac{X}{\lambda}\right) \leq \frac{1}{\lambda}\mathbb{E}(X)$$

□

2.2 Espérance et variance

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. $X = Y$. Ensuite, observons un exemple élémentaire : prenons une v.a U unif. distribuée sur deux valeurs 0 et 1, $U \sim \mathscr{B}(\frac{1}{2})$, $f, g : \{0,1\} \to \mathbb{R}$ et $X = f(U)$ et $Y = g(U)$, on a alors

$$\mathbb{E}((X-Y)^2) = \frac{1}{2}\left((f(0)-g(0))^2 + (f(1)-g(1))^2\right)$$

On tombe sur le carré de la distance Euclidienne entre les vecteurs $(f(0), f(1))$ et $(g(0), g(1))$.

Cette quantité est nulle si et seulement $f = g$, *i.e.* $X = Y$ et un peu d'imagination géométrique devrait montrer que $\mathbb{E}((X-Y)^2)$ mesure un certain écart entre f et g, entre X et Y. Plus $\mathbb{E}((X-Y)^2)$ est grand, plus X et Y sont éloignées.

Variance, dispersion Le cas le plus simple est celui où Y est une constante. Cela soulève les questions suivantes :

1. Si X est une v.a de carré intégrable, quelle est la constante la plus proche de X au sens de l'écart quadratique moyen ?
2. Quelle est la valeur minimale de cet écart quadratique moyen à une constante ?

Proposition 4 (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 5 (Inégalité de BIENAYMÉ–TCHEBYCHEFF(rappel))**.** *Soit X une v.a réelle de carré intégrable, $\lambda > 0$, on a*

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

Dans le langage « courant », on dit souvent que 95% des individus sont à moins de deux écart-type de la moyenne.

Rappelons que l'écart type σ est la racine carrée de la variance. En appliquant l'inégalité de BIENAYMÉ–TCHEBYCHEFF à $\lambda = 2\sigma$, on obtient que

$$\mathbb{P}(|X - \mathbb{E}(X)| > 2\sigma) \leq \frac{1}{4} \textit{ i.e. } \mathbb{P}(|X - \mathbb{E}(X)| \leq 2\sigma) \geq \frac{3}{4}$$

Dans une statistique exhaustive, (les dodos !) on vient donc d'obtenir que, *quelle que soit la loi* de la variable observée, 75% des individus sont à moins de deux écart-type de la moyenne.

Le 95% de la pratique courante provient de l'estimation pour une loi normale dans une statistique d'échantillonnage. Nous reviendrons amplement là-dessus.

2.3 Médiane : (HP)

Une autre façon de mesurer l'écart entre deux v.a réelles (intégrables seulement) X et Y est de considérer l'*écart moyen* $\mathbb{E}(|X-Y|)$. L'analogue de la moyenne pour cette mesure d'écart est la *médiane*.

Proposition 6 ((HP))**.** *Soit X une v.a. réelle admettant ue espérance. La quantité $\mathbb{E}(|X-a|)$ est minimale si et seulement si a est une* médiane *de X.*

On ne détaille pas la définition ni les calculs, assez subtils et hors de notre programme.

3 Proximité en loi : des exemples

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.

3.1 Hypergéométrique et binomiale (HP)

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.

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

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

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

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

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 $\nu \{0,\dots,n\}$,

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

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

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

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 [12] 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 $\mathbb{E}(X_n)$.

12. On peut aussi faire l'exercice avec remise pour comparer

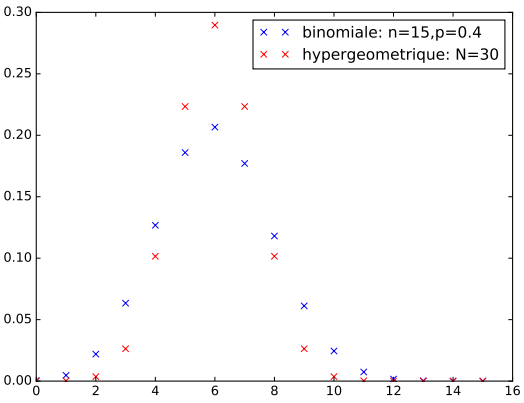


FIGURE 6 – Comparaison Hypergéométrique/Binomiale

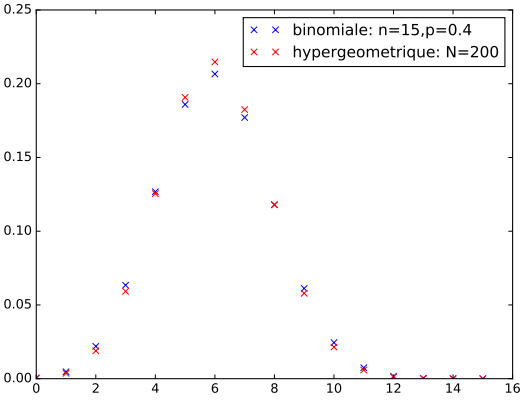


FIGURE 7 – Comparaison Hypergéométrique/Binomiale

3.2 Binomiale et Poisson (fin de l'année)

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 7. *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. Soit $k \in \mathbb{N}$. On a, pour $n \geq k$, puis lorsque $n \to +\infty$,

$$\begin{aligned}
\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \underbrace{(1-p_n)^{-k}}_{\to 1} . \underbrace{\frac{n}{n}\frac{n-1}{n}..\frac{n-k+1}{n}}_{\to 1} . \\
&\quad \underbrace{(1-\frac{p_n.n}{n})^n}_{\to e^{-\lambda}} \frac{1}{k!} \underbrace{(p_n.n)^k}_{\to \lambda^k} \\
&\to e^{-\lambda} \frac{\lambda^k}{k!}
\end{aligned}$$

□

Approximation discrète du processus de POISSON

On considère une suite $(X_k)_{k\geq 1}$ de v.a. de BERNOULLI de paramètre $p = \lambda/n$ que l'on place en chaque point $\frac{k}{n}$ de l'intervalle $]0,+\infty]$ et le nombre N de succès dans l'intervalle $]0,1]$.

Si S_m est la variable donnant l'emplacement du m-ième succès, alors

1. $\{N = m\} = \{S_m \leq 1\} \cap \{S_{m+1} > 1\}$ et $N \sim \mathscr{B}(n,p)$,
2. S_m est distribuée comme
$$\tilde{S}_m := \sum_{\ell=1}^{m} \frac{1}{n}.T_\ell$$
où $(T_\ell)_{\ell\in\mathbb{N}^*}$ est une suite de v.a. indépendantes suivant la loi géométrique sur $\mathbb{N}^*$ de paramètre de succès p.

Lorsque n est grand, chaque $\frac{1}{n}.T_\ell$ est distribuée approximativement suivant une loi $\mathscr{E}(\lambda)$. On a, pour $t = \frac{k}{n}$,

$$\mathbb{P}(\frac{1}{n}.T_\ell > t) = \mathbb{P}(T_\ell > k) = q^k = ((1-\frac{\lambda}{n})^n)^t \simeq e^{-\lambda.t}$$

et donc, lorsque n est grand, S_m se comporte en loi, approximativement comme $\sum_{\ell=0}^{m} E_\ell$ où les E_ℓ sont indépendantes, $E_\ell \sim \mathscr{E}(\lambda)$.

Cette description est celle du processus de POISSON faite dans le chapitre sur les variables discrètes et il est assez naturel que N suive approximativement une loi de POISSON

L'autre nom de la loi de POISSON c'est la loi de comptage des événements rares.

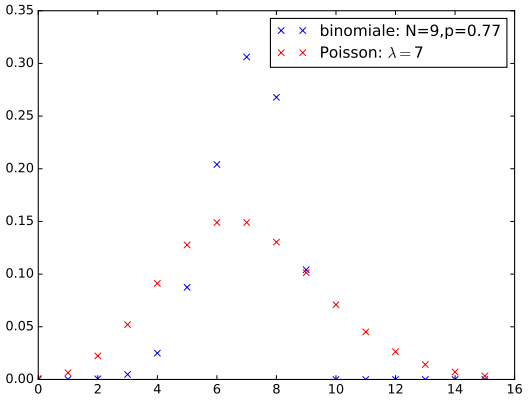


FIGURE 8 – Comparaison Binomiale/Poisson $p = \frac{\lambda}{N}$

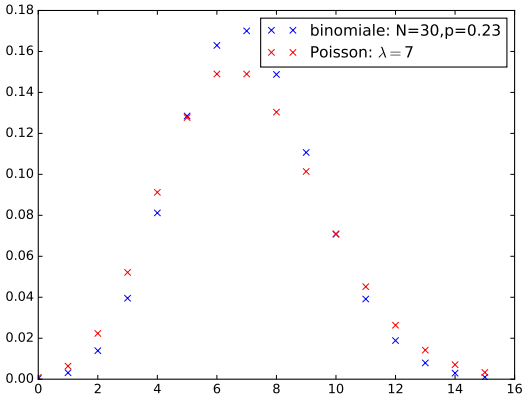


FIGURE 9 – Comparaison Binomiale/Poisson $p = \frac{\lambda}{N}$

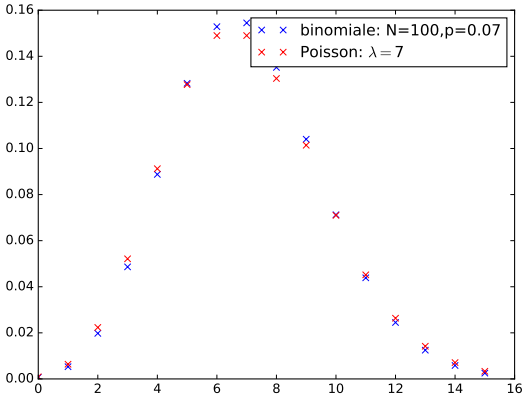


FIGURE 10 – Comparaison Binomiale/Poisson $p = \frac{\lambda}{N}$

3.3 Suites d'extrema

Exercice 3.— 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.

Démonstration. On ne traite que le cas de V_N, le cas de U_N étant similaire.

1. Le premier point est de donner un sens précis à l'assertion finale. Soit F_N la fonction de répartition de V_N, F la fonction de répartition d'une v.a. constante égale à 1. Par V_N est proche de la constante 1, on entend le fait que F_N est proche de F lorsque N est grand. On cherche donc à démontrer que

$$\forall t \in \mathbb{R}, F_N(t) \overset{N\to+\infty}{\to} F(t)$$

   ou une assertion proche de celle-ci.

2. Soit $t \in \mathbb{R}$, on a (calcul usuel de la fonction de répartition d'un max),

$$F_N(t) = \begin{cases} 0 & \text{si } t \leq 0 \\ t^N & \text{si } 0 \leq t \leq 1 \\ 1 & \text{si } t \geq 1 \end{cases} \quad \text{et } F(t) = \begin{cases} 0 & \text{si } t < 1 \\ 1 & \text{si } t \geq 1 \end{cases}$$

   Comme $t^N \to 0$ lorsque $0 \leq t < 1$ et $N \to +\infty$, on a effectivement l'assertion recherchée.

□

Exercice 4.—
1. Donner espérance et variance de V_N (définie dans l'exercice 3) ainsi que des équivalents simples v_N et s_N^2 de ces quantités lorsque $N \to +\infty$.
2. Soit [13] $V_N^* = \frac{1}{s_N}(V_N - v_n)$. 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.

13. Ce n'est pas la centrée réduite de V_N, que valent espérance et variance de V_N^* ? Quelles sont leurs limites lorsque $N \to +\infty$?

3.4 Binomiale et Gaussienne

La forme des coefficients binomiaux

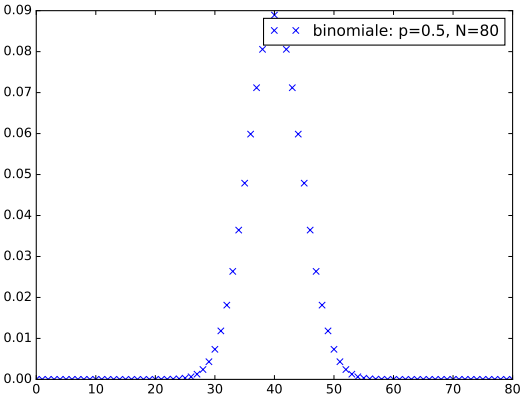


FIGURE 11 – Binomiale

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

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

Plus précisémment, en remarquant que $\mathbb{E}(S)=n.p$, $\mathbb{V}(S)=n.p.(1-p)$,

Proposition 8. *Soit $x \in \mathbb{R}$ fixé, $(k_n)_n$ une suite d'entiers naturels telle que*

$$\forall n \in \mathbb{N}^*, k_n = n.p + \sqrt{n.p.(1-p)}x_n \text{ et } x_n \overset{n\to+\infty}{\to} x$$

On a, lorsque $n \to +\infty$,

$$\binom{n}{k_n} p^{k_n}.(1-p)^{n-k_n} \sim \frac{1}{\sqrt{n.p.(1-p)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Démonstration. Le calcul est difficile (de niveau technique hors-programme). Il est basé sur

— la formule de STIRLING :

$$n! \sim_{n\to+\infty} \sqrt{2\pi} n^n n^{\frac{1}{2}} e^{-n}$$

— les réécritures

$$\frac{k_n}{np} = 1 + \frac{x_n}{\sqrt{n}}\sqrt{\frac{1-p}{p}} \text{ et } \frac{n-k_n}{n(1-p)} = 1 - \frac{x_n}{\sqrt{n}}\sqrt{\frac{p}{1-p}}$$

— et le DL d'ordre 2, $\ln(1+y) = y - \frac{1}{2}y^2 + o\left(y^2\right)$ lorsque $y \to 0$

□

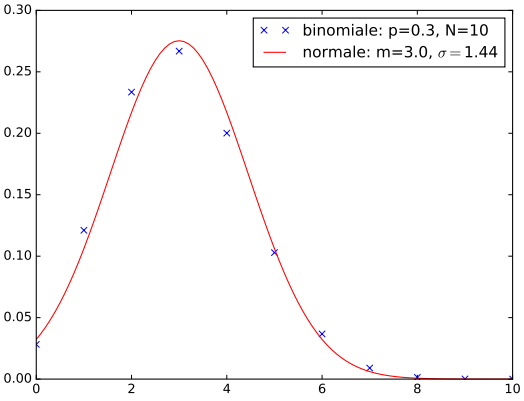


FIGURE 12 – Comparaison Binomiale/Normale

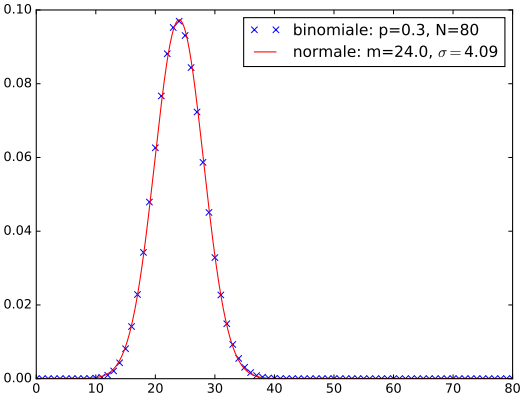


FIGURE 13 – Comparaison Binomiale/Normale

4 LFGN et TCL

4.1 Loi faible des grands nombres

4.1.1 Moyenne empirique et énoncé

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. On commence par un résultat qualitatif

On a vu deux façons de mesurer l'écart d'une v.a.r X à une constante a. L'une utilisant l'écart quadratique $\mathbb{E}(|X-a|^2)$, l'autre l'écart-moyen $\mathbb{E}(|X-a|)$. Une troisième façon, utilisée dans le le théorème suivant est de considérer la famille des quantités $\mathbb{P}(|X-a|>\delta)$ lorsque $\delta > 0$.

Les inégalités de MARKOV, CAUCHY–SCHWARZ et BIENAYMÉ–TCHEBYCHEFF montrent que, pour $\delta > 0$, on a

$$\mathbb{P}(|X-a|>\delta) \leq \frac{\mathbb{E}(|X-a|)}{\delta} \leq \frac{\mathbb{E}(|X-a|^2)^{\frac{1}{2}}}{\delta} \text{ et } \mathbb{P}(|X-a|>\delta) \leq \frac{\mathbb{E}(|X-a|^2)}{\delta^2}$$

Théorème 9 ((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 10 ((Hors programme))**.** *Pour toute fonction h* **continue et bornée** *sur $\mathbb{R}$,*

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

Démonstration. $\Rightarrow$: On a, pour $\delta > 0$ donné, M un majorant de $|h|$ sur $\mathbb{R}$, pour $N \in \mathbb{N}^*$,

$$\begin{aligned} |\mathbb{E}(h(\overline{X}_N)) - \mathbb{E}(h(\mu))| &\leq \mathbb{E}(h(|\overline{X}_N) - h(\mu)|\mathbb{1}_{\{|\overline{X}_N-\mu|\leq\delta\}}) \\ &\quad +2M\mathbb{P}(\{|\overline{X}_N - \mu| > \delta\}) \end{aligned}$$

Soit $\varepsilon > 0$, par continuité de h en μ, il existe $\delta > 0$ tel que

$$\forall x \in \mathbb{R},\ |x-\mu| \leq \delta \Rightarrow |h(x) - h(\mu)| < \frac{\varepsilon}{2}.$$

Pour ce $\delta > 0$, fixé, il existe n_0 tel que

$$\forall N \geq n_0, \mathbb{P}(\{|\overline{X}_N - \mu| > \delta\}) < \frac{\varepsilon}{2(2M+1)}$$

et, l'un dans l'autre, pour tout $N \geq n_0$,

$$|\mathbb{E}(h(\overline{X}_N)) - \mathbb{E}(h(\mu))| < \varepsilon$$

$\Leftarrow$: Pour $\delta > 0$ fixé, on considérons la fonction h_δ définie par

$$h_\delta(x) = 1\!\!1_{\left\{|x-\mu| > \frac{\delta}{2}\right\}}$$

Cette fonction est construite de sorte que

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

Cette fonction est bornée mais pas continue sur $\mathbb{R}$, modifions la en $\tilde{h}_\delta$ (dessin) pour gagner de la continuité en conservant l'inégalité précédente. On peut alors lui appliquer l'hypothèse et, lorsque $N \to +\infty$,

$$\mathbb{E}(\tilde{h}_\delta(\overline{X}_N)) \to \mathbb{E}(h(\mu)) = 0$$

et donc

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

□

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. 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 $Y_{\frac{1}{2}} = 1\!\!1_{\left\{X \leq \frac{1}{2}\right\}}$ 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 $\leq \frac{1}{2}$

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 $Y_x := 1\!\!1_{\{X \leq x\}}$

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. $Y_k = \mathbb{1}\!\mathrm{l}_{\{X=x_k\}}$. Y_k est une v.a de BERNOULLI de paramètre de succès p_k.

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

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

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.

4.1.2 Preuves dans certains cas particuliers

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 11. *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

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

On en déduit

Proposition 12. *Soit $(X_1, \dots, X_N, \dots,)$ 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

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

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. Raffinons un peu cela. On a

$$\begin{aligned} \mathbb{P}(|\overline{X}_N - m| > \delta) &= \frac{2\sqrt{N}}{\sqrt{2\pi}\sigma} \int_{\delta}^{+\infty} e^{-\frac{1}{2}\frac{N.x^2}{\sigma^2}}\, dx \\ = \frac{2}{\sqrt{2\pi}} \int_{\frac{\sqrt{N}\delta}{\sigma}}^{+\infty} e^{-\frac{1}{2}y^2}\, dy &= 2(1 - F_G(\frac{\sqrt{N}\delta}{\sigma})) \end{aligned}$$

où $F_G(t)$ est la fonction de répartition de $G \sim \mathcal{N}(0,1)$,

$$1 - F_G(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{1}{2}y^2}\, dy$$

On sait que $F_G(t) \to 1$ lorsque $t \to +\infty$ mais à quelle vitesse ?

Soit $t > 0$, effectuons une ipp

$$\int_t^{+\infty} \frac{1}{y}.y.e^{-\frac{1}{2}y^2}\, dy \;=\; \left[-\frac{1}{y}e^{-\frac{1}{2}y^2}\right]_t^{+\infty} - \underbrace{\int_t^{+\infty} \frac{1}{y^2}e^{-\frac{1}{2}y^2}\, dy}_{\geq 0} \leq \frac{e^{-\frac{1}{2}t^2}}{t}$$

Et en replaçant dans notre calcul initial

$$\mathbb{P}(|\overline{X}_N - m| > \delta) \;\leq\; \frac{2}{\sqrt{2\pi}} \frac{\sigma}{\delta\sqrt{N}}.e^{-\delta^2 \frac{1}{2}\frac{N}{\sigma^2}}$$

L'intéret de cette estimation est l'apparition de l'exponentielle qui garantit une convergence vers 0 lorsque $N \to +\infty$ beaucoup plus rapide que celle obtenue par BIENAYMÉ–TCHEBYCHEFF.

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,

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

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 $n \to +\infty$,

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

4.1.3 Variance empirique

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

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

Démonstration. Etape de centrage et réduction : Pour faire le calcul, il est plus simple de centrer et réduire les variables. Posons

$$Y = \frac{X-\mu}{\sigma}, X = \mu + \sigma.Y$$

Y est de carré intégrable, d'espérance nulle et de variance 1. On introduit les variables Y_n centrées-réduites des X_n et la variance empirique W_N de ces variables.

La moyenne empirique de ces nouvelles variables vérifie

$$\overline{X}_N = \mu + \sigma.\overline{Y}_N$$

et on a

$$S_N^2 = \sigma^2.W_N$$

Le calcul de l'espérance de S_N^2 se réduit donc au calcul de l'espérance de W_N.

$$\begin{aligned} W_N &= \frac{1}{N}\sum_{n=1}^{N} Y_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N} Y_n\right)^2 \qquad \text{(KOENIG–HUYGHENS !)} \\ \mathbb{E}(W_N) &= \mathbb{V}(Y) - \frac{1}{N^2}\mathbb{V}\left(\sum_{i=1}^{N} Y_n\right) = 1 - \frac{N}{N^2} = \frac{N-1}{N} \end{aligned}$$

(on a utilisé l'indépendance mutuelle des Y_n dans le calcul de $\mathbb{V}\left(\sum_{i=1}^{N} Y_n\right) = \sum_{i=1}^{N}\mathbb{V}(Y_n) = N$) et donc

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

□

Remarque : On trouve, sur certaines calculatrices, deux touches pour calculer la variance (empirique) d'une série statistique. Une touche calculant S_N^2 tel que défini ici, une autre calculant

$$\tilde{V}_N = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X}_N)^2$$

Cette quantité a pour espérance σ^2, la vraie variance de X. C'est un estimateur *non biaisé* de la variance.

Une loi faible des grands nombres pour S_N^2.(HP)

Proposition 13 (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, *cf.* exercice 8.

Exercice 8.— Soit $(X_n)_{n\geq 1}$ un échantillon d'une variable aléatoire X. On suppose que X admet un moment (centré) d'ordre 4, $\mu_4 = \mathbb{E}((X-\mu)^4)$ où $\mu = \mathbb{E}(X)$.

Soit $N \in \mathbb{N}^*$, l'espérance empirique et la variance empirique non-biaisée du N-échantillon $(X_n)_{1\leq n\leq N}$ sont définies respectivement par

$$M_N = \frac{1}{N}\sum_{n=1}^{N} X_n \text{ et } \Sigma_N^2 = \frac{1}{N-1}\left(\sum_{n=1}^{N}(X_n - M_N)^2\right)$$

1. Montrer que $\mathbb{E}(\Sigma_N^2) = \mathbb{V}(X)$.
2. On admet que

$$\mathbb{V}(\Sigma_N^2) = \frac{1}{N}\mu_4 - \frac{(N-3)}{N(N-1)}\mathbb{V}(X)^2$$

Montrer que pour tout $\varepsilon > 0$,

$$\mathbb{P}(|\Sigma_N^2 - \mathbb{V}(X)| > \varepsilon) \overset{N\to+\infty}{\to} 0$$

3. (Facultatif) Démontrer la formule admise en 2.

4.1.4 Versions quantitatives

La démonstration que l'on a faite de la loi faible des grands nombres dans le cas où X est de carré intégrable est instructive au sens où l'on obtient une estimation de la vitesse de convergence. En exercice, on montre que si X est bornée alors, pour tout $\delta > 0$, il existe $\alpha > 0$ tel que

$$\mathbb{P}(|\overline{X}_N - \mu| > \delta) \leq 2e^{-\alpha.N}$$

Cette estimation est bien meilleure que l'estimation obtenue pour X de carré intégrable. Ce type d'estimation a déjà été rencontré dans le cas où X est gaussienne.

Exercice 9.-⋆- Soit $M > 0$ et une variable aléatoire reélle X prenant un nombre fini de valeurs $-M < x_1 < x_2 < \cdots < x_K < M$ dans un intervalle $[-M, +M]$, d'espérance μ.

Soit $\delta > 0$. Il s'agit dans cet exercice d'obtenir une majoration de $\mathbb{P}(|\overline{X}_n - \mu| > \delta)$ qui soit plus fine que celle obtenue via l'inégalité de BIENAYMÉ–TCHEBYCHEFF exhibée dans le cours.

1. On pose, pour $t \in \mathbb{R}$, $\psi_X(t) = \ln \mathbb{E}(e^{tX})$.

1.a. Montrer que ψ_X est de classe $\mathscr{C}^\infty$ sur $\mathbb{R}$. Exprimer ses dérivées premières et secondes en fonction de $\mathbb{E}(Xe^{tX})$, $\mathbb{E}(Xe^{tX})$ et $\mathbb{E}(X^2e^{tX})$.

(On pourra introduire les nombres $p_k = \mathbb{P}(X = x_k)$ pour faire les calculs mais les résultats doivent s'exprimer sans que ces nombres apparaissent)

1.b. Montrer, en utilisant l'inégalité de CAUCHY–SCHWARZ que $\psi''_X \geq 0$. On admettra que, au cas où X n'est pas constante, on a $\psi''_X > 0$.

1.c. Calculer $\psi'_X(0)$.

1.d. Montrer que pour $\delta > 0$ positif donné, il existe $\tau > 0$ tel que $\psi_X(\tau) - \tau(\mu + \delta) < 0$.

1.e. Montrer de même que pour $\delta > 0$ positif donné, il existe $\tau' < 0$ tel que $\psi_X(\tau') - \tau'(\mu - \delta) < 0$.

Indication:On pourra dresser les tableaux de variations de ces fonctions de τ au voisinage de 0.

2. Soit $\delta > 0$.

2.a. En appliquant l'inégalité de MARKOV à la v.a e^{tX}, montrer que pour tout $t \geq 0$,

$$\mathbb{P}(X > \mu + \delta) \leq e^{\psi_X(t) - t(\mu+\delta)}$$

2.b. Montrer de même que pour tout $t \leq 0$,

$$\mathbb{P}(X < \mu - \delta) \leq e^{\psi_X(t) - t(\mu-\delta)}$$

3. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite indépendante de même loi que X, $\overline{X_n}$ la n-ième moyenne empirique.

3.a. Montrer que $\psi_{\overline{X_n}}(t) = n.\psi_X(\frac{t}{n})$.

3.b. En déduire qu'il existe $c = c(\delta) > 0$ tel que pour tout $n \in \mathbb{N}^*$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \delta) \leq 2e^{-c.n}$$

4. Pour X une v.a suivant l'une des lois suivantes (discrète ou à densité) : géométrique, POISSON, exponentielle ou normale, calculer ψ_X et montrer que la méthode précédente reste valable pour obtenir l'inégalité finale.

4.1.5 Un exemple où la loi faible ne s'applique pas

Ma v.a à densité *non intégrable* préférée est une v.a suivant une loi de CAUCHY de paramètre $\lambda > 0$.

$$X \sim \frac{\lambda}{\pi} \frac{1}{1 + (\lambda.x)^2}\, dx$$

Cette variable n'admet pas d'espérance mais vu le caractère symétrique de la densité, l'affirmation que cette variable est centrée n'est pas sans fondement sensible.

Pour cette distribution, il n'y a pas de résultat final du type loi faible des grands nombres.

En exercice : Si X_1 et X_2 sont CAUCHY, indépendantes, de paramètres λ_1 et λ_2 alors $X_1 + X_2$ est CAUCHY de paramètre $\lambda_1 + \lambda_2$.

Si $(X_n)_{n\geq 1}$ sont indépendantes, CAUCHY de paramètre 1, alors $\overline{X}_N$ est CAUCHY, de paramètre 1. La distribution de $\overline{X}_N$ est donc indépendante de N et ne s'approche pas de la distribution d'une v.a constante comme le voudrait la LFGN.

4.2 Le théorème de la limite centrale

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 $\mathscr{N}(0,1)$, *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...

4.2.1 Enoncé I

Théorème 14 ((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 $N \to +\infty$,*

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

Interprétation, reformulation

— La signification de ceci est que la v.a. centrée réduite M_N^*, est approximativement distribuée comme une $\mathscr{N}(0,1)$.

— Au niveau des formules de transfert, pour toute fonction h, *continue, bornée* sur $\mathbb{R}$, lorsque $N \to +\infty$,

$$\mathbb{E}(h(M_N^*)) \to \mathbb{E}(h(G)) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(x) e^{-\frac{x^2}{2}}\, dx$$

où $G \sim \mathscr{N}(0,1)$.

— On a vu ce théorème dans le cas où $X \sim \mathscr{B}(p)$, $N.\overline{X}_N \sim \mathscr{B}(N,p)$. Sous cette forme, il est connu sous le nom de théorème de DE MOIVRE–LAPLACE

— Le TCL a ceci de puissant que la loi « limite » de M_N^* ne dépend pas de la loi de X, juste du caractère « carré intégrable » de X.

— C'est une des raisons pour lequelles, dans de nombreux modèles en physique, chimie, biologie, finance..., on considère que les variables primitives sont gaussiennes, car sommes de nombreux petits phénomènes indépendants.

— En finance, cette hypothèse a mené à quelques krachs boursiers...tout n'est pas de carré intégrable...

Exercice 10.-⋆- 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 $\mathscr{N}\left(0,\sigma^2\right)$.

Exercice 12.— Soit $(T_k)_{k\geq 1}$ une suite de variables indépendantes distribuées suivant une loi exponentielle $\mathscr{E}(\lambda)$. On considère $\overline{T}_n$ la n-ième moyenne empirique.

1. Donner une formule pour la densité de $\overline{T}_n$. (On pourra se servir de l'exercice (déjà fait, forcément déjà fait) reliant densités γ et sommes de v.a. exponentielles indépendantes.)

2. Quelle est son espérance ?

3. Donner ϕ_n une densité continue de $\overline{T}_n-\mathbb{E}(T_n)$.

⋆⋆⋆

La suite de l'exercice nécessite un équivalent fameux (mais hors programme), l'équivalent de STIRLING :

$$n!\sim_{n\to+\infty}\sqrt{n}.\sqrt{2.\pi}.n^n.e^{-n}$$

⋆⋆⋆

4. Donner, pour $x\in\mathbb{R}$, la limite de $\phi_n(x)$. Faire le lien avec la LFGN.

5. Pour $x\in\mathbb{R}$ donné, donner une équivalent simple de $\phi_n(\frac{x}{\sqrt{n}})$ lorsque $n\to+\infty$? Faites le lien avec le TCL.

Exercice 13.— Médiane empirique. Soit $n\in\mathbb{N}^*$, $(x_1,\ldots,x_n,\ldots,x_{2n+1})$ un $2n+1$-uplet de nombres réels tous distincts. Pour calculer la médiane de ce $2n+1$-uplet on classe ces nombres par ordre croissant, et on prend le nombre se retrouvant en position $n+1$.

Soit X une v.a suivant la loi $\mathscr{U}_{[-1,1]}$. $(X_k)_{k\geq 1}$ une suite de v.a indépendantes, de même loi que X.

Pour $n\geq 1$, on pose M_n la médiane de $(X_1,\ldots,X_n,\ldots,X_{2n+1})$. On admet que la probabilité que deux de ces nombres soient égaux est nulle et M_n est donc définie presque sûrement.

Le but de cet exercice est de déterminer la loi de M_n ainsi que son comportement asymptotique.

1. Déterminer la fonction de répartition F de la variable X. Tracer son graphe.

2. Soit $n\in\mathbb{N}^*$ fixé. On définit, pour $m\in\mathbb{R}$, la variable aléatoire

$$C_m=\sum_{k=1}^{2n+1}\mathbb{1}_{\{X_k\leq m\}}$$

2.a. Quelle est, en fonction de n et $F(m)$, la loi de C_m ?

Indication:Voir C_m comme une somme de BERNOULLI indépendantes.

2.b. Montrer l'égalité d'événements $(C_m\geq n+1)=(M_n\leq m)$.

2.c. En déduire une formule pour $\phi_n(m)$ en fonction de n et $F(m)$ où ϕ_n est la fonction de répartition de M_n.

3. Montrer que M_n est une variable à densité, dont une densité est

$$\phi_n'(m) = \frac{1}{2^{2n+1}} \frac{(2n+1)!}{(n!)^2} (1-m^2)^n \mathbb{1}_{\{|m|\leq 1\}}$$

Quelle est son espérance ? sa variance ?
Indication:On donne

$$\int_{-1}^{+1} x^2 (1-x^2)^n \, dx = 2^{2n+1} \frac{(n!)^2.(2n+2)}{(2n+3)!}$$

4. Est-ce la même loi que $\overline{X}_{2n+1}$?
5. Montrer que pour tout $\delta > 0$, lorsque $n \to +\infty$, $\mathbb{P}(|M_n| > \delta) \to 0$.
6. Soit $N_n = \sqrt{2n} M_n$. Donner une densité ψ_n de N_n continue sur $\mathbb{R}$. A x fixé, quelle est la limite de $\psi_n(x)$ lorsque $n \to +\infty$? Quel serait un énoncé raisonnable dans l'esprit du TCL ?

4.2.2 Enoncé II

Théorème 15. *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.

4.2.3 Construction d'intervalles de confiance et tests de conformité

Fonction de répartition de la Gaussienne normalisée

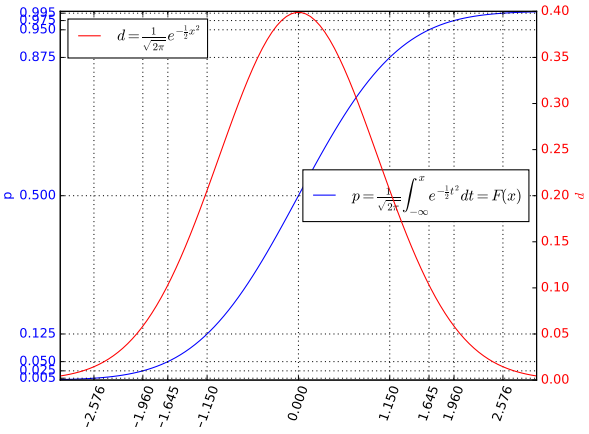


FIGURE 14 – La densité normale $\mathcal{N}(0,1)$ sa fonction de répartition

Notons F_G la fonction de répartition de la gaussienne normalisée ; Il s'agit d'une fonction s'exprimant sous forme d'intégrale [14]. Elle est très directement liée à la *fonction d'erreur gaussienne* [15], une fonction tabulée depuis longtemps et présente [16] dans tous les langages et logiciels sérieux de calcul.

Ce qui nous intéresse ici, ce sont les quantiles de F_G. Plus précisément, si $1-\alpha \in]0,1[$ est un niveau de confiance statistique, on note $u_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ défini par

$$F_G(u_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2} \text{ } i.e. \text{ } u_{1-\frac{\alpha}{2}} = Q_G(1-\frac{\alpha}{2})$$

On a [17], si $G \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(|G| < u_{1-\frac{\alpha}{2}}) = 1-\alpha \text{ et } \mathbb{P}(|G| > u_{1-\frac{\alpha}{2}}) = \alpha$$

Exercice 14.—
1. Déterminer, d'après le graphe, la valeur de u correspondant aux niveaux de confiance 99%, 95% et 75%.
2. Seriez vous capables de les déterminer d'après la version tabulée de la fonction F_G donnée dans la table 1 ?
3. Quelle instruction Python utiliser pour retrouver ces valeurs ?

Le théorème 15 a pour conséquence (on conserve les notations)

Théorème 16. *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 $N \to +\infty$,*

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

14. dont *Il n'existe pas de formule* l'exprimant de façon algèbrique $(+, . \circ)$ en termes des fonctions usuelles. Un tel résultat, c'est de l'algèbre très difficile.
15. *cf.* `http://fr.wikipedia.org/wiki/Fonction_d\%27erreur`
16. En Python, dans le module `scipy.stats`, on accède à ses valeurs par `norm.cdf` et aux valeurs de son inverse, la fonction des quantiles, par `norm.ppf`
17. Utiliser la symétrie naturelle de F_G

u	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999

TABLE 1 – Table des valeurs de $F_G(u)$, $u \in [0, 2.99]$

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 16.

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 [18], 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 $\left|\sqrt{N}.\frac{\bar{x}-\mu}{\sigma_x}\right| > u_{1-\frac{\alpha}{2}}$: 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 $\left|\sqrt{N}.\frac{\bar{x}-\mu}{\sigma_x}\right| < u_{1-\frac{\alpha}{2}}$: on retient (H_0) au niveau de confiance $1-\alpha$: notre population est conforme à cette hypothèse à ce niveau de confiance.

4.2.4 Construction d'intervalle 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 $\bar{x}$ 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

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

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).

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% ?

Exercice 16.— Le diamètre D de *Corbicula fluminea* suit une loi inconnue.

Pour un individu adulte, le diamètre d de cet individu est la valeur prise par une v.a aléatoire D. En un sens, la nature « calcule » D à partir de paramètres (aléatoires) du mileu de culture ainsi que de paramètres (aléatoires) propres à l'individu. La loi de D ne nous est pas accessible.

On mesure le diamètre d_k de chaque individu d'une population de 100 individus numérotés de 1 à $n = 100$. Ces individus sont supposés avoir évolué indépendamment les uns des autres.

Pour récapituler le résultat de ces mesures, on calcule la moyenne $\overline{d} = 3cm$ des nombres d_k ainsi que leur écart-type $\sigma = 0.5cm$.

18. et pourvu que N soit suffisamment grand pour que la limite dans le théorème 16 issue puisse être considérée comme une égalité

1. En supposant que D est une variable de carré intégrable, d'espérance inconnue m et de variance connue s^2 et en vous servant du graphe de la fonction de répartition de la loi normale $\mathscr{N}(0,1)$, évaluer

$$\mathbb{P}(m \in [\overline{d} - 2s, \overline{d} + 2s])$$

Quel théorème d'approximation utilisez-vous ?
2. On suppose que D est une variable de carré intégrable, d'espérance inconnue m et de variance s^2 inconnue.
3.a. Comment évaluer s^2 ?
3.b. Evaluer

$$\mathbb{P}(m \in [\overline{d} - 2\sigma, \overline{d} + 2\sigma])$$

Quel théorème d'approximation utilisez-vous ?
3.c. (Question ouverte) Pour $\varepsilon > 0$ donné, comment évaluer

$$\mathbb{P}(s^2 \in [\sigma^2 - \varepsilon, \sigma^2 + \varepsilon])$$

Exercice 17.—
1. Afin d'étudier le pourcentage p de consommateurs satisfait p ar le produit A, on interroge 100 consommateurs et 56 déclarent être satisfaits. Est-ce suffisant pour continuer l'exploitation du produit A ? (on cherchera un intervalle de confiance à 95%)
2. En supposant qu'on garde la même moyenne empirique de $0,56$, et le même risque $\alpha = 0,05$, combien de personnes doit-on interroger pour prendre une décision ?

Exercice 18.—Chez le petit lapin, la durée moyenne de gestation est de 30 jours. On étudie un échantillon de 66 familles de gros lapins, pour lesquelles on observe une durée moyenne de gestation de 30,83 jours avec un écart-type de 4,07 jours. Peut-on conclure que la durée de gestation est significativement différente chez les petits et les gros lapins ?

Exercice 19.—On utilise un dé cubique parfait. Combien faut-il effectuer de lancers pour pouvoir affirmer avec un risque d'erreur inférieur à 5%, que la fréquence d'apparition du numéro 6 au cours de ces lancers differera de $\frac{1}{6}$ d'au plus 0, 01 ?

On donnera une réponse en utilisant l 'inégalité de BT, et une autre (plus précise ?) en utilisant le TCL.

Exercice 20.—Une usine fabrique des cables. On suppose que la charge maximale supportée par un cable, exprimée en tonnes, est une V. A. qui suit une loi normale $\mathscr{N}\left(\mu, 0,5^2\right)$). Une étude portant sur 50 cables a donné une moyenne des charges maximales supportées égales à 12,2 tonnes.
1. Déterminer l'intervalle de confiance à 99% de la charge maximale moyenne de tous les cables fabriqués par l'usine.
2. Déterminer la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance à 99% soit inférieure ou égale à $0,2$?

Exercice 21.—Un producteur de coton produit 950 Kg au km^2. Il veut améliorer sa production en testant une nouvelle espèce de coton mais avant cela il demande conseil à un statisticien. Ce dernier récupère les données de 7 champs et obtient une moyenne empirique de 1 032 kg au km^2 pour une variance empirique de 3 420.
1. En quelle unité est la variance ?
2. Déterminer un intervalle de confiance au risque de 0, 01 pour la production moyenne de cette nouvelle espèce de coton. On arrondira au kg/km^2.

4.2.5 Barres d'erreurs sous Exxxel et autres embrouilles

Barres d'erreurs Lorsque l'on trace un diagramme sous Exxxel ou un autre tableur, on peut lui faire afficher des « barres d'erreur [19] ». Maintenant si nous avons deux séries de données (imaginons : les tailles des hommes et des femmes dans deux échantillons de N individus), on peut afficher côte à côte ces segments afin de se donner une idée quant à la coïncidence réelle des distributions de cette variable dans chacune des populations.

Certains, nommons les « les amateurs de barres d'erreur », considèrent que

— « barres d'erreurs chevauchantes ? » : il n'y a pas de différence significative dans les distributions ;

— « barres d'erreurs non chevauchantes ? » : les distributions sont différentes.

L'idée n'est pas bête mais mérite d'être examinée à la lumière de notre philosophie des tests.

Supposons par exemple que la population d'hommes, resp. de femmes, dont la taille T_h, resp T_f, est distribuée suivant une loi normale $\mathscr{N}\left(173,\sigma^2\right)$ avec $\sigma = 10$, resp. $\mathscr{N}\left(163,\sigma^2\right)$ et que pour un échantillon de N individus de chaque population, on obtienne $\overline{t_h} = 172, \overline{t_f} = 164$ avec écart-types valant exactement $\sigma = 10$. Les barres d'erreurs dans ce cas sont chevauchantes.

— Si les populations testées comportent, disons $N = 9$ individus, les intervalles de confiance à 95% sont repectivement $\left[\overline{t_h} - 2.\frac{\sigma}{\sqrt{9}}, \overline{t_h} + 2.\frac{\sigma}{\sqrt{9}}\right] = [165.33, 178.66]$ et $\left[\overline{t_f} - 2.\frac{\sigma}{\sqrt{9}}, \overline{t_f} + 2.\frac{\sigma}{\sqrt{9}}\right] = [157.33, 170.66]$. L'hypothèse $(H_0) : \mu_h = \mu_f = 168$ ne sera pas rejetée par le test et donc, il est possible que T_h et T_f aient même distribution ; On est conduit à accepter (H_0) alors que celle-ci est fausse : il s'agit d'une erreur de deuxième espèce, la même que fait l'amateur de barres d'erreurs.

— Si les populations testées comportent, disons $N = 100$ individus, les intervalles de confiance à 95% sont respectivement $\left[\overline{t_h} - 2.\frac{\sigma}{\sqrt{100}}, \overline{t_h} + 2.\frac{\sigma}{\sqrt{100}}\right] = [170, 174]$ et $\left[\overline{t_f} - 2.\frac{\sigma}{\sqrt{100}}, \overline{t_f} + 2.\frac{\sigma}{\sqrt{100}}\right] = [162, 166]$. Toute hypothèse $(H_0) : \mu_h = \mu_f = \mu_0$ est rejetée et on rejettera donc l'hypothèse « T_h et T_f ont même distribution ».

L'acceptation ou le rejet de l'hypothèse de même distribution dépend donc de la *puissance* du test, ici, directement fonction de nombre d'individus.

Ce qu'il faut en tirer, c'est qu'un test de conformité d'une population à une moyenne est d'autant plus puissant que le nombre d'individus est important. Le test de chevauchement est un test très peu puissant pour conclure à l'égalité de distributions.

N doit être grand ! (Si on utilise un théorème limite...) Il faut se méfier de l'application aveugle de la formule de l'intervalle de confiance et notamment la signification du 1.96.

L'hypothèse de travail primitive pour mettre en oeuvre intervalles de confiance et tests de conformité de moyennes est le caractère Gaussien ou presque de la distribution de $T_\mu = \frac{M_N - \mu}{\frac{S_N}{\sqrt{N}}}$. Même lorsque la distribution du caractère est supposée Gaussienne, cela nécéssite des valeurs de N assez grandes. Dans le cas Gaussien, la variable T_μ suit une loi connue [20] et on a [21] , pour $N = 10$, que $\mathbb{P}\left(\left|T_\mu\right| \geq 2.228\right) \simeq 5\%$. Ce 2.228 n'est pas vraiment le 1.960 de la loi normale et pour descendre à 2 (ce qu'on utilise en général comme arrondi de 1.960), il faut monter à $N = 60$. Pour $N = 10$, les vrais intervalles de confiance à 95%, sous hypothèse Gaussienne, sont *grosso modo* 10% plus larges que ceux donnés par l'application aveugle du théorème 16. On ne parle même pas de ce qui arrive quand on n'est pas sous hypothèse Gaussienne.

19. On suppose ici que l'on représente un intervalle de 1/2-largeur un écart-type, centré en la moyenne empirique.
20. la loi de STUDENT (*cf.* TP 5)
21. En prenant pour S_N^2 la variance empirique *non biaisée*, *cf.* `http://fr.wikipedia.org/wiki/Test_de_Student`