

Chapitre 2 – La théorie des probabilités (2^e partie)

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

Contexte

Il existe plusieurs lois de probabilité bien connues, qui ont même un **nom**, et pour lesquelles on connaît bien le comportement et les caractéristiques (espérance, variance, forme).

Plusieurs de ces lois sont reliées à des expériences aléatoires simples, par exemple une expérience qui résulte en un **succès** ou un **échec**.

Épreuve de Bernoulli

Expérience aléatoire ayant deux résultats possibles que l'on désigne arbitrairement par SUCCÈS et ÉCHEC. On note p la probabilité de succès.

Exemples : le sexe d'un nouveau-né (mâle / femelle); l'état d'un arbre (mort / vivant); une tumeur (bénigne / maligne); la germination d'une graine (oui / non).

Quelques lois classiques

Loi (section des notes)	Contexte typique
Binomiale (2.6.1)	Nombre de succès parmi n épreuves de Bernoulli indépendantes
Géométrique (2.6.2)	Nombre d'épreuves de Bernoulli indépendantes pour obtenir un succès
Hypergéométrique (2.6.3)	Nombre de succès parmi n épreuves (tirages sans remise)
Poisson (2.6.4)	Nombre d'événements dans un intervalle de temps (ou d'espace)
Uniforme continue (2.6.5)	Densité constante sur $[a, b]$
Normale (2.6.6)	Plusieurs phénomènes; loi d'une somme ou d'une moyenne (TLC)
Student (3.5.2)	Moyenne d'un échantillon normal de variance inconnue
χ^2 (3.6.2)	Variance d'un échantillon normal; tableaux de fréquences
Fisher (4.2.2)	Comparer des variances

Dans ce chapitre : les lois **binomiale**, de **Poisson** et **normale**.

Un concept important : v.a. indépendantes

Dans ce chapitre, on parlera souvent de variables aléatoires indépendantes.

Variables aléatoires indépendantes

Deux variables aléatoires sont **indépendantes** si connaître la valeur de l'une ne nous renseigne en rien sur la valeur de l'autre.

Rappel (chapitre 2, 1^{re} partie) : si X et Y sont indépendantes,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Exemple : je note le sexe d'un veau ($X = 1$ si femelle) et je lance un dé ($Y =$ face obtenue). X et Y sont indépendantes.

2.6.1 La loi binomiale

Loi de Bernoulli

Soit Y une variable aléatoire telle que $Y = 1$ si l'épreuve est un succès et $Y = 0$ si c'est un échec.

Définition. On dit que Y suit une **loi de Bernoulli** de paramètre p , noté $Y \sim \text{Bernoulli}(p)$, si

$$p(y) = \begin{cases} p & \text{si } y = 1, \\ 1 - p & \text{si } y = 0, \\ 0 & \text{sinon.} \end{cases}$$

Espérance :

Variance :

Exemple 1

Selon la directive européenne pour la commercialisation des semences de céréales, le taux de germination minimal pour les grains d'avoine doit être de 85 %. Vous plantez 5 grains d'un lot qui suit cette norme.

Quelle est la probabilité qu'un seul grain ne germe pas ?

- Épreuve de Bernoulli : planter un grain ; succès = il germe.
- Probabilité de succès : $p = 0.85$.
- Nombre d'épreuves (indépendantes) : $n = 5$.

Résultats avec exactement un échec (E) :

EGGGG, GEGGG, GGEGG, GGEGE, GGGGE

Chacun a probabilité $0.85^4 \times 0.15$... et il y en a combien ?

Rappel : coefficient binomial (section 2.3.4)

Coefficient binomial

Le nombre de sous-ensembles de r objets choisis sans remise parmi n objets distincts est

$$\binom{n}{r} = \frac{n!}{r! (n - r)!}.$$

- Combien de sous-ensembles de 2 éléments peut-on former à partir de l'ensemble $\{A, B, C, D, E\}$?
- Combien d'échantillons différents de taille 6 peut-on tirer d'une population de taille 49?

Loi binomiale

Soit $Y_1, \dots, Y_n$ des v.a. **indépendantes** de loi Bernoulli(p), et $X = Y_1 + Y_2 + \dots + Y_n$ le nombre total de succès.

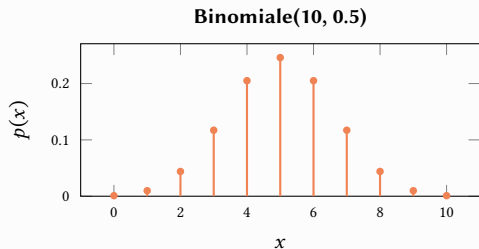
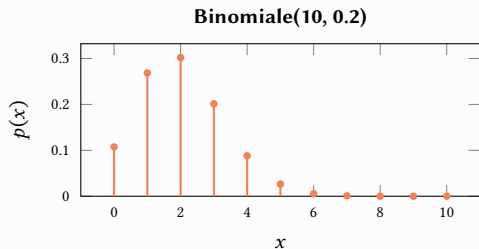
On cherche la probabilité d'avoir k succès, donc $n - k$ échecs.

- Chaque résultat avec k succès a probabilité $p^k(1 - p)^{n-k}$.
- Il y a $\binom{n}{k}$ façons de placer les k succès parmi les n épreuves.

Définition. X suit une **loi binomiale** de paramètres n et p , noté $X \sim \text{Binomiale}(n, p)$:

$$p(k) = P(X = k) = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{si } k \in \{0, 1, \dots, n\}, \\ 0 & \text{sinon.} \end{cases}$$

Allure de la loi binomiale



Dans tous les cas :

- les valeurs possibles sont $0, 1, 2, \dots, n$;
- les valeurs les plus probables sont autour de np ;
- la fonction de masse croît jusqu'à environ np , puis décroît.

Espérance et variance

Proposition. Si $X \sim \text{Binomiale}(n, p)$, alors

$$\mu = E(X) = np \quad \text{et} \quad \sigma^2 = \text{Var}(X) = np(1 - p).$$

Pourquoi? On écrit $X = Y_1 + \dots + Y_n$ avec $Y_i \sim \text{Bernoulli}(p)$ indépendantes :

$$\begin{aligned} E(X) &= E(Y_1) + \dots + E(Y_n) = p + \dots + p = np, \\ \text{Var}(X) &= \text{Var}(Y_1) + \dots + \text{Var}(Y_n) = np(1 - p). \end{aligned}$$

[propriétés 2 et 4 de l'espérance et de la variance, 1^{re} partie]

Exemple 1 (suite)

Vous plantez 5 grains d'un lot dont le taux de germination est de 85 %.

Quelle est la probabilité qu'un seul grain ne germe pas ?

Exemple 2

Une étude révèle que 95 % des canards colverts ont des parasites internes. Un biologiste capture 7 canards lors d'une séance d'échantillonnage.

- a) Quelle est la probabilité qu'il capture plus de 5 canards infestés?
- b) Si plusieurs biologistes capturent 7 canards chacun, le nombre moyen de canards infestés devrait être environ...
- c) L'écart type du nombre de canards infestés par biologiste devrait valoir...

Calculs dans R

Pour $X \sim \text{Binomiale}(n, p)$: `dbinom(k, n, p)` donne $P(X = k)$ et `pbinom(k, n, p)` donne $P(X \leq k)$.

Exemple 1 : $P(4 \text{ grains germent sur } 5)$

```
> dbinom(4, size = 5, prob = 0.85)
[1] 0.3915047
```

Exemple 2 : $P(X > 5) = 1 - P(X \leq 5)$

```
> 1 - pbinom(5, size = 7, prob = 0.95)
[1] 0.9556195
```

QUIZ!

Deux parents sont porteurs sains (génotype Aa) de l'allèle récessif responsable de la fibrose kystique. Chacun de leurs enfants a une probabilité $1/4$ d'être atteint (génotype aa), indépendamment des autres. Le couple a 4 enfants; soit X le nombre d'enfants atteints.

- Quelle est la loi de X ? Vérifiez les conditions.
- Combien d'enfants atteints ce type de famille a-t-il en moyenne?
- Quelle est la probabilité qu'aucun des 4 enfants ne soit atteint?

2.6.4 La loi de Poisson

Contexte

La loi de Poisson sert à modéliser une **fréquence** : un nombre d'événements qui se produisent pendant un certain temps ou dans un certain espace.

Exemples :

- X = nombre d'animaux qui s'abreuvent à une source en 1 h ;
- Y = nombre de décès dus à la tuberculose en un mois ;
- Z = nombre de chevreuils par km^2 dans un parc national ;
- W = nombre de colonies de bactéries sur une plaque de gélose.

On l'obtient aussi comme **loi des événements rares** : une binomiale avec n très grand et p très petit (on y revient plus loin).

Conditions d'application

Pour que la loi de Poisson soit un bon modèle, il faut que la réalité s'approche des conditions suivantes :

- les nombres d'événements dans des intervalles qui ne se chevauchent pas sont des v.a. **indépendantes** ;
- les événements se produisent au hasard, mais à un **taux moyen constant** sur tout l'intervalle considéré ;
- la probabilité que deux événements ou plus surviennent **simultanément** est nulle.

Loi de Poisson

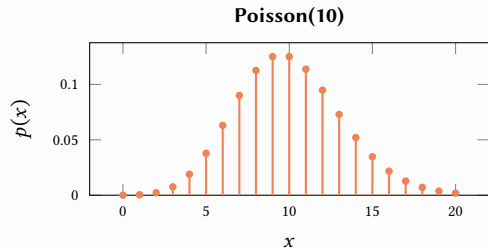
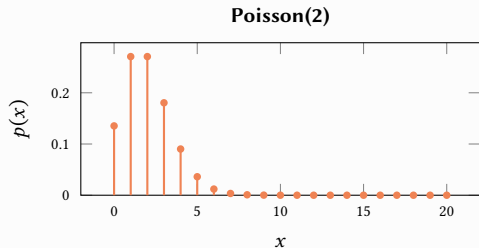
La loi de Poisson est caractérisée par un paramètre λ , le taux d'occurrence (nombre moyen d'événements) sur l'intervalle considéré.

Définition. X suit une **loi de Poisson** de paramètre $\lambda > 0$, noté $X \sim \text{Poisson}(\lambda)$, si

$$p(k) = P(X = k) = \begin{cases} \frac{e^{-\lambda} \lambda^k}{k!} & \text{si } k \in \{0, 1, 2, \dots\}, \\ 0 & \text{sinon.} \end{cases}$$

Proposition. Si $X \sim \text{Poisson}(\lambda)$, alors $E(X) = \lambda$ et $\text{Var}(X) = \lambda$.

Allure de la loi de Poisson



- Les valeurs possibles sont tous les entiers non négatifs 0, 1, 2,
- Les valeurs les plus probables sont concentrées autour de l'espérance λ .
- Plus λ est grand, plus la distribution est étalée (car $\text{Var}(X) = \lambda$) et symétrique.

Siméon-Denis Poisson (1781–1840)

Le chirurgien manqué... mais excellent mathématicien et physicien.



Propriété d'additivité

Proposition. Si $X_1, \dots, X_n$ sont des v.a. indépendantes avec $X_i \sim \text{Poisson}(\lambda_i)$, alors

$$X_1 + \dots + X_n \sim \text{Poisson}(\lambda_1 + \dots + \lambda_n).$$

En pratique : si on observe en moyenne λ événements par unité de temps (ou d'espace), le nombre d'événements dans t unités suit une loi $\text{Poisson}(\lambda t)$.

Exemple 3

Vous dénombrez les globules blancs sur une lame quadrillée présentant 25 carrés de $0,1\ \mu\text{L}$ chacun. Supposons qu'il y a en moyenne 5 globules blancs par carré.

a) Quelle est la probabilité d'observer moins de 3 globules blancs sur un carré?

b) Quelle est la probabilité d'observer 45 globules blancs dans un microlitre?

Calculs dans R

Pour $X \sim \text{Poisson}(\lambda)$: `dpois(k, lambda)` donne $P(X = k)$ et `ppois(k, lambda)` donne $P(X \leq k)$.

Exemple 3 a) : $P(X < 3) = P(X \leq 2)$ avec $\lambda = 5$

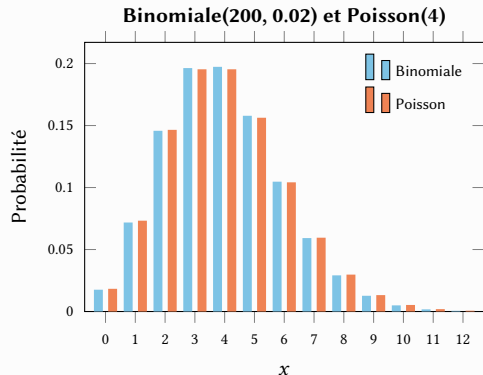
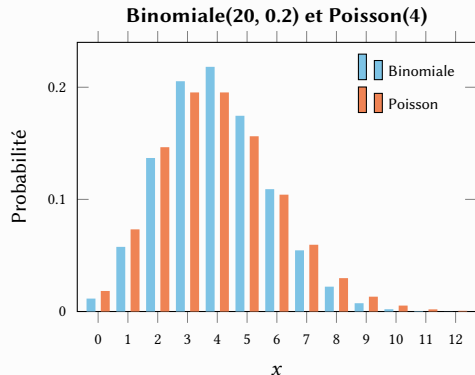
```
> ppois(2, lambda = 5)
[1] 0.1246520
```

Exemple 3 b) : un microlitre = 10 carrés, donc $\lambda = 10 \times 5 = 50$

```
> dpois(45, lambda = 50)
[1] 0.04582624
```

Loi des événements rares

La loi de Poisson peut servir d'approximation de la loi binomiale lorsque n est grand et p petit, avec $\lambda = np$:



QUIZ!

Une mutation donnée survient avec une probabilité de 10^{-5} à chaque division cellulaire, indépendamment d'une division à l'autre. Une culture bactérienne compte 300 000 divisions. Soit X le nombre de mutants obtenus.

- Quelle est la loi exacte de X ?
- Par quelle loi plus simple peut-on l'approximer? Avec quel paramètre?
- Quelle est la probabilité d'obtenir au moins un mutant?

QUIZ!

Pour chaque variable, quelle loi utiliser : **binomiale**, **Poisson**, ou **ni l'une ni l'autre** ?

- a) Le nombre de graines qui germent parmi 20 graines d'un même lot.
- b) Le nombre de colonies bactériennes sur une boîte de Petri.
- c) Le nombre de mutations dans un gène de 1 500 paires de bases, si chaque base mute avec probabilité 10^{-4} , indépendamment des autres.
- d) La masse (en g) d'un nouveau-né.
- e) Le nombre de porcelets malades dans une portée de 10 lors d'une épidémie de grippe porcine (maladie contagieuse).

2.6.6 La loi normale

La loi normale

La loi normale est probablement la loi la plus utilisée en statistique. Quelques raisons de sa popularité :

- elle modélise bien la distribution de beaucoup de variables d'intérêt (tailles, masses, rendements...);
- elle approxime la loi d'une somme ou d'une moyenne de plusieurs variables aléatoires (**théorème limite central**);
- elle mène à des résultats simples dans plusieurs problèmes d'inférence (chapitres 3 à 6).

Loi normale : définition

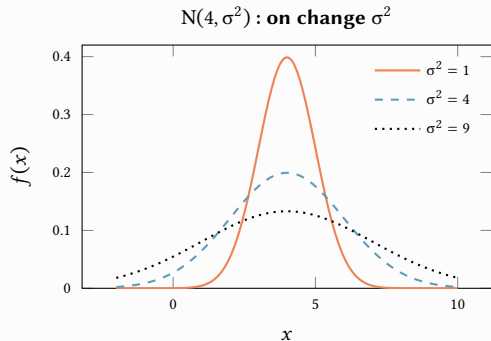
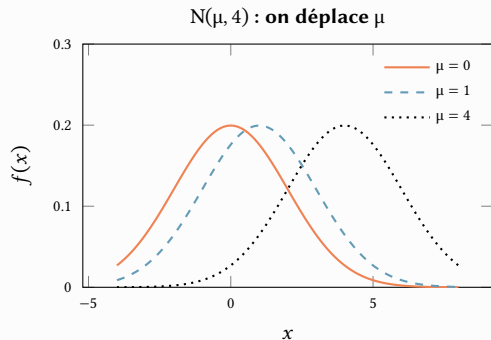
Définition. X suit une **loi normale** d'espérance μ et de variance σ^2 , noté $X \sim N(\mu, \sigma^2)$, si sa densité est

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}, \quad -\infty < x < \infty.$$

Proposition. Si $X \sim N(\mu, \sigma^2)$, alors $E(X) = \mu$ et $\text{Var}(X) = \sigma^2$.

Les deux paramètres de la loi normale sont donc directement son espérance et sa variance.

Allure de la densité



μ est un paramètre de **position** : il déplace la cloche sans la déformer. σ est un paramètre d'**échelle** : plus il est grand, plus la cloche est large et basse (l'aire totale reste égale à 1).

GD9674175N9

Deutsche Bundesbank

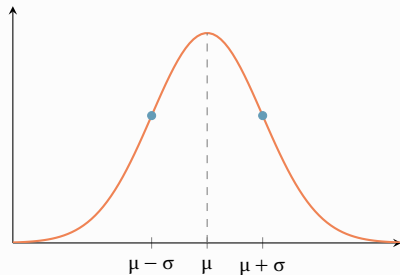
Heinrich *Heine*
Frankfurt am Main
1 Oktober 1993



GD9674175N9

Propriétés de la densité $N(\mu, \sigma^2)$

- $f(x) > 0$ pour tout x et l'aire totale vaut 1 ;
- cloche **symétrique** autour de μ :
 $f(\mu - x) = f(\mu + x)$;
- maximum unique en $x = \mu$;
- deux points d'inflexion, en $\mu - \sigma$ et $\mu + \sigma$.



Mais la fonction de répartition $F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$ **n'a pas de forme explicite** : on utilise un logiciel (R) ou une table.

QUIZ!

Le poids à la naissance X (en g) des bébés nés à terme suit approximativement une loi normale de paramètres $\mu = 3400$ g et $\sigma^2 = 500^2$ g². Sans faire de calcul, répondez :

1. Laquelle est la plus grande : $P(X < 2400)$ ou $P(X > 4400)$?
2. Laquelle est la plus grande : $P(2400 < X < 2500)$ ou $P(3400 < X < 3500)$?
3. Si l'écart type était plutôt de 400 g (même moyenne), $P(X < 2400)$ augmenterait-elle ou diminuerait-elle?

Loi normale standard

Loi normale standard (centrée réduite)

La loi $N(0, 1)$ est appelée **loi normale standard**. Si $Z \sim N(0, 1)$, on note $\Phi(z) = P(Z \leq z)$ sa fonction de répartition.

Par symétrie de la densité autour de 0 :

$$\Phi(-z) = 1 - \Phi(z).$$

Les tables donnent donc $\Phi(z)$ seulement pour $z \geq 0$.

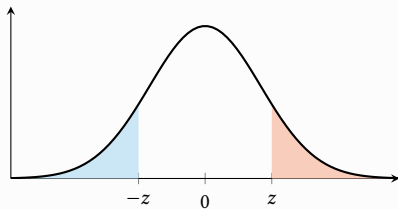


Table de la loi normale standard : $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Table de la loi normale standard (suite)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

$\Phi(z)$	0.80	0.90	0.95	0.975	0.98	0.99	0.995	0.9975	0.999	0.9995
z	0.841	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

QUIZ!

Soit $Z \sim N(0, 1)$. À l'aide de la table, trouvez :

a) $P(Z < 1.15)$

e) $P(|Z| < 2)$

b) $P(Z > 1.15)$

f) la valeur c telle que $P(Z < c) = 0.78$

c) $P(0.4 < Z < 2.12)$

g) la valeur c telle que
 $P(Z > c) = 0.025$

d) $P(Z < -1.4)$

Standardisation

Théorème.

$$X \sim N(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$\Rightarrow$ Une seule table (celle de la $N(0, 1)$) suffit !

$$P(a < X \leq b) = P\left(\frac{a - \mu}{\sigma} < Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Un calcul de probabilité pour une loi normale se fait donc en deux étapes :

1. **standardiser** la v.a. d'intérêt pour se ramener à la $N(0, 1)$;
2. aller chercher les valeurs de $\Phi(\cdot)$ dans la table (ou dans R).

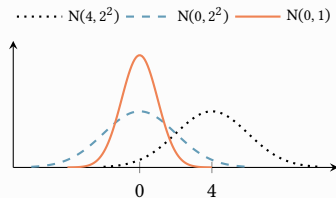
Standardisation : centrer, puis réduire

1. **Centrer** : $X - \mu \sim N(0, \sigma^2)$.

On déplace la cloche pour qu'elle soit centrée en 0.

2. **Réduire** : $\frac{X - \mu}{\sigma} \sim N(0, 1)$.

On change d'unité : la nouvelle unité est l'écart type.



Justification. $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$ est de la forme $aX + b$. Par les propriétés de la 1^{re} partie, $E(Z) = \frac{E(X) - \mu}{\sigma} = 0$ et $\text{Var}(Z) = \frac{\text{Var}(X)}{\sigma^2} = 1$; de plus, une transformation linéaire d'une v.a. normale est encore normale.

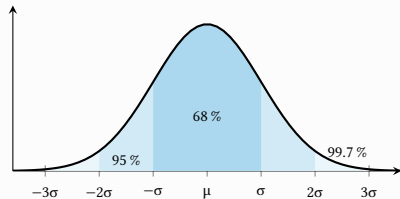
Interprétation. $z = (x - \mu)/\sigma$ est le **nombre d'écarts types** qui séparent x de la moyenne. Ex. : un bébé de 2 650 g a $z = (2650 - 3400)/500 = -1.5$.

Règle empirique : 68 – 95 – 99.7

Règle empirique

Si $X \sim N(\mu, \sigma^2)$, environ

- 68 % des valeurs sont dans $\mu \pm \sigma$;
- 95 % sont dans $\mu \pm 2\sigma$;
- 99.7 % sont dans $\mu \pm 3\sigma$.



Valeurs exactes : $\Phi(1) - \Phi(-1) = 0.6827$, $\Phi(2) - \Phi(-2) = 0.9545$, $\Phi(3) - \Phi(-3) = 0.9973$. Pour exactement 95 %, il faut $\pm 1.96 \sigma$.

Poids à la naissance ($\mu = 3400$ g, $\sigma = 500$ g) : environ 68 % des bébés pèsent entre 2 900 et 3 900 g, 95 % entre 2 400 et 4 400 g, et presque tous (99.7 %) entre 1 900 et 4 900 g. Une valeur à plus de 3 écarts types de la moyenne est donc **très inhabituelle**.

Exemple 4

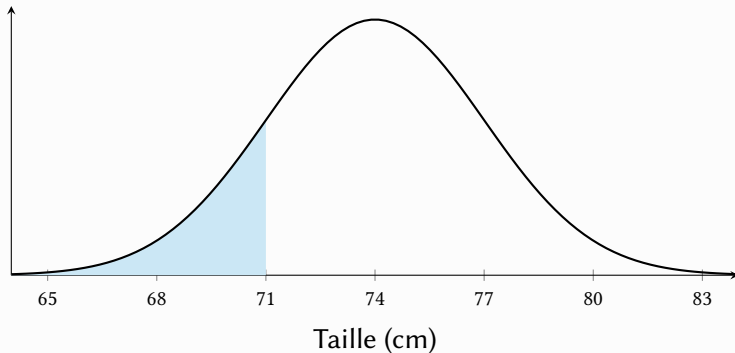
La taille d'une fille de 12 mois (X) suit une distribution normale de moyenne 74 cm et de variance 9 cm^2 : $X \sim N(74, 9)$.

- a) À 12 mois, une fillette mesure 71 cm. Quelle proportion des fillettes sont plus petites qu'elle ?

- b) Quelle est la probabilité qu'une fillette choisie au hasard mesure entre 70 et 75 cm ?

- c) On considère que les 1 % des fillettes les plus petites devraient faire l'objet d'un suivi spécial. À partir de quelle taille les fillettes doivent-elles faire l'objet d'un suivi ?

Exemple 4 : commencer par un croquis



a) $P(X < 71)$: aire à gauche de 71, soit à 1 écart type (σ) sous la moyenne.

Exemple 4 : calculs dans R

- a) $P(X < 71)$

```
> pnorm(71, mean = 74, sd = 3)
[1] 0.1586553
```

- b) $P(70 \leq X \leq 75)$

```
> pnorm(75, mean = 74, sd = 3) - pnorm(70, mean = 74, sd = 3)
[1] 0.5393474
```

- c) Pour quel x^* a-t-on $P(X \leq x^*) = 0.01$?

```
> qnorm(0.01, mean = 74, sd = 3)
[1] 67.02096
```

Attention : dans R, on donne l'écart type `sd`, pas la variance!

QUIZ!

Chez les adultes en santé, la glycémie à jeun X (en mmol/L) suit approximativement une loi $N(5.0, 0.5^2)$.

- Un patient a une glycémie de 6.1 mmol/L. À combien d'écarts types de la moyenne se situe-t-il?
- Quelle proportion des adultes en santé ont une glycémie supérieure à la sienne?
- Sous quelle valeur se trouve la glycémie de 95 % des adultes en santé?

Combinaisons linéaires de v.a. normales

Proposition.

1. Si $X \sim N(\mu, \sigma^2)$ et $Y = aX + b$, alors $Y \sim N(a\mu + b, a^2\sigma^2)$.
2. Si $X_1, \dots, X_n$ sont **indépendantes** avec $X_i \sim N(\mu_i, \sigma_i^2)$, alors

$$a_1X_1 + \dots + a_nX_n \sim N(a_1\mu_1 + \dots + a_n\mu_n, a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2).$$

Cas particuliers importants (X et Y normales **indépendantes**) :

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2), \quad X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Attention! La variance d'une *différence* est une *somme* : $\text{Var}(X - Y) = 1^2\sigma_X^2 + (-1)^2\sigma_Y^2$.
Soustraire deux quantités aléatoires ajoute de l'incertitude, ça ne l'enlève pas. (Utile au chapitre 4 pour comparer deux moyennes.)

Exemple 5

La masse d'une orange est une variable aléatoire **de distribution normale** d'espérance 204 g et de variance 256 g^2 . On crée des sacs de 7 oranges. La masse de l'emballage est une variable normale de moyenne 50 g et de variance 10 g^2 . Toutes ces masses sont indépendantes.

Quelle est la probabilité que la masse d'un sac choisi au hasard soit inférieure à 1400 g?

Rappel : on a calculé l'espérance et la variance d'un sac de 7 oranges à la fin de la 1^{re} partie. Ce qui est nouveau : on connaît maintenant la *loi* de la masse du sac.

Exemple 6

Dans une animalerie, la masse (en g) des souris adultes d'une lignée suit une loi $N(30, 9)$ chez les mâles et une loi $N(26, 7)$ chez les femelles. On choisit au hasard un mâle et une femelle, dont les masses M et F sont indépendantes.

Quelle est la probabilité que la femelle soit plus lourde que le mâle ?

Piste : $P(F > M) = P(F - M > 0)$. Quelle est la loi de $D = F - M$?

Et si les v.a. ne sont pas indépendantes ?

Les formules de variance précédentes supposent l'**indépendance**. Sinon, il faut tenir compte de la **covariance** :

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2 \text{Cov}(X, Y), \quad \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

- $\text{Cov}(X, Y) > 0$: X et Y ont tendance à être grandes (ou petites) *en même temps* (ex. : longueur et masse d'un poisson).
- $\text{Cov}(X, Y) < 0$: quand l'une est grande, l'autre a tendance à être petite.
- La **corrélation** $\rho = \text{Cov}(X, Y)/(\sigma_X \sigma_Y)$ est sans unité et toujours entre -1 et 1 .
- Si X et Y sont indépendantes, alors $\text{Cov}(X, Y) = 0$ (mais la réciproque est fausse).

Dans ce cours, on supposera presque toujours l'indépendance. On retrouvera ces idées avec les données appariées (chapitre 4) et la corrélation échantillonnale (chapitre 6).

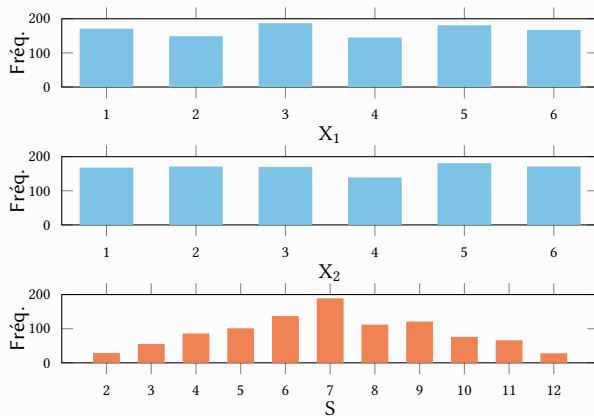
Le théorème limite central

Exemple exploratoire

$\mathcal{E}$: lancer deux dés ($n = 2$) et faire la somme. On répète l'expérience 1000 fois.

Extrait des résultats

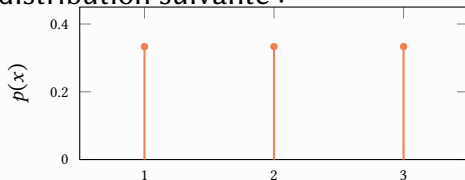
X_1	X_2	$S = X_1 + X_2$
6	4	10
3	2	5
4	1	5
3	1	4
5	6	11
6	6	12
3	5	8
6	1	7
3	5	8
1	3	4
1	5	6
1	1	2



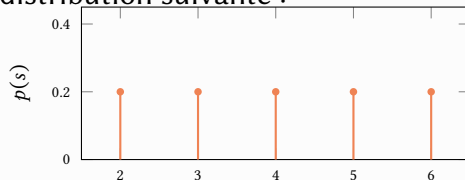
QUIZ!

Une urne contient 3 boules numérotées 1, 2, 3. On pige une boule, on la remet, puis on en pige une autre. Vrai ou faux?

1) La valeur de la première boule a la distribution suivante :



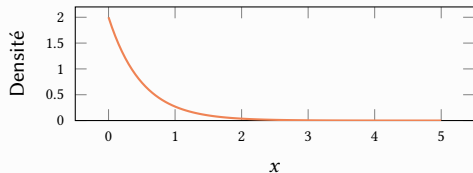
2) La somme des deux numéros a la distribution suivante :



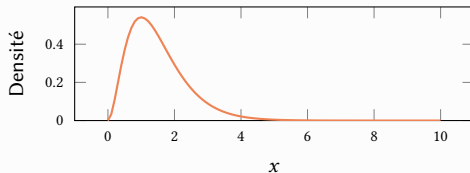
Sommes de variables exponentielles

Même en partant d'une loi très asymétrique, la somme de plusieurs v.a. indépendantes prend une forme de cloche.

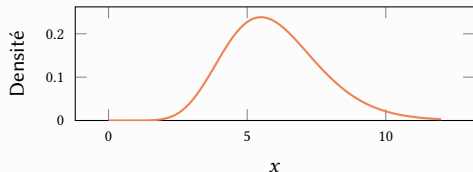
1 exponentielle



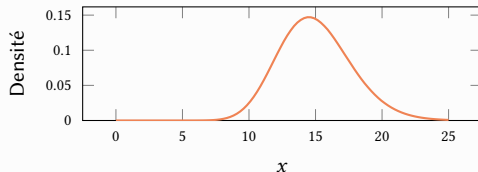
Somme de 3



Somme de 12



Somme de 30



Un résultat très (très) utile

Théorème. Soit $X_1, X_2, \dots, X_n$ des variables aléatoires **indépendantes et identiquement distribuées** de moyenne μ et de variance σ^2 , de loi quelconque. Alors, si n est assez grand,

$$S_n = X_1 + X_2 + \dots + X_n \approx N(n\mu, n\sigma^2) \quad \text{et}$$
$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

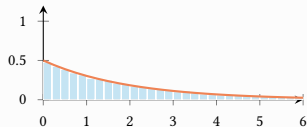
C'est le **théorème limite central** (TLC).

L'espérance et la variance de S_n et de $\bar{X}$ découlent des propriétés vues en 1^{re} partie ; le TLC ajoute que la loi est *à peu près normale*. En pratique, on considère souvent que $n \geq 30$ suffit.

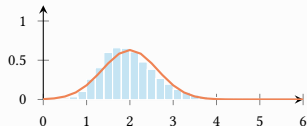
Le TLC en action : simulation

Temps d'attente X (en min) entre deux visites de pollinisateurs sur une fleur : loi exponentielle, très asymétrique, avec $\mu = 2$ et $\sigma = 2$. On simule 10 000 échantillons de taille n et on trace l'histogramme des 10 000 moyennes $\bar{X}$.

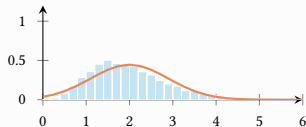
$n = 1$ (écart type 2)



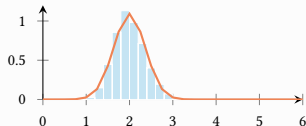
$n = 10$ (écart type 0.63)



$n = 5$ (écart type 0.89)



$n = 30$ (écart type 0.37)



Quand n augmente :

- le **centre** reste à $\mu = 2$;
- l'**étalement** diminue comme $\sigma/\sqrt{n}$;
- la **forme** devient celle d'une cloche.

Courbe : densité de départ ($n = 1$), puis $N(\mu, \sigma^2/n)$ prédite par le TLC.

Conditions d'application du TLC

- Les observations $X_1, \dots, X_n$ sont **indépendantes** (échantillon aléatoire) et de **même loi**.
- n est **assez grand**. Règle du pouce : $n \geq 30$. Mais cela dépend de la loi de départ :
 - à peu près symétrique : n plus petit suffit (somme de 2 dés!);
 - très asymétrique : il faut un n plus grand.
- Si les X_i sont déjà **normales**, $\bar{X}$ est *exactement* normale pour tout n (combinaison linéaire de normales indépendantes) : pas besoin du TLC.

Attention : ce que le TLC ne dit pas

Le TLC porte sur la loi de la **moyenne** (ou de la somme), pas sur celle des observations : peser 1 000 bébés au lieu de 30 ne rend pas les poids individuels « plus normaux ».

QUIZ!

Dans chaque cas, peut-on considérer que $\bar{X}$ suit (approximativement) une loi normale ? Pourquoi ?

- a) $\bar{X}$ = nombre moyen de colonies bactériennes sur 50 boîtes de Petri indépendantes (nombre par boîte $\sim$ Poisson(12)).
- b) $\bar{X}$ = glycémie moyenne de 4 patients choisis au hasard, la glycémie étant de loi normale.
- c) $\bar{X}$ = temps de survie moyen de 6 souris, le temps de survie ayant une loi très asymétrique.
- d) Vrai ou faux : avec $n = 500$ bébés, l'histogramme des 500 poids à la naissance ressemblera à une cloche grâce au TLC.

Exemple 7

On sait que la concentration d'atrazine (un herbicide) dans les étangs d'une région a une moyenne de 4 mg/l et un écart type de 1,5 mg/l. On prend un échantillon d'eau d'un litre dans 40 de ces étangs.

Quelle est la probabilité que la concentration moyenne d'atrazine dans ces échantillons dépasse 4,2 mg/l ?

La loi de la concentration dans un étang n'est pas donnée : c'est le TLC qui nous permet de conclure.

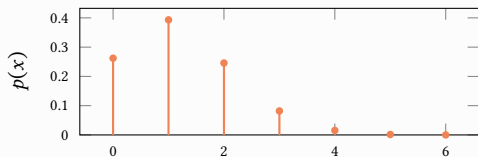
Application : approximation normale de la binomiale

Une v.a. binomiale est une somme de v.a. Bernoulli(p) indépendantes :
 $X = Y_1 + \dots + Y_n$. Par le TLC, si n est grand,

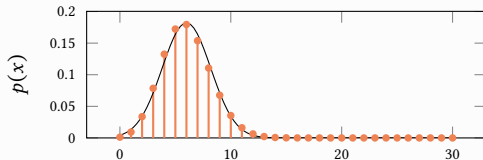
$$\text{Binomiale}(n, p) \approx N(np, np(1 - p)).$$

Règle pratique : l'approximation est bonne si $np \geq 5$ et $n(1 - p) \geq 5$.

Binomiale(6, 0.2)



Binomiale(30, 0.2)

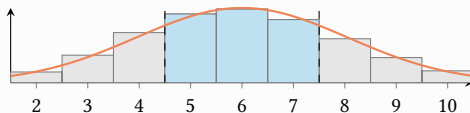


Correction pour la continuité

La binomiale est **discrète**, la normale est **continue** : on améliore l'approximation en élargissant de $1/2$ de chaque côté.

Proposition. Si $X \sim \text{Binomiale}(n, p)$ et n est grand, pour des entiers $j \leq \ell$,

$$P(j \leq X \leq \ell) \approx \Phi\left(\frac{(\ell + \frac{1}{2}) - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{(j - \frac{1}{2}) - np}{\sqrt{np(1-p)}}\right).$$



$P(5 \leq X \leq 7)$: on intègre la densité normale de 4.5 à 7.5.

Exemple 8

Environ 40 % des étudiants de la FSG ont déjà suivi un cours de statistique avant d'entrer à l'université. Sur la cohorte de 400 étudiants admis l'an dernier, quelle est la probabilité qu'entre 150 et 160 étudiants (inclusivement) aient déjà fait un cours de statistique ?

Exemple 8 : vérification dans R

Valeur exacte (binomiale) :

```
> sum(dbinom(150:160, size = 400, prob = 0.4))  
[1] 0.3798861
```

Approximation normale ($np = 160$, $np(1 - p) = 96$), avec correction pour la continuité :

```
> pnorm(160.5, 160, sqrt(96)) - pnorm(149.5, 160, sqrt(96))  
[1] 0.3784114
```

... et sans correction pour la continuité :

```
> pnorm(160, 160, sqrt(96)) - pnorm(150, 160, sqrt(96))  
[1] 0.3462829
```

Avec la correction, l'erreur est d'environ 0.0015; sans elle, d'environ 0.034. La correction vaut la peine!

Réponses aux exemples

Loi de Bernoulli : $E(Y) = p$, $\text{Var}(Y) = p(1 - p)$.

Coefficient binomial : $\binom{5}{2} = 10$; $\binom{49}{6} = 13\,983\,816$.

Exemple 1 : $X \sim \text{Bin}(5, 0.85)$;

$P(X = 4) = 5 \times 0.85^4 \times 0.15 = 0.3915$.

Quiz (fibrose kystique) : $X \sim \text{Bin}(4, 0.25)$; $E(X) = 1$;

$P(X = 0) = 0.75^4 = 0.3164$.

Exemple 2 : a) $P(X > 5) = 0.9556$; b) $np = 6.65$; c)

$\sqrt{7(0.95)(0.05)} = 0.577$.

Exemple 3 : a) $P(X \leq 2) = 0.1247$ ($\lambda = 5$); b)

$P(X = 45) = 0.0458$ ($\lambda = 50$).

Quiz (mutations) : $\text{Bin}(300\,000, 10^{-5}) \approx \text{Poisson}(3)$;

$P(X \geq 1) = 1 - e^{-3} = 0.9502$.

Quiz (binomiale ou Poisson ?) : a) binomiale($20, p$); b)

Poisson; c) binomiale($1500, 10^{-4}$) $\approx$ Poisson(0.15); d) ni

l'une ni l'autre (v.a. continue, plutôt normale); e) ni l'une ni

l'autre (contagion : les porcelets ne sont pas indépendants).

Quiz (poids à la naissance) : 1. égales par symétrie (0.0228 chacune); 2. $P(3400 < X < 3500) = 0.0793$ est plus grande que $P(2400 < X < 2500) = 0.0132$ (plus près de μ); 3. elle diminuerait ($0.0228 \rightarrow 0.0062$).

Quiz (table) : a) 0.8749 ; b) 0.1251 ; c) 0.3276 ; d) 0.0808 ; e) 0.9544 ; f) $c \approx 0.77$; g) $c = 1.96$.

Exemple 4 : a) $\Phi(-1) = 0.1587$; b)

$\Phi(0.33) - \Phi(-1.33) = 0.5375$ avec la table (0.5393 avec R);

c) $74 - 2.326 \times 3 = 67.02$ cm.

Quiz (glycémie) : $z = 2.2$; $P(X > 6.1) = 0.0139$;

$5.0 + 1.645 \times 0.5 = 5.82$ mmol/L.

Exemple 5 : masse du sac $\sim N(1478, 1802)$;

$P(\text{sac} < 1400) = \Phi(-1.84) = 0.0329$ (0.0331 avec R).

Exemple 6 : $D = F - M \sim N(-4, 16)$;

$P(D > 0) = P(Z > 1) = 0.1587$.

Réponses aux exemples (suite)

Quiz (urne) : 1) vrai ; 2) faux : la somme vaut 2, 3, 4, 5, 6 avec probabilités $1/9, 2/9, 3/9, 2/9, 1/9$ (déjà une forme en « triangle »).

Quiz (conditions du TLC) : a) oui, par le TLC ($n = 50$ observations indépendantes); b) oui, et même exactement (moyenne de v.a. normales indépendantes); c) non : $n = 6$ est trop petit pour une loi très asymétrique; d) faux : le TLC porte sur la loi de la *moyenne*, pas sur celle des observations individuelles.

Exemple 7 : $\bar{X} \approx N(4, 1.5^2/40)$; $P(\bar{X} > 4.2) = P(Z > 0.84) = 0.2005$ avec la table (0.1995 avec R).

Exemple 8 : valeur exacte 0.3799; approximation normale avec correction pour la continuité 0.3784 (sans correction : 0.3463).

Références

Ce document est rédigé avec Beamer ainsi qu'une classe .cls fournie par Jérôme Soucy.

Les diapositives sont adaptées des diapositives créées par Emmanuelle Reny-Nolin pour le cours STT-1920 ainsi que de celles des cours STT-1000 et STT-1900.

Référence : C. Bélisle, *STT-1920 Méthodes statistiques*, notes de cours, Université Laval, 2020, section 2.6.