Cours de biostatistique ⋈ Illustrations dans 

Vraisemblance d'une hypothèse

A.B. Dufour & D. Chessel

23 février 2017

La fiche définit quelques éléments de base du raisonnement statistique : événements, variables aléatoires, échantillons, hypothèses, intervalles de confiance.

Table des matières

1 Espace d'événements

Les notions essentielles sont décrites sur des exemples.
Enumérer les résultats possibles d'une expérience. Jouer à pile ou face. Lancer un dé. Lancer deux dés. Placer trois objets dans trois cases. Tirer 7 boules dans une urne de 49 boules. Placer 7 boules dans 49 boîtes n'en pouvant contenir qu'une.

Et par généralisation,
Placer r boules dans n boîtes. L'anniversaire de n personnes sur 365 jours. Le résultat du jet de r dés à 6 faces. Le type de r individus pour un locus à n génotypes. r fautes d'orthographe dans n pages. Le sexe, la catégorie socio-professionnelle de r personnes. r arbres de n espèces.

Boules indistinctes.
Ensembles discrets fini ou dénombrable.
Jouer au loto.

1.1 Bases

Evénement élémentaire $\omega \in \Omega$. Événement $A \subseteq \Omega$, $\{\omega\} \subseteq \Omega$

Espace des événements $\mathrm{P}(\Omega)$.
Événement impossible $\varnothing \in \mathrm{P}(\Omega)$

Événements incompatibles $A \cap B = \varnothing$.

Système complet d'événements $[i \neq j \Rightarrow A_i \cap A_j = \varnothing]$, $\bigcup_{i=1}^{n} A_i = \Omega$

Espace probabilisé $P : \mathrm{P}(\Omega) \to [0,1]$ avec :
$P(\Omega) = 1$
$[i \neq j \Rightarrow A_i \cap A_j = \varnothing] \Rightarrow P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

Propriétés élémentaires :
$P(\varnothing) = 0 \quad ; \quad P(\overline{A}) = 1 - P(A) \quad ; \quad A \subseteq B \Rightarrow P(A) \leq P(B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Si $\{B_i\}_{i=1}^{n}$ est un système complet alors $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$

Probabilité conditionnelle

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Evénements indépendants : A est indépendant de B si $P(A/B) = P(A)$

1.2 Formule de Bayes

$$\boxed{P(B_k/A) = \frac{P(B_k)\ P(A/B_k)}{\sum_{i=1}^{n} P(B_i)\ P(A/B_i)}} \tag{1}$$

http ://www.stpi.com/loto/stplot01.html
Résultats du mois de Janvier 1999

Date	No_année	No	Ordre_de_sortie	Ordre_numérique
02-janv	53c	2825	14-28-25-02-12-15-18	02-12-14-15-25-28-18
02-janv	53d	2826	16-27-02-42-32-05-19	02-05-16-27-32-42-19
06-janv	01a	2827	45-30-22-10-28-32-02	10-22-28-30-32-45-02
06-janv	01b	2828	19-28-33-03-10-42-25	03-10-19-28-33-42-25
09-janv	01c	2829	15-44-16-38-03-23-34	03-15-16-23-38-44-34
09-janv	01d	2830	25-14-41-35-23-13-16	13-14-23-25-35-41-16
13-janv	02a	2831	06-42-13-49-43-32-15	06-13-32-42-43-49-15
13-janv	02b	2832	08-22-11-06-49-43-35	06-08-11-22-43-49-35
16-janv	02c	2833	15-42-37-30-47-19-14	15-19-30-37-42-47-14
16-janv	02d	2834	19-48-34-21-15-40-13	15-19-21-34-40-48-13
20-janv	03a	2835	03-22-13-45-25-07-29	03-07-13-22-25-45-29
20-janv	03b	2836	34-04-46-36-03-07-02	03-04-07-34-36-46-02
23-janv	03c	2837	14-09-01-49-30-03-15	01-03-09-14-30-49-15
23-janv	03d	2838	35-04-29-18-01-28-08	01-04-18-28-29-35-08
27-janv	04a	2839	32-42-38-03-34-27-33	03-27-32-34-38-42-33
27-janv	04b	2840	40-01-38-02-47-04-39	01-02-04-38-40-47-39
30-janv	04c	2841	01-33-41-08-39-38-18	01-08-33-38-39-41-18
30-janv	04d	2842	19-15-33-16-07-11-06	19-15-33-16-07-11-06

...

Quelques événements élémentaires dans l'espace probabilisé associé au jeu du loto (tirage au hasard sans remise de 7 boules dans une urne qui en contient 49). L'espace est connu et il n'y a rien à apprendre de ces résultats. Ceci n'est pas une statistique expérimentale destinée à acquérir une information nouvelle.

1.3 Espaces d'équiprobabilité : exemples

Pièce de Monnaie : $\Omega = \{pile, face\}$
$P(\varnothing) = 0 \quad P(\{pile\}) = \frac{1}{2} \quad P(\{face\}) = \frac{1}{2} \quad P(\{pile, face\}) = 1$

Dé à six faces : $\Omega = \{1, 2, 3, 4, 5, 6\} \quad P(\{i\}) = \frac{1}{6} \quad P(\{2, 3, 6\}) = \frac{1}{2}$

Espace équiprobable à n événements élémentaires :

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\} \quad P(\{\omega_j\}) = \frac{1}{n} \quad P(\{\omega_{j_1}, \omega_{j_2}, \cdots, \omega_{j_k}\}) = \frac{k}{n}$$

Espace [4!] des permutations de {1,2,3,4}. Il y en a 24 :

1234	1243	1324	1342	1423	1432
2134	2143	2314	2341	2413	2431
3124	3142	3214	3241	3412	3421
4123	4132	4213	4231	4312	4321

Espace $\left[4^2\right]$ des choix de 2 objets parmi 4 *avec remise* (ou des placements de 2 objets dans 4 boîtes). Il y a 16 événements élémentaires équiprobables.

aa	ab	ac	ad
ba	bb	bc	bd
ca	cb	cc	cd
da	db	dc	dd

12			
2	1		
2		1	
2			1

1	2		
	12		
	2	1	
	2		1

1		2	
	1	2	
		12	
		2	1

1			2
	1		2
		1	2
			12

Espace $[(6)_2]$ des choix de 2 objets sur 6 *sans remise*.

12	13	14	15	16
21	23	24	25	26
31	32	34	35	36
41	42	43	45	46
51	52	53	54	56
61	62	63	64	65

Espace $\left[\begin{pmatrix} 5 \\ 3 \end{pmatrix}\right]$ des choix de 3 positions sur 5 places. Il y a 10 événements élémentaires de probabilité $\frac{1}{10}$.

x	x	x		
x	x		x	
x	x			x
x		x	x	
x		x		x
x			x	x
	x	x	x	
	x	x		x
	x		x	x
		x	x	x

Plus généralement :
Il y a $n!$ écritures ordonnées des n premiers entiers (permutations) :

$$n! = 1 \times 2 \times 3 \times \cdots \times (n-1) \times n$$

Il y a 3 628 800 manières de ranger 10 objets.

Il y a n^r choix de r objets parmi n avec remise (mots).

$$n^r = \underbrace{n \times n \times \cdots \times n}_{r \ fois}$$

Il y a 1 048 576 de suites de longueur 10 avec 4 signes ACGT du genre ACGTTCGCGT.

Il y a $(n)_r$ choix de r objets parmi n sans remise (sr-sélections) :

$$\boxed{(n)_r = n\,(n-1)\,(n-2)\cdots(n-r+1)}$$

Il y a 432 938 943 360 tirages possibles au loto (7 boules sur 49).

Il y a $\begin{pmatrix} n \\ r \end{pmatrix}$ sous-populations de taille r dans une population de taille n (combinaisons) :

$$\boxed{\begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!\,(n-r)!} = \frac{(n)_r}{r!} = \frac{n\,(n-1)\,(n-2)\cdots(n-r+1)}{r\,(r-1)\cdots 3\times 2\times 1}}$$

Il y a 13 983 816 grilles de loto (6 cases sur 49).

Important (convention) :

$$\boxed{1^0 = 0! = (n)_0 = \begin{pmatrix} n \\ 0 \end{pmatrix} = 1}$$

Dans chaque cas, on définit un espace probabilisé Ω en affectant à chaque événement la probabilité

$$P(\omega) = \frac{1}{Card(\Omega)} \quad A \in \mathrm{P}(\Omega) \Rightarrow P(A) = \frac{Card(A)}{Card(\Omega)}$$

On note ces espaces : $[n!]$; $[n^r]$; $[(n)_r]$; $\left[\begin{pmatrix} n \\ r \end{pmatrix}\right]$.

1.4 Variables aléatoires

(Ω, P) est un espace probabilisé. Ψ est un ensemble.
$X : \Omega \to \Psi$ est une application. $B \in \mathrm{P}(\Psi)$ est une partie de Ψ.
$X^{-1}(B) = \{\omega \in \Omega / X(\omega) \in B\}$ est une partie de Ω.

La règle $Q(B) = P\left(X^{-1}(B)\right)$ définit un nouvel espace probabilisé (Ψ, Q).

Exemple. On jette deux dés.
$\Omega = \{(1,1),(1,2),...,(1,6),(2,1),...,(2,6),...,(6,1),...,(6,6)\}$.
Chaque événement a une probabilité de 1/36.
$\Psi = \{2,...,12\}$.
X associe à chaque événement la somme des résultats *e.g.* $X((2,6)) = 8$.

$$X^{-1}(6) = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} \Rightarrow Q(6) = 5/36$$

Le nouvel espace probabilisé est défini par la probabilité des événements élémentaires :

	$Q(4) = 3/36$	$Q(7) = 6/36$	$Q(10) = 3/36$
$Q(2) = 1/36$	$Q(5) = 4/36$	$Q(8) = 5/36$	$Q(11) = 2/36$
$Q(3) = 2/36$	$Q(6) = 5/36$	$Q(9) = 4/36$	$Q(12) = 1/36$

$$B \in \mathrm{P}(\Psi) \Rightarrow Q(B) = \sum_{\psi \in B} Q(\psi)$$

X est appelée variable aléatoire.
L'espace (Ψ, Q) est appelé loi de probabilité.

1.5 Espaces des échantillons aléatoires simples

Une autre manière simple de générer des espaces probabilisés utilise le produit. $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ est un espace d'éventualités probabilisé par $P(\omega_j) = p_j$.

$$p_j \geq 0, \quad \sum_{j=1}^{n} p_j = 1$$

$\Omega^2 = \{(\omega_j, \omega_k)\}_{1 \leq j \leq n}^{1 \leq k \leq n}$ est l'ensemble des couples d'éventualités. Il est probabilisé par la loi de deux tirages aléatoires simples $P_2((\omega_j, \omega_k)) = p_j p_k$.

$\Omega^r = \{(\omega_{j_1}, \omega_{j_2}, ..., \omega_{j_r})\}_{1 \leq j_k \leq n}$ est l'ensemble des r échantillons aléatoires simples dans (Ω, P). Il est probabilisé par $P_r((\omega_{j_1}, \omega_{j_2}, ..., \omega_{j_r})) = p_{j_1} p_{j_2} ... p_{j_r}$.

Exemples.
Le jeu de Pile ou Face avec une pièce parfaite est modélisé par $\Omega = \{0, 1\}$ et $P(0) = P(1) = \frac{1}{2}$.

Quand on joue $r = 10$ fois, $\Omega^{10} = \{a_1 a_2 ... a_{10}\}$ est l'ensemble des mots de longueur 10 écrits avec 0 et 1. Chaque mot est équiprobable ($1/2^{10} = 0.0009765625$).

$\Omega = \{0, 1\}$ peut être probabilisé par $P(1) = p, \quad P(0) = 1 - p = q$ (p est la probabilité du succès, q est la probabilité de l'échec). (Ω, P) est appelée loi de Bernoulli (1654-1705).

Chaque mot de longueur n (dans Ω^n) comportant k succès et $n - k$ échecs a la même probabilité $p^k q^{n-k}$. La variable aléatoire X définie par le nombre de succès prend ses valeurs dans {0,1,...,n}. Comme il y a $\binom{n}{k}$ mots de ce type, on a une nouvelle loi définie par $\Psi = \{0,1,...,n\}$ et $Q(k) = \binom{n}{k} p^k q^{n-k}$. (Ψ, Q) est appelée loi binomiale.

L'ensemble des espaces d'équiprobabilité, de leurs puissances et des variables aléatoires associées génère une infinité de cas particuliers. Mais les principes communs sont simples.

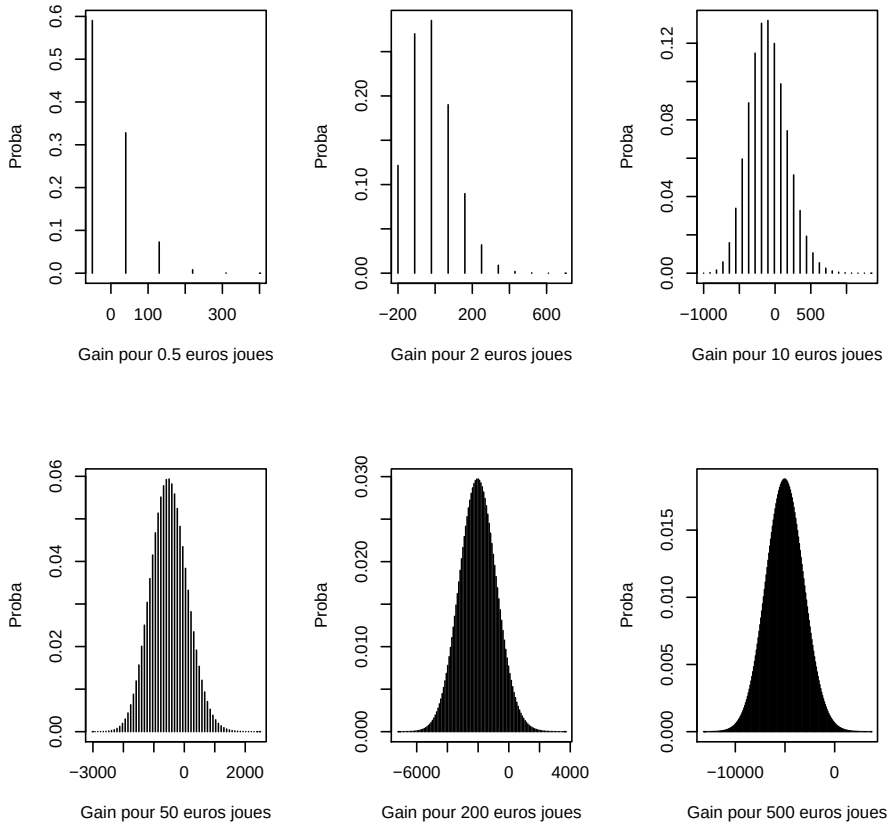
2 Variables aléatoires

Jusqu'à présent, les exemples énumèrent les possibles. Par exemple, un joueur va au casino avec n pièces de 0.10 €. Une machine prend une pièce. 1 fois sur 10, elle en rejette 9 (gain 8) et 9 fois sur 10 elle ne rend rien (perte 1 = gain -1). Le joueur décide de jouer exactement n fois et de sortir. Quelle est la loi de probabilité qui régira son gain ? Pour $n = 5$, en utilisant ce qui précède, on peut dire que la variable aléatoire X nombre de succès suit la loi :

X	0	1	2	3	4	5
Gain	-0.50	0.40	1.30	2.20	3.10	4.00
$P(X)$	$1 \times 0.1^0 \times 0.9^5$	$5 \times 0.1^1 \times 0.9^4$	$10 \times 0.1^2 \times 0.9^3$	$10 \times 0.1^3 \times 0.9^2$	$5 \times 0.1^4 \times 0.9^1$	$1 \times 0.1^5 \times 0.9^0$
$P(X)$	0.59049	0.32805	0.07290	0.00810	0.00045	0.00001

Et pour $n = 20, 100, 500, 2000, 10\,000$?

```
g <- function(n) {
    y <- diff(c(0, pbinom(0:n,n,0.1)))
    x <- (0:n)*80-(n-0:n)*10
    l0 <- paste("Gain pour", 0.10*n,"euros joues")
plot(x[y>=0.000001], y[y>=0.000001], xlab=l0, ylab="Proba", type="h")
}
par(mfrow=c(2,3))
g(5);g(20);g(100);g(500);g(2000);g(5000)
```



2.1 Moyenne et variance d'une variable discrète

$(\Phi = \{\varphi_1, ..., \varphi_n\}, P)$ est une loi de probabilité. Les φ_i (valeurs d'une variable aléatoire) sont des nombres. L'espérance ou moyenne de la loi est :

$$\boxed{\mu = \sum_{i=1}^{n} \varphi_i P(\varphi_i)}$$

Sa variance est :

$$\boxed{\sigma^2 = \sum_{i=1}^{n} (\varphi_i - \mu)^2 P(\varphi_i)}$$

Exemple. Une loi binomiale est définie par :

$$\Phi = \{0, 1, ..., n\} \text{ et } P(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \quad 0 \leq k \leq n$$

On démontre que $\mu = np$ et $\sigma^2 = npq$

La loi binomiale normalisée est définie par :

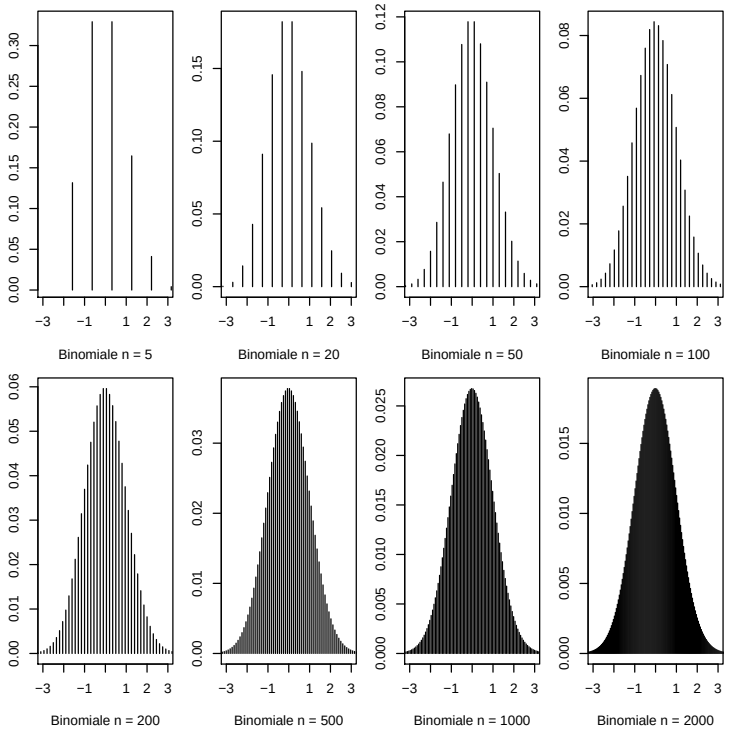
$$\Phi^\bullet = \left\{ \varphi_0^\bullet = \frac{0-\mu}{\sigma}, \varphi_1^\bullet = \frac{1-\mu}{\sigma}, ..., \varphi_n^\bullet = \frac{n-\mu}{\sigma} \right\}$$

et $P(\varphi_k^\bullet) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \quad 0 \leq k \leq n$

L'espérance vaut 0 et la variance vaut 1.

Faisons alors l'expérience suivante. Il s'agit de comparer les binomiales normalisées pour $p = 1/3$ et $n = 5$, 20, 50, 100, 200, 500, 1000, 5000.

```
g <- function(n,p) {
     x <- ((0:n)-n*p)/sqrt(n*p*(1-p))
     l0 <- paste("Binomiale n =",n)
     plot(x, dbinom(0:n,n,p), xlab = l0, ylab = "Proba", type = "h", xlim = c(-3,3))
}
```



Le nombre de valeurs possibles tend vers l'infini, chacune des probabilités tend vers 0, mais ce qui devient constant c'est la probabilité d'être dans un intervalle :

$$P\left([\alpha,\beta[\right) = \sum_{\alpha \leq \varphi_i < \beta} P\left(\varphi_i\right)$$

Il y a dans un intervalle, de plus en plus d'événements mais la somme de leur probabilité tend vers une quantité fixée :

$$P\left([\alpha,\beta[\right) \underset{n\to\infty}{\longrightarrow} \int_\alpha^\beta \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

La forme de la distribution se stabilise. Pour rendre compte de ce phénomène, il faut utiliser la fonction de répartition.

2.2 Fonctions de répartition

$(\Phi = \{\varphi_1, ..., \varphi_n\}, P)$ est une loi de probabilité. Les φ_i (valeurs d'une variable aléatoire) sont des nombres. Les $\varphi_{(i)}$ sont ces nombres rangés par ordre croissant :

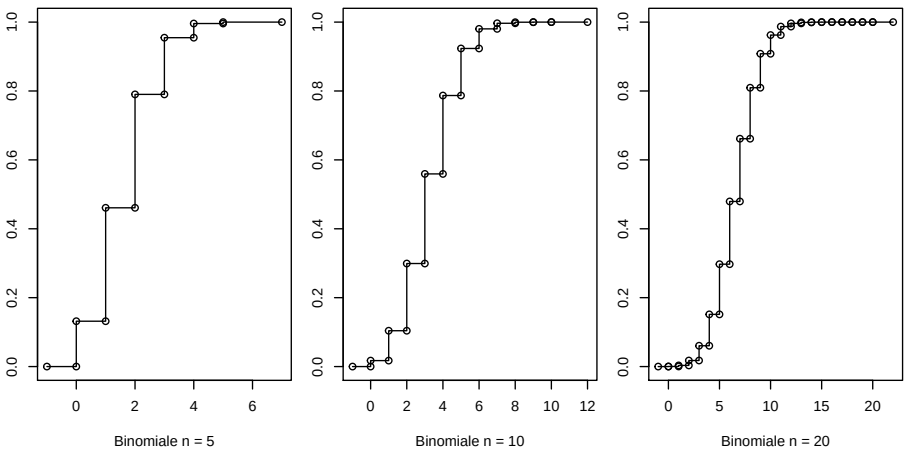
$$\varphi_{(1)} \leq \varphi_{(2)} \leq ... \leq \varphi_{(n)}$$

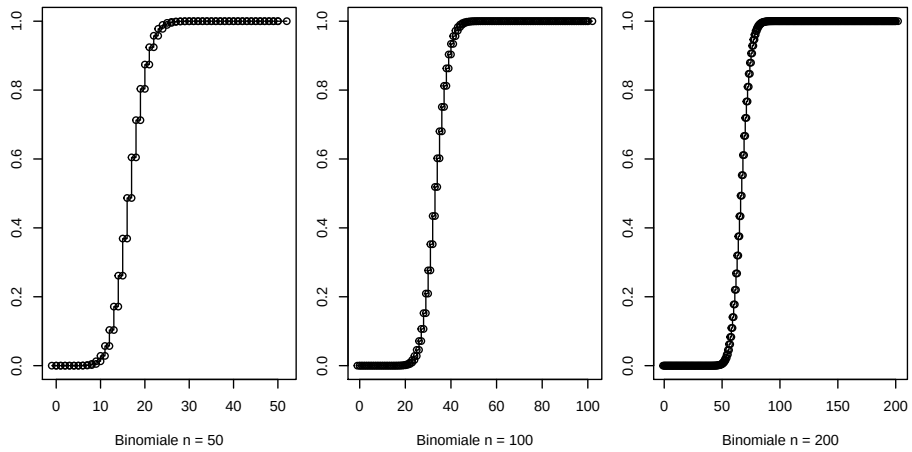
La fonction de répartition de la loi est définie par $F_P : \rightarrow [0,1]$ et :

$$\begin{array}{l} x < \varphi_{(1)} \Rightarrow F_P\left(x\right) = 0 \\ \varphi_{(1)} \leq x < \varphi_{(2)} \Rightarrow F_P\left(x\right) = P\left(\varphi_{(1)}\right) \\ \varphi_{(2)} \leq x < \varphi_{(3)} \Rightarrow F_P\left(x\right) = P\left(\varphi_{(1)}\right) + P\left(\varphi_{(2)}\right) \\ ... \\ \varphi_{(n-1)} \leq x < \varphi_{(n)} \Rightarrow F_P\left(x\right) = P\left(\varphi_{(1)}\right) + P\left(\varphi_{(2)}\right) + ... + P\left(\varphi_{(n-1)}\right) \\ \varphi_{(n)} \leq x \Rightarrow F_P\left(x\right) = 1 \end{array}$$

Les fonctions de répartition de la loi binomiale $\mathcal{B}\left(n, p = 1/3\right)$ sont :

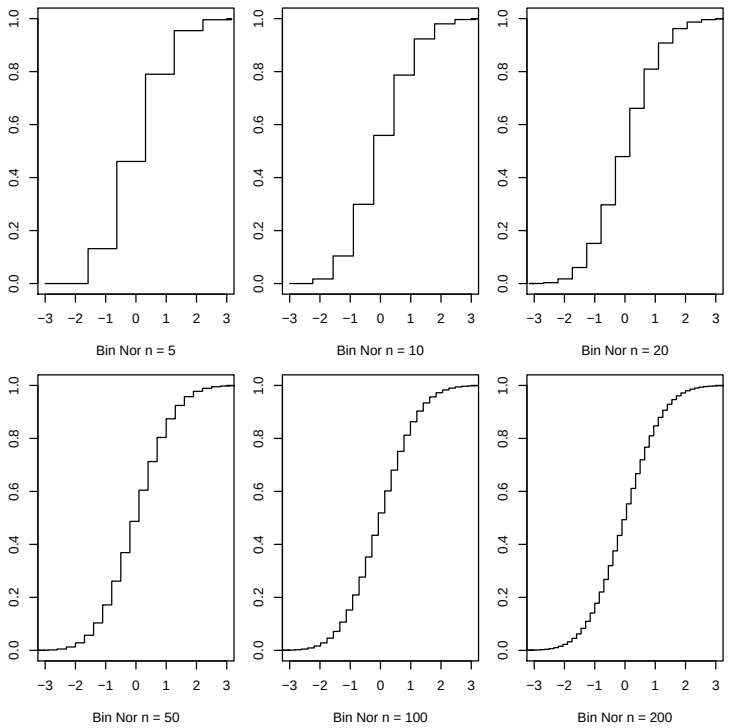
```
g <- function(n,p) {
    x <- c(-1,rep(0:n,rep(2,n+1) ),n+2)
    z <- rep(c(-1,0:n),rep(2,n+2) )
    y <- pbinom(z,n,p)
    l0 <- paste("Binomiale n =",n)
plot(x,y,type="o",xlab=l0)
}
```

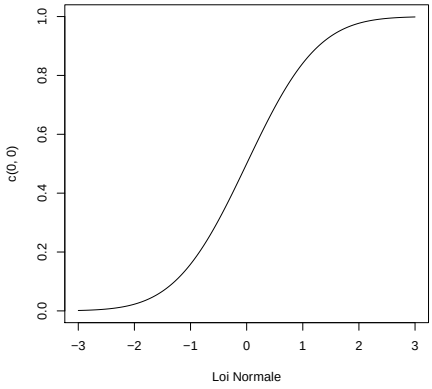




Les fonctions de répartition de la loi binomiale normalisée $\mathcal{B}(n, p = 1/3)$ sont :

```
g <- function(n,p) {
    mu <- n*p ; sigma <- sqrt(n*p*(1-p))
    w0 <- (0:n-mu)/sigma
    x <- c(-3,rep(w0,rep(2,n+1)),3)
    z <- rep(c(-1,0:n),rep(2,n+2) )
    y <- pbinom(z,n,p)
    l0 <- paste("Bin Nor n =",n)
    plot(c(0,0),xlim=c(-3,3),ylim=c(0,1), type="n",xlab=l0)
    lines(x,y)
    return (cbind(x,y))
}
```





Cette dernière courbe est celle de la loi normale :

$$F_{\mathcal{N}}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

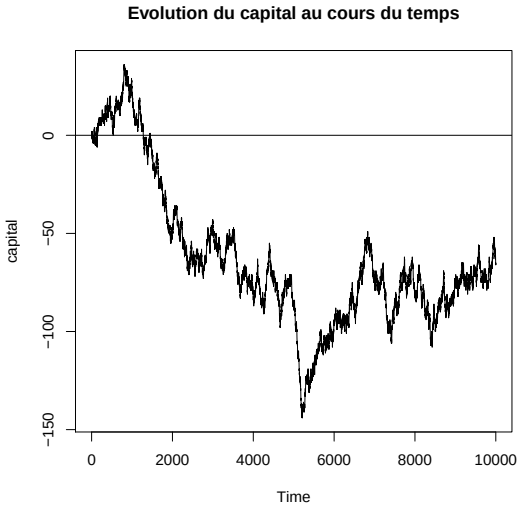
On appelle ce phénomène la convergence en loi de la loi binomiale normalisée $B_{nor}(n, p)$ vers la loi normale quand $n \to \infty$ et p est constante. La fonction F_N ou fonction de répartition de la loi normale est éditée dans tous les livres de statistique. On s'en servira souvent. L'expérience ci-dessus est un cas particulier.

2.3 Le théorème central limite

Jouons à Pile ou Face. Pile, je gagne 1 €. Face, je perd 1 €. Je regarde évoluer mon capital.

Gain X_i	1 (pile)	-1 (face)
$P(X_i)$	$\frac{1}{2}$	$\frac{1}{2}$

Au départ j'ai 0 € sur mon compte. Il y a des bons et des mauvais moments.



n, le nombre de coups, peut être aussi grand qu'on veut. On dit que $n \to +\infty$. A chaque coup, on utilise une variable aléatoire X_i et une famille dénombrable de variables aléatoires.
La moyenne de X_i vaut : $\mu_{X_i} = +1\left(\frac{1}{2}\right) - 1\left(\frac{1}{2}\right) = 0$
La variance de X_i vaut $\sigma^2_{X_i} = +1^2\left(\frac{1}{2}\right) + 1^2\left(\frac{1}{2}\right) = 1$
Le gain cumulé est la variable aléatoire :

$$Y_n = X_1 + ... + X_i + ... + X_n$$

On a :

$$\mu_{Y_n} = \mu_{X_1} + ... + \mu_{X_n} = 0$$

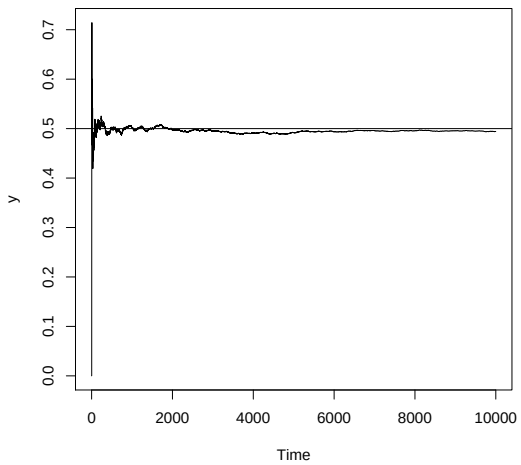
et parce que les X_i sont indépendantes

$$\sigma^2_{Y_n} = \sigma^2_{X_1} + ... + \sigma^2_{X_n} = n.$$

Le théorème central limite dit qu'en général, quand n est assez grand,
Y_n suit une loi Normale de moyenne $\mu_{Y_n} = \mu_{X_1} + ... + \mu_{X_n}$ et de variance $\sigma^2_{Y_n} = \sigma^2_{X_1} + ... + \sigma^2_{X_n}$.
Nous savons que dès $n \geq 30$, l'approximation est bonne.
Pour $n = 10000$, j'aurai gagné en moyenne 0 avec un écart-type de 100 soit presque sûrement entre -200 € et +200 €, pour $n = 1000000$ entre -2000 € et + 2000 €, ... La variance grandit sans arrêt.
On démontre qu'en attendant suffisamment, on passera toujours par une perte excédant le capital initial du joueur.

Mais, si on considère la variable aléatoire "Fréquence de piles" $Z_n = \frac{T_1 + ... + T_i + ... + T_n}{n}$ où T_i est l'obtention de *pile* au lancer i. Ses paramètres sont les suivants :

$$\mu_{Z_n} = \frac{\mu_{T_1} + ... + \mu_{T_n}}{n} = \frac{1}{2} \quad \sigma^2_{Z_n} = \frac{\sigma^2_{T_1} + ... + \sigma^2_{T_n}}{n^2} = \frac{1}{4n}$$



La variance tend vers 0 et la fréquence tend vers la probabilité. On retiendra simplement que, si X_i et une famille de variables aléatoires indépendantes suivant toutes la même loi de moyenne μ_X et de variance σ_X^2, les lois de :

$$S_n = \frac{(X_1 + ... + X_n) - n\mu_X}{\sqrt{n}\sigma_X} \rightsquigarrow \mathcal{N}(0,1)$$

et

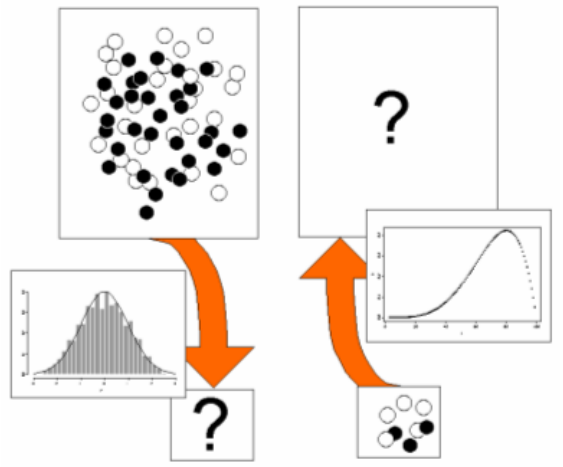
$$\bar{X}_n = \frac{\frac{X_1+...+X_n}{n} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1)$$

Ce théorème fondamental indique qu'on trouvera souvent la loi normale dans la pratique de la statistique pour des raisons de fond.

> D'une façon générale, toute somme de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) tend vers une variable aléatoire gaussienne.

3 Echantillons

Jusqu'à présent la situation était la suivante. De la connaissance d'un espace probabilisé, on peut déduire ce qui va se passer sur un échantillon, du genre si on tire une boule au hasard dans une urne contenant 7 rouges et 3 bleues, 7 fois sur 10 on aura une rouge ! On va maintenant inverser totalement la question : si une urne contient 10 boules et qu'en tirant au hasard on obtient 3 rouges qu'est-ce qu'on peut dire des autres ? C'est ce qu'on appelle l'inférence statistique. Le premier point de vue est celui des mathématiques, le second est celui des sciences expérimentales. Les deux sont très liés.



3.1 Vraisemblance d'une hypothèse

La base du raisonnement est dans la fonction de vraisemblance. Prenons un exemple. Il y a dans l'amphi 100 personnes et l'enseignant X veut savoir combien d'étudiants ont une opinion favorable de ce qu'il raconte. Le plus efficace

est de les interroger tous un par un. Mais c'est long et pénible (Remarque : c'est encore plus long et pénible si on veut savoir combien d'individus dans une population de 100 ours blancs possèdent ou non le gène Z, ont mangé ou non du phoque, souffrent ou non de la maladie A). Alors X en prend 5 au hasard et pose la question " Pensez vous que la statistique est intéressante ? ". C'est OUI 4 fois et NON 1 fois. On peut dire que cet échantillon représente tout à fait la réalité et que 80% des étudiants s'intéressent à la statistique. On peut dire aussi bien que X a eu beaucoup de chance et que la petite minorité de ceux qui s'intéressent à la statistique est sur-représentée.

Il y a exactement 101 hypothèses *a priori* : dans l'amphi, il y a m étudiants qui répondent OUI à la question. Ces hypothèses notées $H_0, H_1, ..., H_{100}$ correspondent aux valeurs 0, 1, 2, ..., 100, valeurs que peut prendre l'inconnue m (*a posteriori*, après l'échantillonnage, on peut éliminer *certainement* les cas 0, 1, 2, 3 et 100) mais m est inconnu et peut prendre *a priori* les valeurs 0 à 100. Nous allons étudier ces 101 hypothèses.

Soit, à titre d'exemple, l'hypothèse H_{30}. Si m vaut 30, en tirant au hasard 5 étudiants sur 100 on aura l'observation 0, 1, 2, 3, 4 ou 5 avec les probabilités $P(0), P(1), ..., P(5)$. Il y a $\begin{pmatrix} 100 \\ 5 \end{pmatrix}$ façons de choisir 5 étudiants et :

$$P(j) = \frac{\begin{pmatrix} 30 \\ j \end{pmatrix}\begin{pmatrix} 70 \\ 5-j \end{pmatrix}}{\begin{pmatrix} 100 \\ 5 \end{pmatrix}}$$

La probabilité de trouver 4 (le résultat observé) est alors 0.02548.
On peut programmer la fonction sous R.

```
proba <- function(j, m = 30, n = 100, r = 5) {
  w0 <- 1
  if (j>0) { for(i in 1:j) w0 <- w0 * (m - i + 1)/i} # j parmi m
  if (j<r) { for(i in 1:(r - j)) w0 <- w0 * (n - m - i + 1)/i} # r-j parmi n-m
  for(i in 1:r) w0 <- w0*i/(n - i + 1) # r parmi n
  return(w0)
}
proba(4)
```

```
[1] 0.02548032
```

On peut aussi noter que c'est une loi hypergéométrique déjà programmée (`help.search("dhyper")`).

```
dhyper(4,30,70,5)
```

```
[1] 0.02548032
```

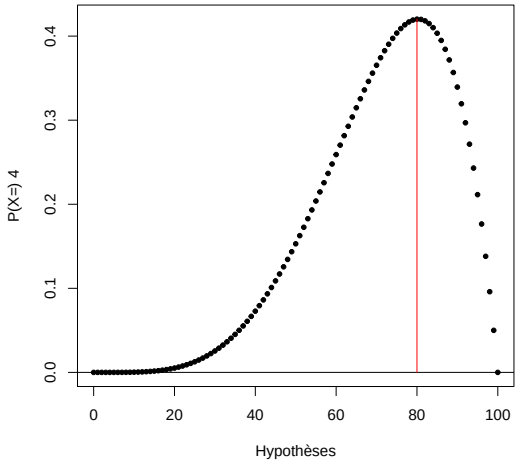
La probabilité d'observer le résultat sous une hypothèse *H* arbitraire est la vraisemblance de cette hypothèse pour cette observation.

Nous pouvons tracer la valeur de la vraisemblance en fonction de l'hypothèse. Ce n'est pas une loi de probabilité mais une fonction dont les valeurs sont des probabilités pour des lois différentes :

```
tracevraisemblance <- function(j) {
   hypothese <- 0:100
```

```
    valproba <- dhyper(j,hypothese,100-hypothese,5)
    plot(hypothese, valproba, xlab="Hypothèses", ylab=paste("P(X=)",j), pch=20)
    abline(h=0)
    hypossvrai <- hypothese[valproba == max(valproba)]
    print(hypossvrai)
    segments(hypossvrai,0,hypossvrai,dhyper(j,hypossvrai,100-hypossvrai,5),col="red")
}
tracevraisemblance(4)
```

```
[1] 80
```



Estimer, c'est choisir une des hypothèses. Estimer au maximum de vraisemblance, c'est choisir l'hypothèse qui donne à l'observation la plus grande vraisemblance.

On trouve ici 80. C'est le plus vraisemblable. On notera la vraisemblance par :

$$L(H) = P_{H\ vraie}(\text{Observation})$$

L renvoie à Likelihood inventé par Sir Ronald Fisher (1890-1962). On parle de statistique fishérienne.

3.2 Rejet d'une hypothèse

On a choisi une hypothèse : elle a toutes les chances d'être fausse ! 75, 78 ,81 ou 83 sont presque aussi vraisemblables. Par contre 30 l'est beaucoup moins. Quand peut-on dire que l'hypothèse rend aberrant un résultat ? Peut-on être sûr que H_{30} n'est pas acceptable ?

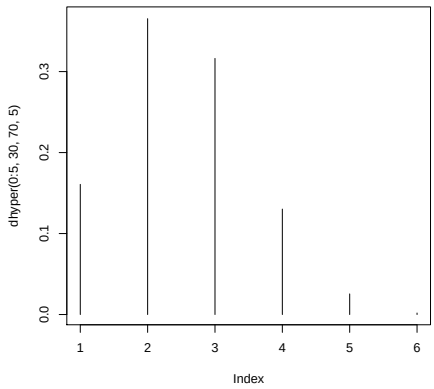
Tester une hypothèse, c'est décider au vu du résultat si elle est vraie ou si elle est fausse.

Décider si elle est vraie est impossible. X ne pourra jamais savoir si il y a 80 ou 79 ou 81 personnes ayant une opinion positive (sauf à les interroger toutes mais penser aux ours blancs). X est déjà certain qu'il y en a au moins 4 et pas

plus de 99. Ce serait bien étonnant qu'il y en ait eu 5, ou 6, ou 7, mais quand doit-on s'arrêter ?
S'il y en a 30, la loi du résultat est :

```
plot(dhyper(0:5,30,70,5),type="h")
round(dhyper(0:5,30,70,5),4)
```

```
[1] 0.1608 0.3654 0.3163 0.1302 0.0255 0.0019
```



Sous l'hypothèse H_{30} l'événement " L'échantillon contient au moins 4 étudiants d'opinion positive " a une probabilité de 2.7% :

```
sum(dhyper(4:5,30,70,5))
```

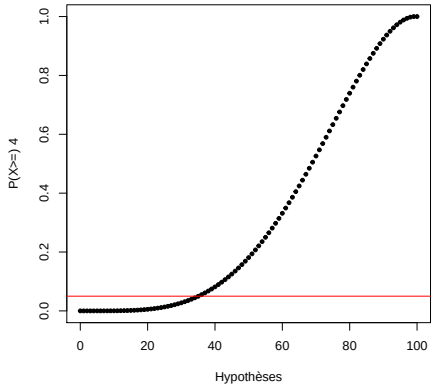
```
[1] 0.02737314
```

Si H_{30} est vraie, un résultat aussi favorable à X est peu plausible. On dit :

> L'hypothèse H_{30} est rejetée à gauche avec un risque d'erreur de première espèce de 5%.

On peut calculer ce risque pour chacune des hypothèses et tracer la courbe correspondante.

```
tracegauche <- function(j) {
        hypothese <- 0:100
        sumprobagauche <- sapply(hypothese, function(x) sum(dhyper(j:5,x,100-x,5)))
        plot(hypothese, sumprobagauche, xlab="Hypothèses", ylab=paste("P(X>=)",j), pch=20)
        abline(h=0.05,col="red")
}
tracegauche(4)
```



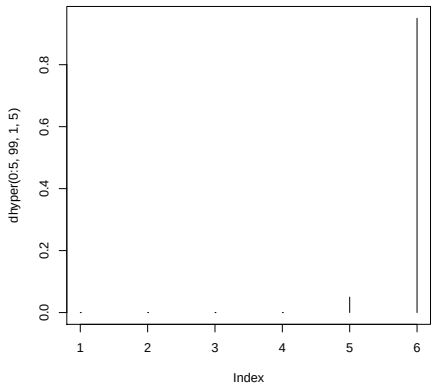
Pour les hypothèses de H_0 jusqu'à H_{35} le résultat est trop favorable au risque de 5% pour qu'on puisse accepter l'hypothèse.

```
     [,1]   [,2]
[1,]   33 0.0352
[2,]   34 0.0396
[3,]   35 0.0444
[4,]   36 0.0495
[5,]   37 0.0551
[6,]   38 0.0611
```

A l'inverse, on peut se demander pour quelles hypothèses le résultat est trop défavorable. La loi pour H_{99} est :

```
plot(dhyper(0:5,99,1,5),type="h")
round(dhyper(0:5,99,1,5),4)
```

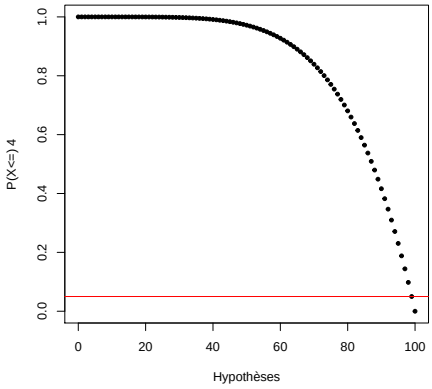
```
[1] 0.00 0.00 0.00 0.00 0.05 0.95
```



L'hypothèse H_{99} est rejetée à droite avec un risque d'erreur de première espèce de 5%.

```
tracedroite <- function(j) {
       hypothese <- 0:100
```

```
        sumprobadroite <- sapply(hypothese, function(x) sum(dhyper(0:j,x,100-x,5)))
        plot(hypothese, sumprobadroite, xlab="Hypothèses", ylab=paste("P(X<=)",j), pch=20)
        abline(h=0.05,col="red")
}
tracedroite(4)
```



3.3 Intervalle de confiance

Les hypothèses H_0 jusqu'à H_{35} et H_{99} jusqu'à H_{100} sont rejetées au risque d'erreur de 5%. Les hypothèses H_{36} jusqu'à H_{98} ne peuvent être exclues. On dit que :

> Les valeurs de 36 à 98 forment l'intervalle de confiance $[36, 98]$ de l'estimation de m au risque de première espèce de 10%.

Trouver dans un échantillon de taille $r = 5$ dans une population de taille $n = 100$ un résultat de $j = 4$ opinions positives conduit à penser que dans la population toute entière il y a entre $m = 36$ et $m = 98$ personnes d'opinion positive. La maîtrise, pour l'utilisation, des concepts de vraisemblance, test d'hypothèse et estimation est le seul objectif de ce cours. A retenir :

> Le calcul des probabilités parle de l'échantillon à partir de la population, la statistique inférentielle parle de la population à partir de l'échantillon.

4 Stratégies de calculs

On comprend la difficulté associée aux calculs. Les mathématiciens ont proposé des solutions générales basées sur des approximations. On parle de méthodes asymptotiques. Les numériciens ont proposé des solutions générales basées sur le principe des simulations. On parle de méthode de Monte-Carlo. Qu'on utilise l'une ou l'autre, le raisonnement statistique reste le même.

4.1 L'urne U_{NB} vue par les mathématiques

Une boîte contient N boules noires et B boules blanches. Quand on prélève sans remise des boules dedans, on trouve n boules noires et b boules blanches.

La fréquence inconnue de boules blanches est $T = \frac{B}{N+B}$.
La fréquence observée, après l'expérience, est $t = \frac{b}{n+b}$.
Inférer, c'est parler de T en partant de t.

L'amphi du paragraphe précédent est une urne de ce type. L'ensemble des français âgés de 18 ans et plus en est une autre. Un soir d'élections, on ne compte que ceux qui ont voté - en général on utilise G pour gauche et D pour droite - mais l'image noir/blanc nous vient des Grecs anciens. Si on a le temps d'attendre le décompte complet, on connaît T. Mais comme l'émission commence à 20 h, on veut le résultat tout de suite. On l'estime sur une partie des comptages. C'est une situation ordinaire. Nous avons vu que, dans notre amphi, 4 étudiants répondent OUI à "la statistique est intéressante" (boules blanches b) et un étudiant NON (boules noires n). On a vu que pour $\frac{b}{n+b} = \frac{4}{5} = 0.8$, on pouvait inférer (au risque de 5%) que

$$0.36 \leq T = \frac{B}{N+B} \leq 0.98$$

Supposons maintenant que l'urne contienne plusieurs dizaines de millions de boules et qu'on ait un échantillon de 1000 boules : $n = 620$ sont noires et $b = 380$ sont blanches. Que peut-on dire de T à partir de $t = 0.38$ et $n + b = 1000$?

Chaque valeur possible comprise entre 0 et 1 est une hypothèse. L'hypothèse H_T considère que le taux de boules blanches dans l'urne géante est T. Sous cette hypothèse, la probabilité d'observer x boules blanches est donnée par une loi binomiale de paramètres $n+b = 1000$ et T. La valeur de b est donc la réalisation d'une variable aléatoire de moyenne $1000T$ et de variance $1000T(1-T)$, en général $(n+b)T$ et $(n+b)T(1-T)$ pour un échantillon de taille $(n+b)$. La probabilité du résultat b pour l'hypothèse T est :

$$P(b) = \left(\begin{array}{c} n+b \\ b \end{array} \right) T^b (1-T)^n$$

La vraisemblance de l'hypothèse est :

$$L(T) = \left(\begin{array}{c} n+b \\ b \end{array} \right) T^b (1-T)^n$$

Si on cherche une valeur de T qui maximise cette quantité, il est équivalent d'étudier :

$$LL(T) = -\ln\left(\left(\begin{array}{c} n+b \\ b \end{array} \right) T^b (1-T)^n \right)$$

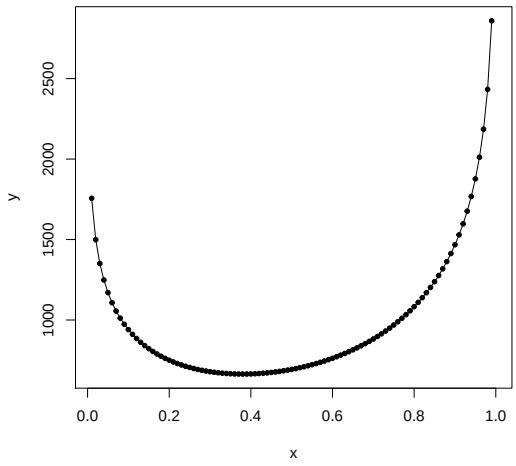
où ln désigne la fonction logarithme néperien. ln est une fonction monotone. LL signifie Log-Likelihood. Le signe - est utilisé car $L(T)$ est comprise entre 0 et 1 et son logarithme est négatif. La Log-Vraisemblance est alors positive et on cherche à la minimiser.

$$LL(T) = -\ln\left(C^{te}\right) - b\ln(T) - n\ln(1-T)$$

Dans le problème, la constante ne joue aucun rôle et on peut ne s'intéresser qu'à la fonction :

$$f(T) = -b\ln(T) - n\ln(1-T)$$

```
x <- seq(0.01,1,le=100)
b <- 380 ; n <- 620
y <- -b*log(x)-n*log(1-x)
plot(x,y,type="o", pch=20)
```



Le minimum est atteint pour :

$$f'(T) = -b\frac{1}{T} + n\frac{1}{1-T} = 0 \Rightarrow b(1-T) = nT \Rightarrow T = \frac{b}{b+n}$$

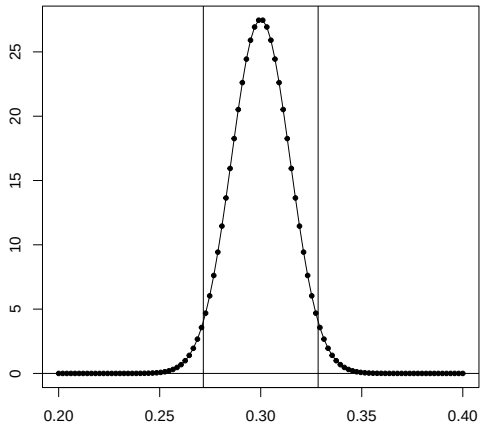
L'estimation au maximum de vraisemblance de la fréquence théorique est encore la fréquence observée. Le calcul est très différent mais le principe est le même.

Comment rejeter une hypothèse ?
Prenons par exemple, $T_1 = 0.30$. T_1 est-elle acceptable pour un tirage $b = 380, n = 620$? Le résultat d'un tirage est une variable aléatoire de Bernoulli qui prend la valeur 1 (blanc / OUI) avec la probabilité T_1 et la valeur 0 (noir / NON) avec la probabilité $1 - T_1$.

Le nombre de boules blanches est la somme de 1000 variables de ce type c'est-à-dire toujours une loi binomiale $\mathcal{B}(n+b,T_1)$. D'après le théorème central limite (voir 2.3.), la loi binomiale converge, en loi, vers la loi Normale de moyenne $(n+b)T_1$ et de variance $(n+b)T_1(1-T_1)$.

Le taux observé, quant à lui, tend vers une loi normale de moyenne $\mu = T_1$ et de variance $\sigma^2 = \dfrac{T_1(1-T_1)}{(n+b)}$.

```
x <- seq(0.20,0.40,le=100)
b <- 380 ; n <- 620 ;
m0 <- 0.30 ; sd0 <- sqrt(0.30*0.70/1000)
y <- dnorm(x,mean=m0,sd=sd0)
plot(x,y,type="o",pch=20, xlab="",ylab="")
x1 <- m0-qnorm(0.975)*sd0 ; x2 <- m0+qnorm(0.975)*sd0
abline(v=x1)
abline(v=x2)
abline(h=0)
```



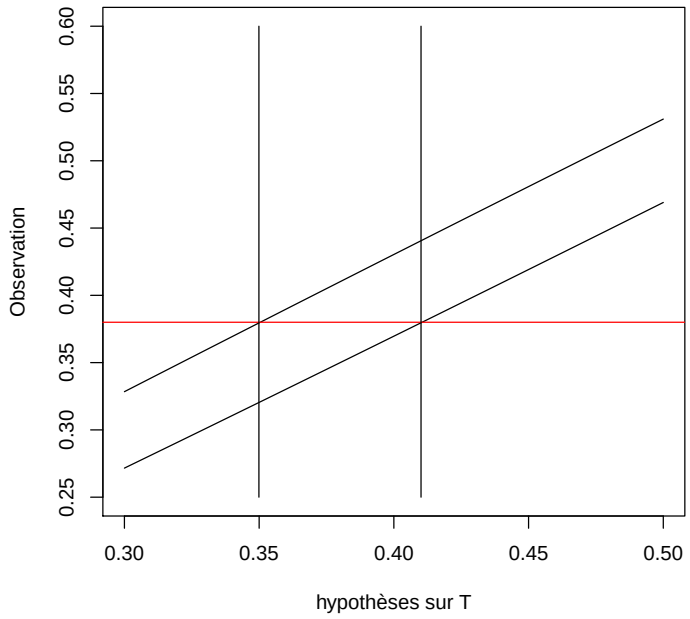
Dans une loi normale, la probabilité d'être dans l'intervalle

$$[\mu - 1.96 \times \sigma, \mu + 1.96 \times \sigma]$$

vaut 0.95. Si l'observation tombe en dehors de cet intervalle, au risque de se tromper de 5% quand c'est vrai, on pense qu'elle est anormale, c'est-à-dire que l'hypothèse est fausse. Ici l'observation 0.38 est dans la zone de rejet. L'hypothèse $T_1 = 0.30$ n'est pas acceptable.

Pour constituer un intervalle de confiance, on cherche enfin toutes les hypothèses inacceptables. Si t est le taux observé et T_1 le taux théorique, T_1 est inacceptable si :

$$t < T_1 - 1.96 \times \sqrt{\frac{T_1(1-T_1)}{(n+b)}}$$
$$t > T_1 + 1.96 \times \sqrt{\frac{T_1(1-T_1)}{(n+b)}}$$



Ce n'est pas un problème simple mais on peut utiliser une approximation des solutions [0.35,0.41] par :

$$\left[t - 1.96 \times \sqrt{\frac{t(1-t)}{(n+b)}}, t + 1.96 \times \sqrt{\frac{t(1-t)}{(n+b)}}\right]$$

A retenir : dans des conditions de calcul très différentes, nous avons utilisé les mêmes idées pour dire :

> 0.38 est l'estimation au maximum de vraisemblance du taux de boules blanches.
> L'hypothèse 0.30 est inacceptable au risque de 5%.
> L'intervalle de confiance au risque de 5% est voisin de [0.35,0.41].

4.2 L'urne U_{NB} vue par l'informatique

Comme les calculs sont souvent très difficiles, l'idée est venue de les simuler par ordinateur. On peut fabriquer l'urne.

```
urne <- rep(c(0,1),c(80,240))
urne
```

```
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [41] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [81] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[161] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[201] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[241] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[281] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

On peut alors l'échantillonner, sans remise, 20 boules de l'urne. On rappelle que 1 est une boule blanche et 0 est une noire.

```
sample(urne,20)
[1] 1 1 1 1 0 1 1 0 1 0 0 1 0 1 1 1 1 1 1 0
sample(urne,20)
[1] 0 0 1 0 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1
sample(urne,20)
[1] 1 1 1 1 1 0 0 1 0 1 1 1 1 1 0 1 1 1 0 0
```

Pour compter les boules blanches, il suffit de faire la somme :

```
(nbrB <- sample(urne,20))
[1] 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0
sum(nbrB)
[1] 16
```

Pour avoir la loi du nombre de boules blanches, on fait alors 1000 fois l'expérience :

```
res <- rep(0,1000)
for (i in 1:1000) {
    res[i] <- sum(sample(urne,20))
}
head(res, n=15)
[1] 15 15 17 10 14 15 15 14 16 16 15 18 13 16 13
tail(res, n=15)
[1] 16 13 17 15 14 14 12 16 14 14 15 18 14 15 17
```

On compte les résultats :

```
table(res)
res
  9  10  11  12  13  14  15  16  17  18  19  20
  4  10  22  59  98 174 221 180 148  65  16   3
```

Quand on tire 20 boules dans une urne avec 240 blanches et 80 noires, on obtient 98 écahntillons ayant 13 boules blanches soit une probabilité approchée de 0.098. Mathématiquement, on aurait obtenu :

$$P(13) = \frac{\begin{pmatrix} 240 \\ 13 \end{pmatrix}\begin{pmatrix} 80 \\ 7 \end{pmatrix}}{\begin{pmatrix} 300 \\ 20 \end{pmatrix}} = 0.113 \ \textit{Loi hypergéométrique}$$

```
dhyper(13,240,80,20)
[1] 0.1130246
```

$$P(13) = \begin{pmatrix} 20 \\ 13 \end{pmatrix} \left(\tfrac{240}{320}\right)^{13} \left(\tfrac{80}{320}\right)^{7} = 0.112 \ \textit{Loi binomiale}$$

```
dbinom(13,20,240/320)
[1] 0.1124062
```

Quand on sait faire le calcul exactement, c'est parfait. Quand on a une approximation mathématique, si les conditions d'approximations sont respectées, c'est bon. Quand on ignore la solution, on peut l'approcher et la précision ne dépend que du temps de calcul.

```
res
   7    8    9   10   11   12   13   14   15   16   17   18   19   20
   1    8   24   93  229  577 1116 1725 2085 1956 1395  582  178   31
```

Si on prend 10000 échantillons, on trouve 0.1116.

```
res
    5     7     8     9    10    11    12    13    14    15    16    17    18    19
    1    13    55   219   852  2540  5831 11342 17490 20914 19210 13080  6315  1876
   20
  262
```

Si on prend 100000 échantillons, on trouve 0.1134.

La totalité du raisonnement statistique peut donc se faire avec un ordinateur. Le plus important c'est d'en connaître le principe. Donnons un exemple. Soit un amphi de 100 étudiants de la génération de vos parents. On demande à chacun quel est le jour de son anniversaire et on note qu'il existe 80 dates de naissance différentes. Peut-on choisir entre les deux hypothèses :

H_0 : cela n'a rien d'extraordinaire et provient directement du hasard,
H_1 : ce résultat est trop faible et pourrait indiquer qu'il y a des périodes de l'année pour lesquelles le nombre de naissances est plus grand.

La première hypothèse, dite hypothèse nulle, définit un modèle probabiliste. L'ensemble des résultats possibles est celui des manières de distribuer $r = 100$ objets dans $n = 365$ cases. L'expérience définit une variable aléatoire sur ce modèle : c'est le nombre de cases occupées. Si la seconde hypothèse est vraie, le nombre de jours anniversaires différents sera plus petit que prévu. Avant de faire le moindre calcul, nous décidons d'étudier l'événement E " il y a au plus 80 dates d'anniversaire différentes ". Si cet événement a une faible probabilité de survenir sous l'hypothèse du hasard, disons α, nous dirons H_0 est rejetée au profit de H_1 avec un risque d'erreur de α, sinon nous dirons que le test n'est pas significatif. Reste à faire le calcul.

Distribuer des objets dans des cases, c'est échantillonner avec remise. Echantillonner sans remise interdit de reprendre un objet :

```
sample (1:10,8)
[1]  8  3 10  4  6  5  9  2
```

Echantillonner avec remise permet de reprendre un objet :

```
sample (1:10,8,replace=T)
[1] 2 9 4 3 7 9 9 8
```

Il nous faut des échantillons avec remise de 100 objets parmi 365 :

```
sample(1:365,100,replace=T)
  [1] 292  98 361 229  39 257  40 212 105   9  14 257 356 223  84 169 171 298 191 253
 [21] 326 308 278 187 227 255 240 173 100 346 220 266 338 235  19 181 360  59 337 143
 [41]  40 335   5 146 256 313 115 320  67 347 164 127 264 139 220 327 139  13 330 271
 [61] 222 244 166 169 323 355  62 349 119  52 114  77 363 120 236 209  29 357 326 210
 [81] 292  43 181 159 146 225  15 213 178 142 201 333  53 120 254 327  39 344 359 170
```

Il faut compter le nombre de résultats distincts :

```
length(unique(sample(1:365,100,replace=T)))
[1] 85
```

Nous faisons 1000 tirages et comptons les cas observés. Ceci donne une approximation de la loi de la variable désirée :

```
res <- rep(0,1000)
for (i in 1:1000) {
    res[i] <- length(unique(sample(1:365,100,replace=T)))
}
table(res)
```

```
res
 79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97
  4   8   9  24  45  67  82 104 134 149 106 104  74  46  23  10   9   1   1
```

La probabilité de E est estimée par simulation à 0.012. Pour plus de sûreté, nous en refaisons 10000 :

```
res <- rep(0,10000)
for (i in 1:10000) {
    res[i] <- length(unique(sample(1:365,100,replace=T)))
}
table(res)
```

```
res
  76   77   78   79   80   81   82   83   84   85   86   87   88   89   90   91   92
   1    3    7   21   44  118  240  418  615  855 1138 1320 1331 1240 1044  736  467
  93   94   95   96   97   99
 253  105   31   10    2    1
```

La probabilité de E est estimée par simulation à 0.0076. La conclusion est :

H_0 est rejetée au profit de H_1 avec un risque d'erreur de $\alpha < 1\%$