DESINF

Cours 1 et 2

Julien CLAUDE

January 3, 2025

Généralités et définitions

- Objectif des biostatistiques
- Population
- Echantillon
- Variable statistique
- Variable ordonnée et non ordonnée
- Variable quantitative et qualitative
- Variables continues et discrètes
- Distribution

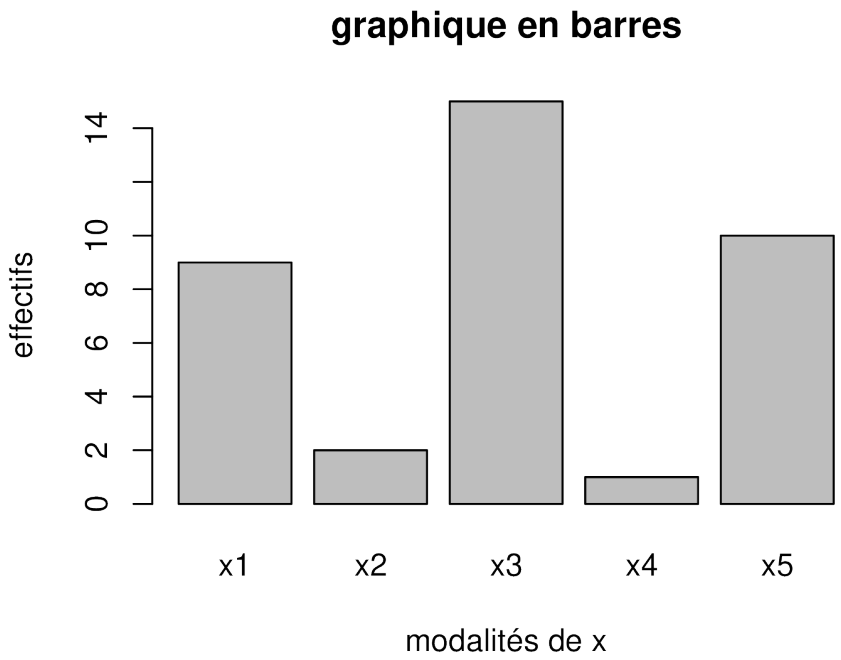
Example: un échantillon d'escargots de la famille des Conidae

Représentation des variables

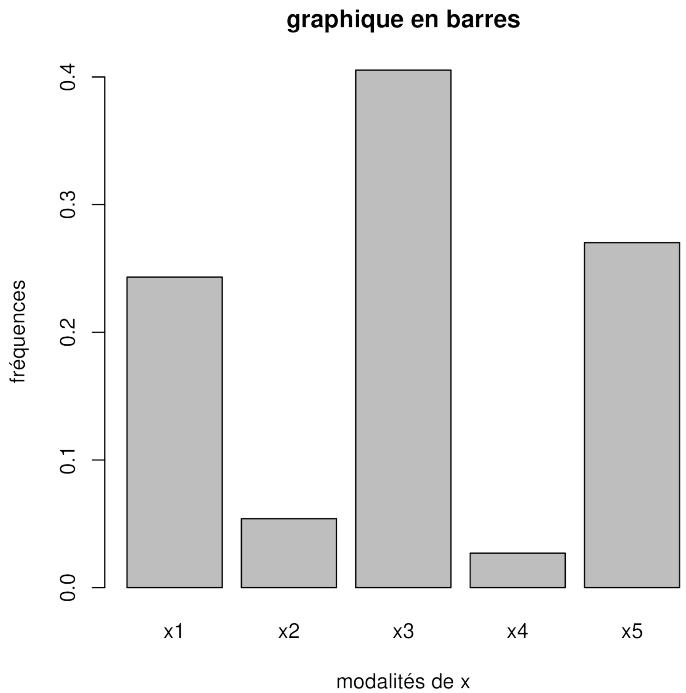
Tableau effectif et valeurs (ou modalités)

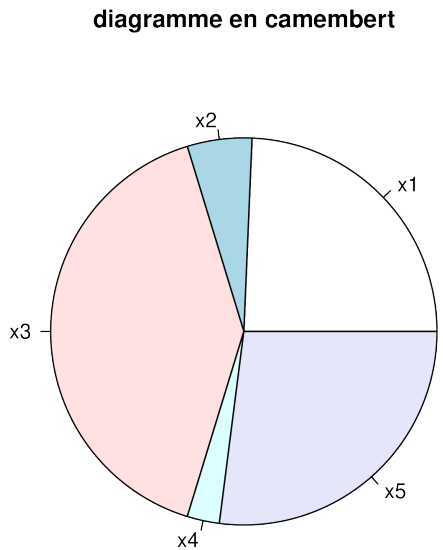
valeur	effectif	fréquence
x_1	n_1	$f_i = n_1 / \sum_{i=1}^{i=p} n_i$
x_2	n_2	$f_2 = n_2 / \sum_{i=1}^{i=p} n_i$
...	...	...
x_i	n_i	$f_i = n_i / \sum_{i=1}^{i=p} n_i$
...	...	...
x_p	n_p	$f_i = n_p / \sum_{i=1}^{i=p} n_i$

$$N = \sum_{i=1}^{i=p} n_i$$



Représentation en fréquences





Fréquences cumulées: soit une distribution d'une variable ordonnée $(x_i; n_i)$, avec $i \in [1; p]$, et k un entier appartenant à $[1; p]$, alors la fréquence cumulée d'ordre k vaut:

$$F_k = f_1 + f_2 + ... + f_i + ... + f_k = \sum_{i=1}^{i=k} f_i$$

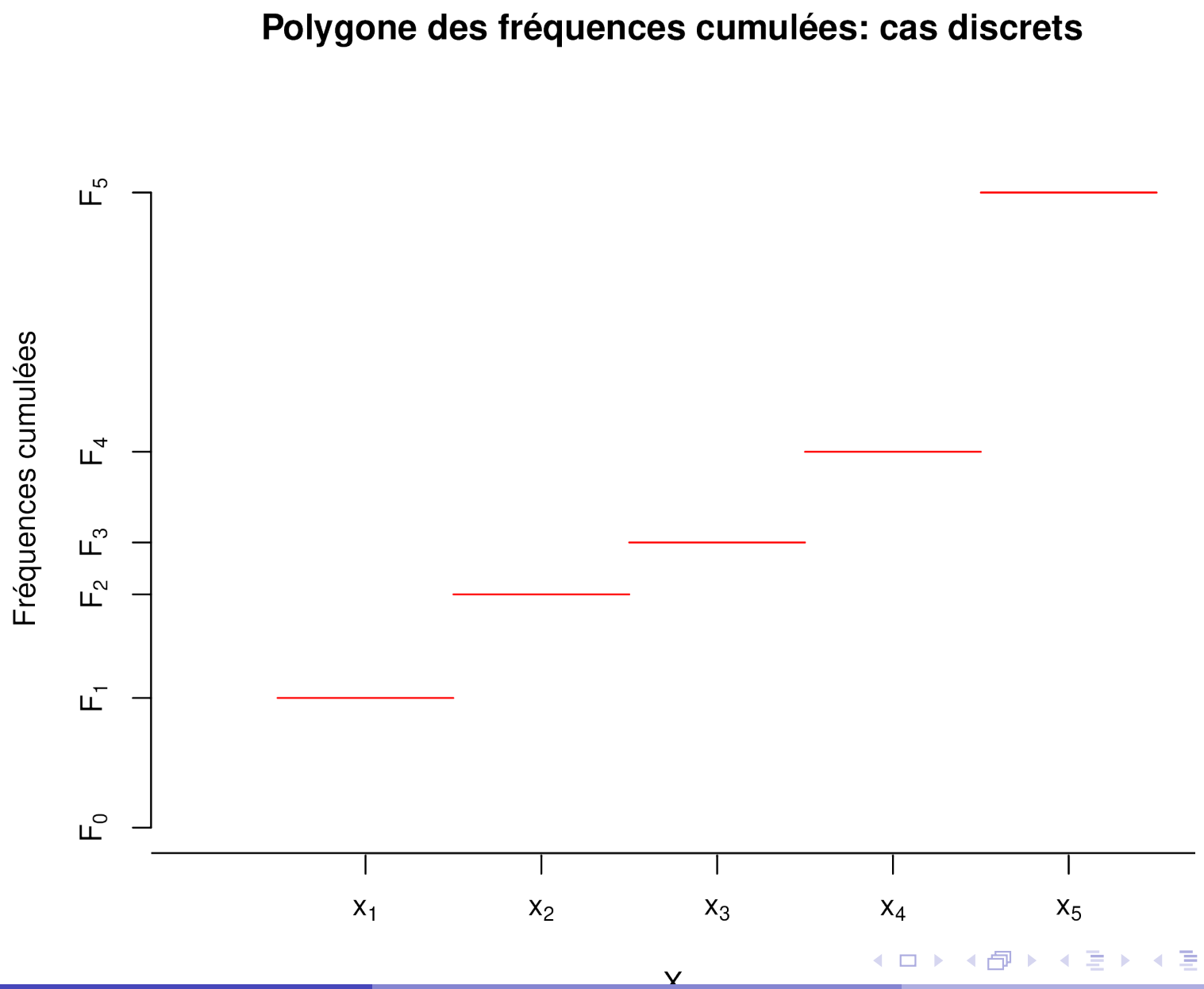
La valeur x_i correspondant à F_i porte le nom de quantile de X à $F_i \times 100\%$

Variables discrètes ordonnées

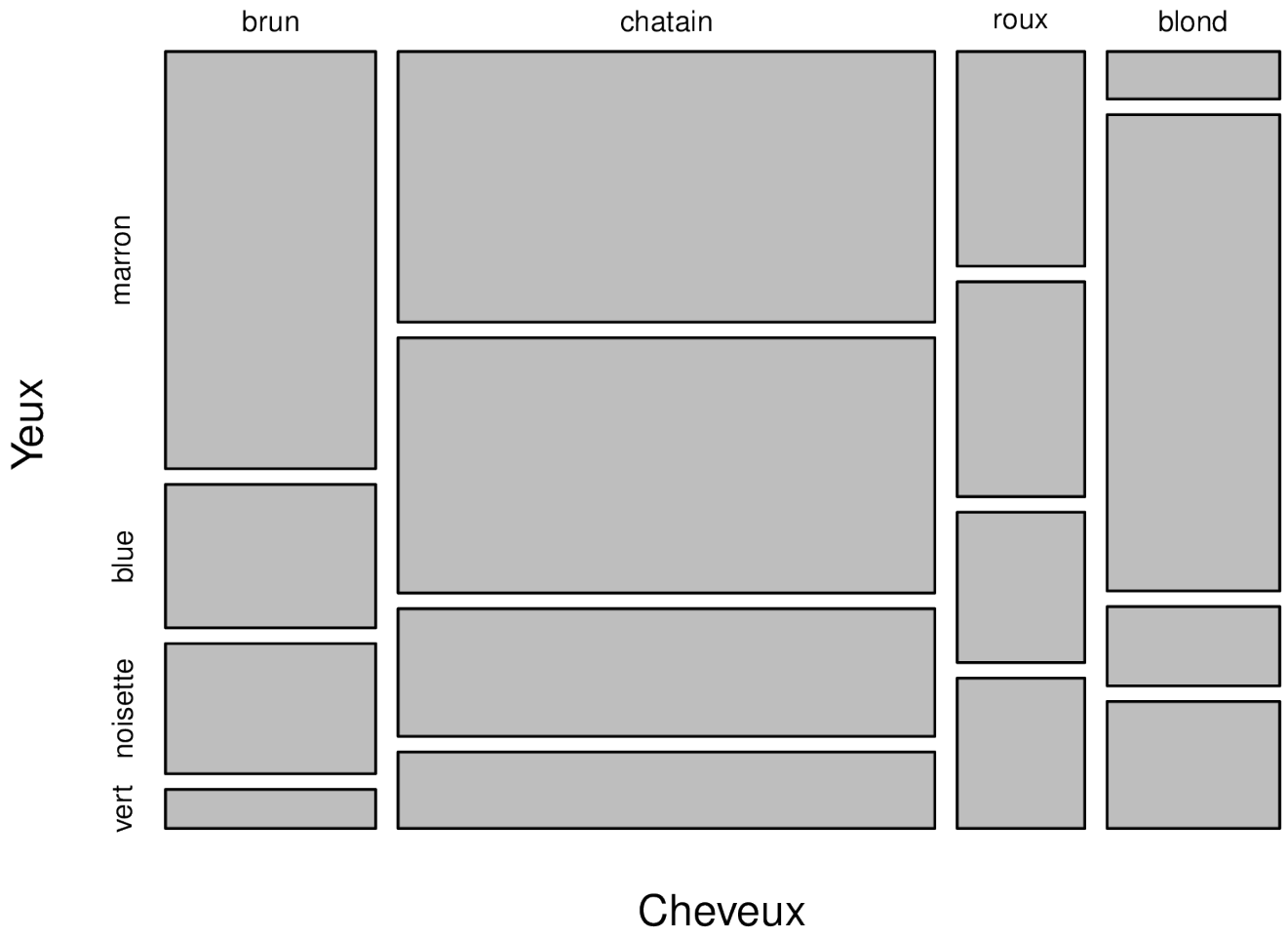
valeur	effectif	fréquence	Fréquences cumulées
x_1	n_1	$f_1 = n_1 / \sum_{i=1}^{i=p} n_i$	$F_1 = n_1/N$
x_2	n_2	$f_2 = n_2 / \sum_{i=1}^{i=p} n_i$	$F_2 = n_1/N + n_2/N$
...	...	...	...
x_i	n_i	$f_i = n_i / \sum_{i=1}^{i=p} n_i$	$F_i = \sum_{i=1}^{i=i} n_i/N$
...	...	...	...
x_p	n_p	$f_p = n_p / \sum_{i=1}^{i=p} n_i$	$F_p = 1$

$N = \sum_{i=1}^{i=p} n_i$

Polygone des fréquences cumulées pour variables discrètes



Graphique en mosaique

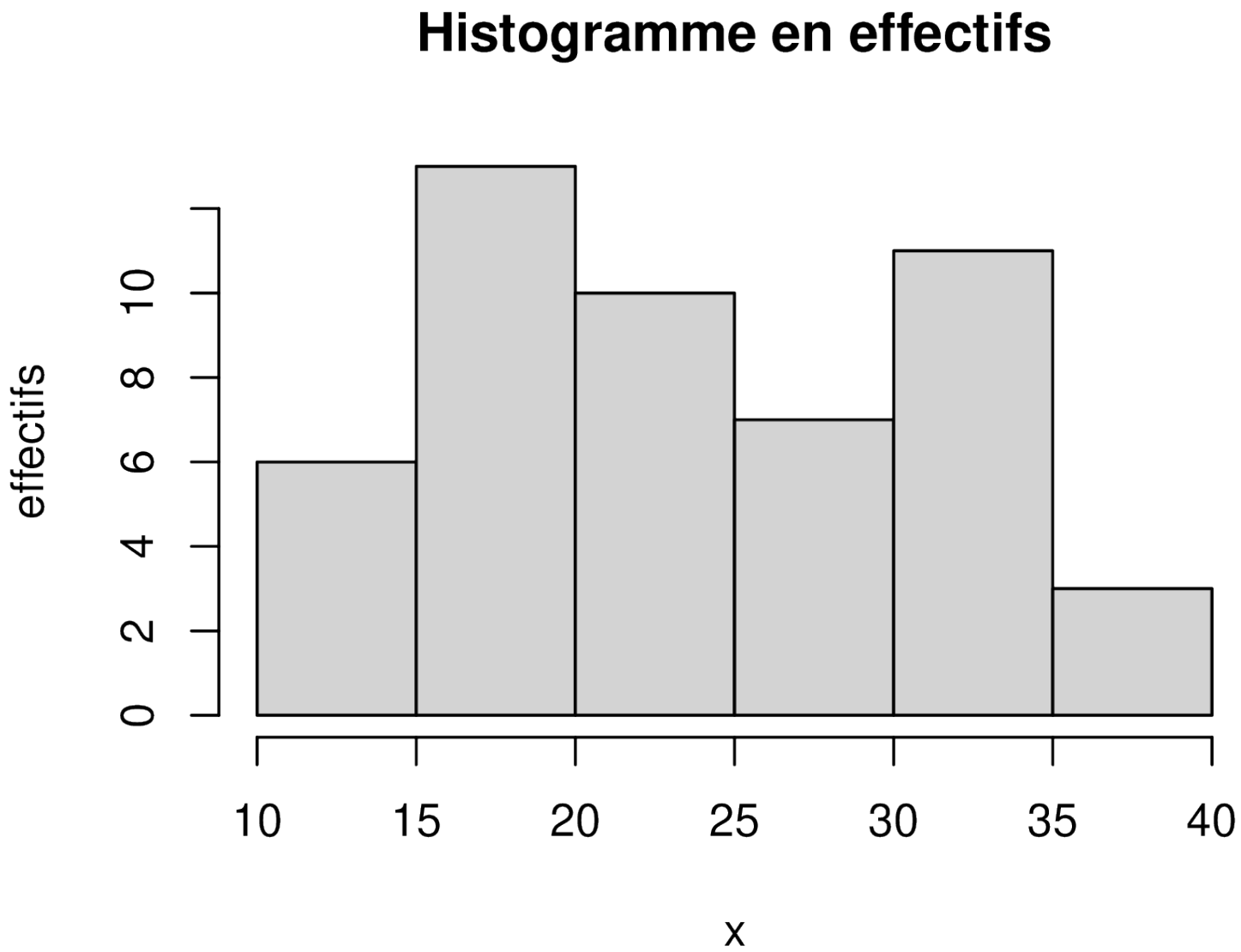


Cas des variables quantitatives continues

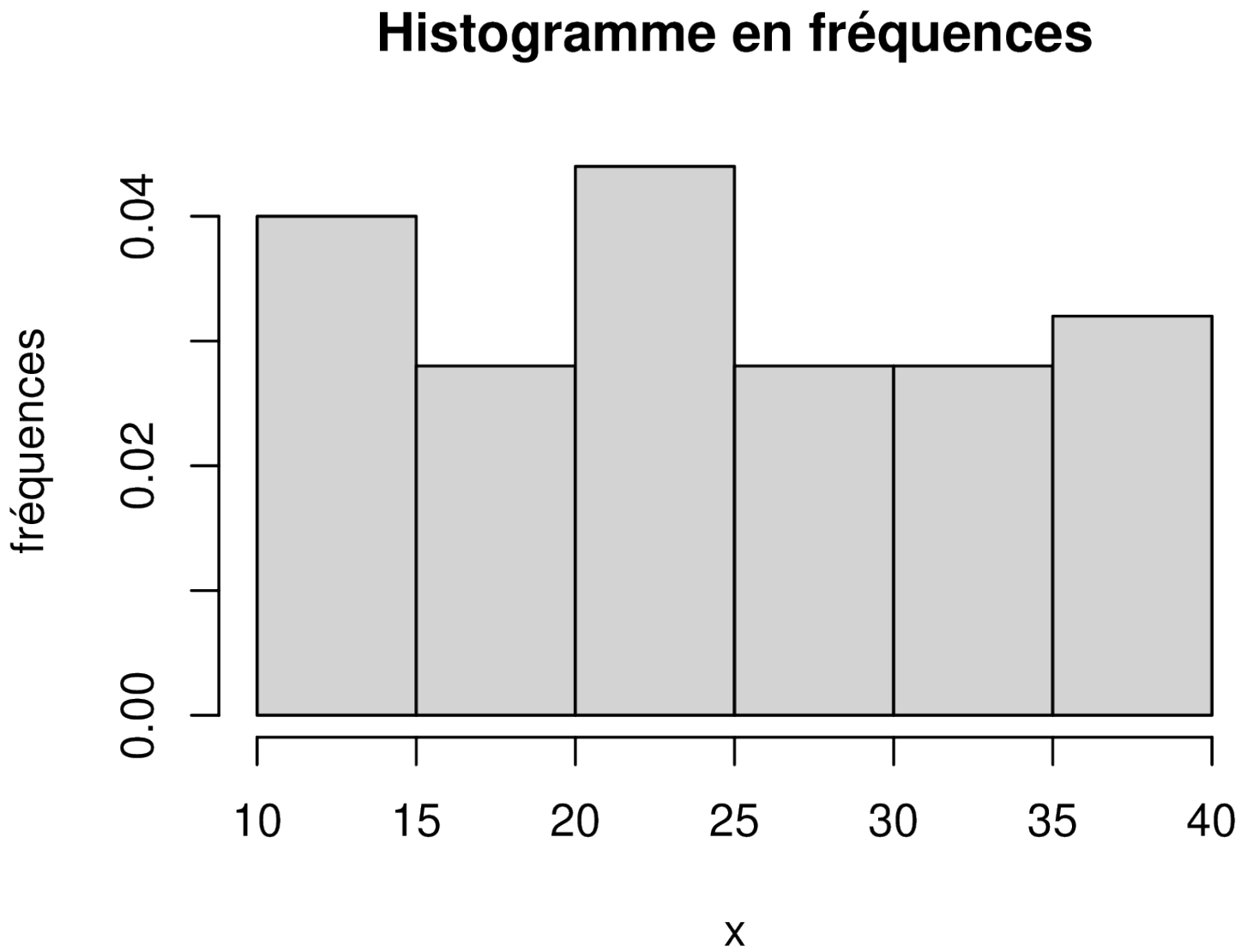
Tableau classes, effectifs, fréquences et fréquences cumulées

classe	effectif	fréquence	Fréquence cumulée
$[x_0; x_1]$	n_1	$f_1 = n_1 / \sum_{i=1}^{i=p} n_i$	$F_1 = f_1$
$]x_1; x_2]$	n_2	$f_2 = n_2 / \sum_{i=1}^{i=p} n_i$	$F_2 = f_1 + f_2$
...	...	...	...
$]x_{i-1}; x_i]$	n_i	$f_i = n_i / \sum_{i=1}^{i=p} n_i$	$F_i = \sum_{i=1}^{i=i} f_i$
...	...	...	...
$]x_{p-1}; x_p]$	n_p	$f_p = n_p / \sum_{i=1}^{i=p} n_i$	$F_p = 1$

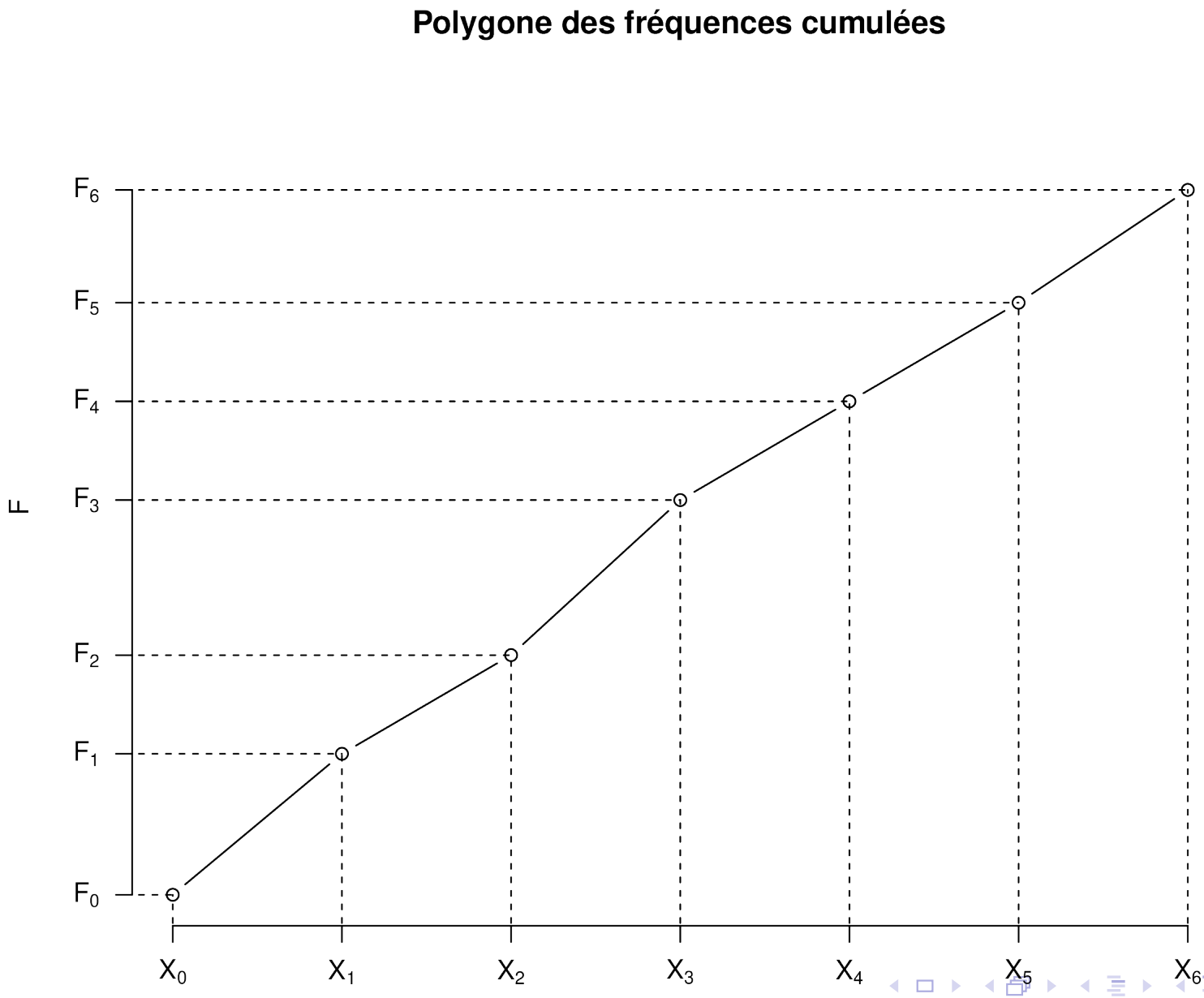
Distribution des variables continues



Distribution des variables continues

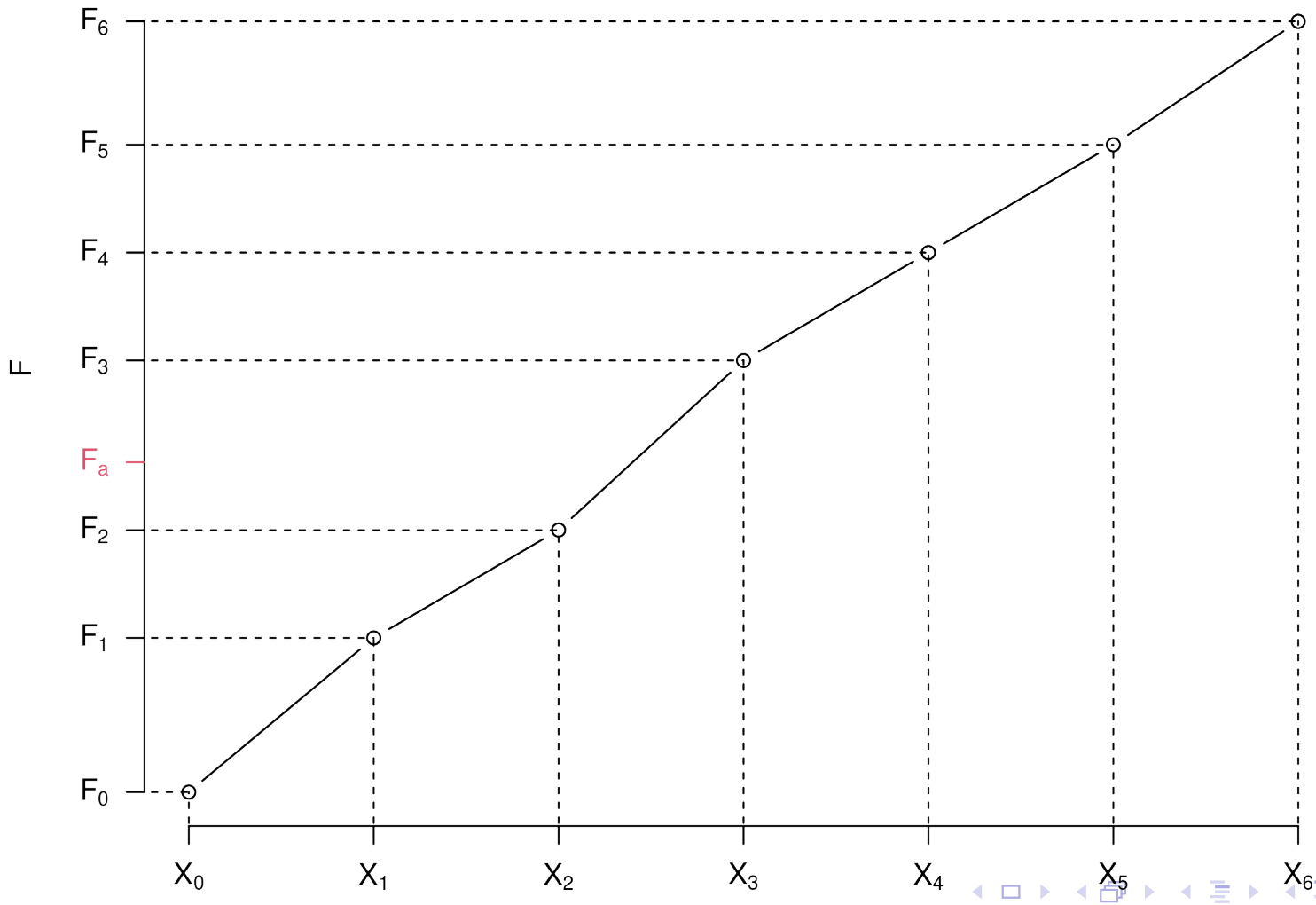


Fréquences cumulées et quantiles



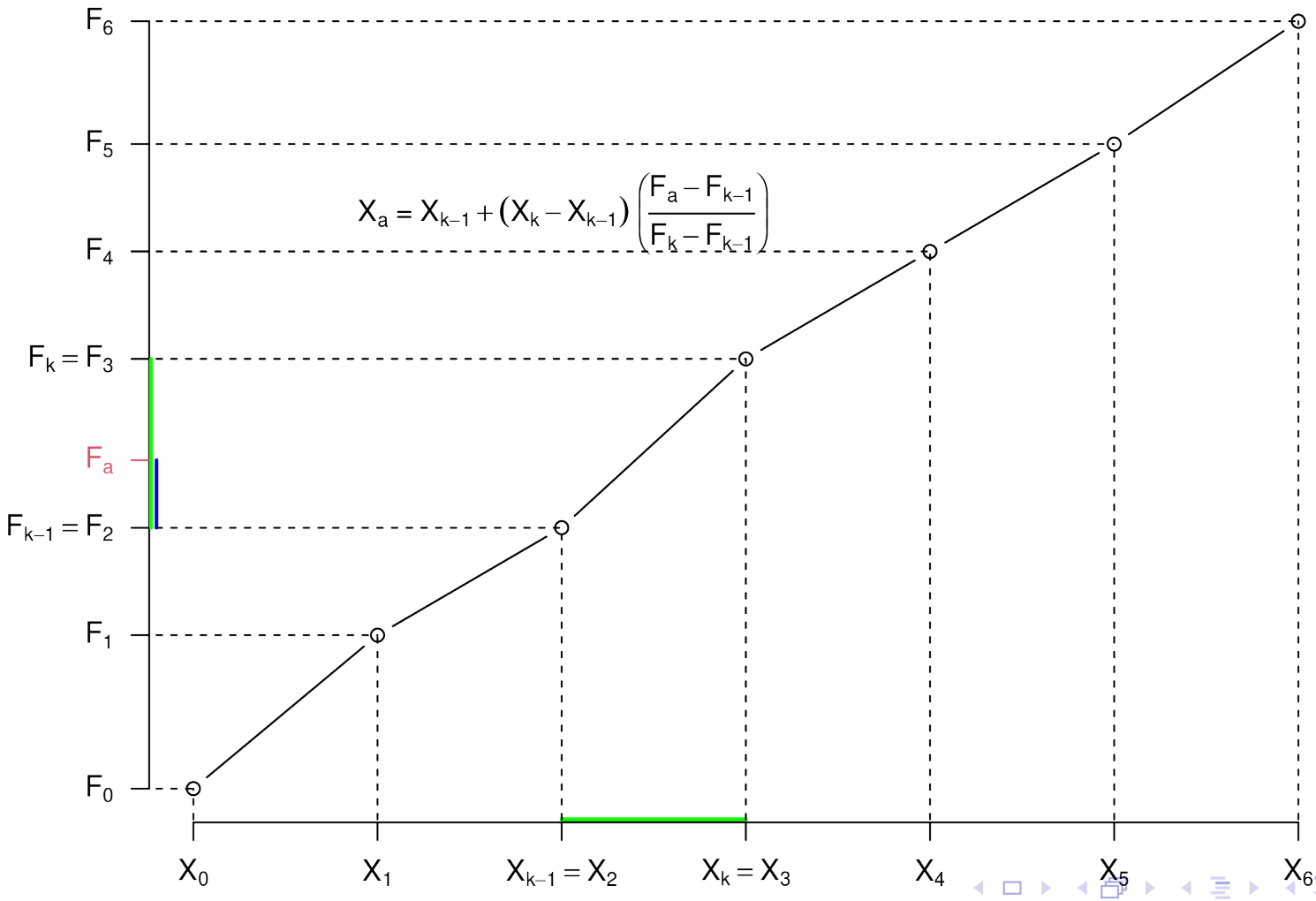
Fréquences cumulées et quantiles

Recherche d'un quantile par interpolation linéaire



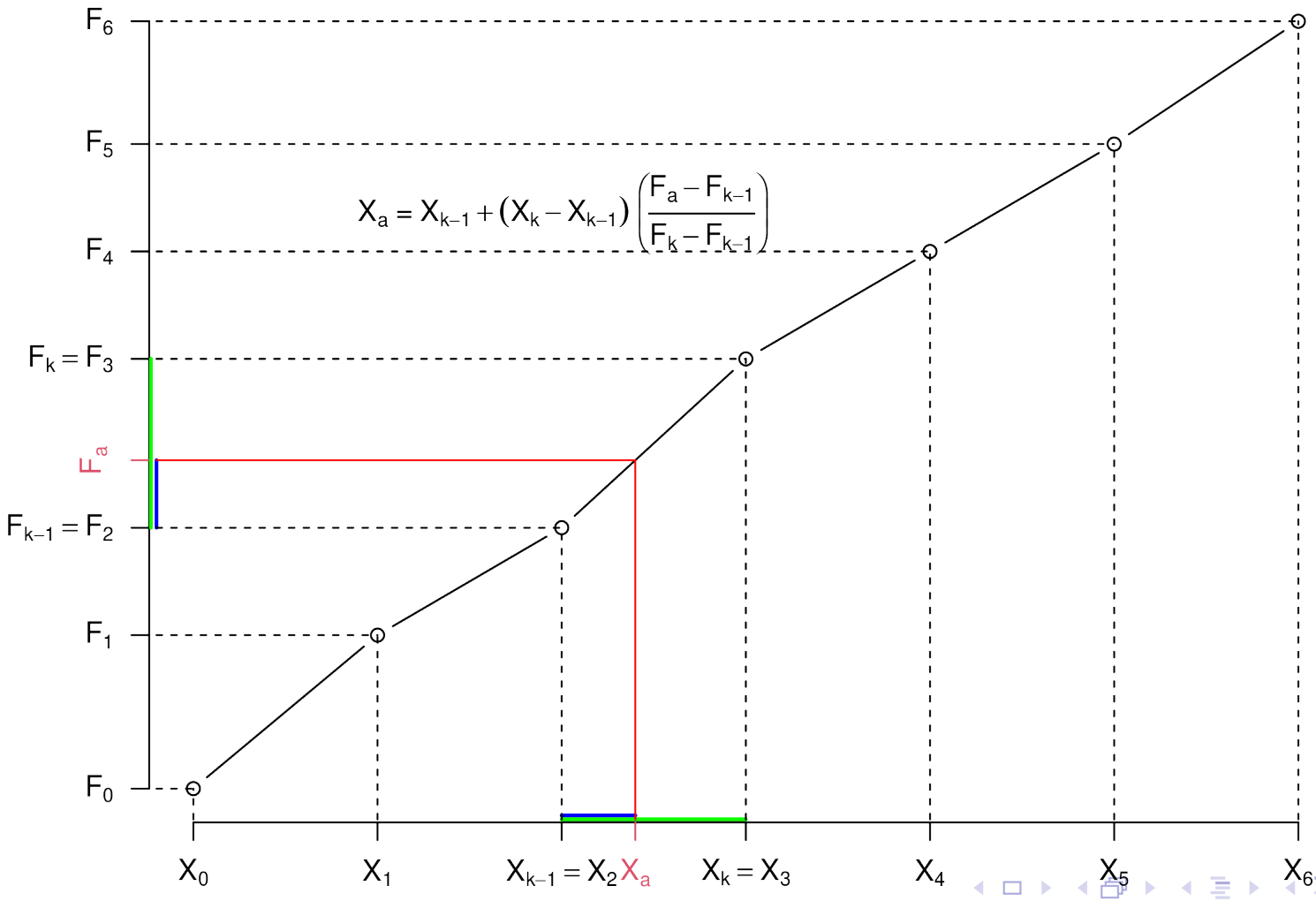
Fréquences cumulées et quantiles

Recherche d'un quantile par interpolation linéaire



Fréquences cumulées et quantiles

Recherche d'un quantile par interpolation linéaire



Paramètres de tendance centrale

- La médiane ou quantile à 50%
- Le mode : classe ayant la plus forte fréquence dans la distribution
- La moyenne

$$\mu = \frac{1}{N}\sum_{i=1}^{i=p} n_i x_i \quad ; \quad \text{avec} \quad N = \sum_{i=1}^{i=p} n_i$$

Ou bien si on dispose des fréquences

$$\mu = \sum_{i=1}^{i=p} f_i x_i \quad ; \quad \text{avec} \quad f_i = \frac{n_i}{N}$$

Quand les fréquences correspondent à la probabilité de réalisation de X, cette expression désigne l'espérance de X notée $\mathbb{E}(X)$.

Moyenne de moyennes

$$\mu = \frac{1}{N}(n_1\mu_1 + n_2\mu_2); \quad \text{avec} \quad N = n_1 + n_2$$

Estimer la moyenne à partir d'un échantillon

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i \quad ; \quad \text{avec} \quad N = \sum_{i=1}^{i=p} n_i$$

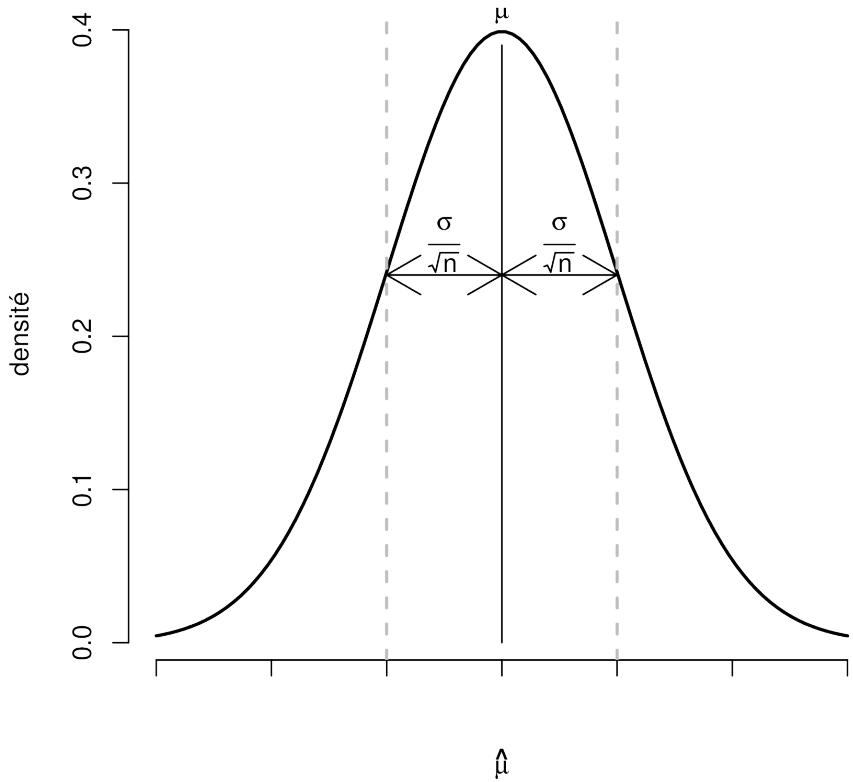
Loi des grands Nombres: Cette loi stipule que plus on augmente la taille de l'échantillon, plus les caractères statistiques de l'échantillon se rapprochent des caractères de la population. Donc plus l'échantillon sera grand plus ces paramètres seront représentatifs de la population, et plus il sera petit et plus le risque d'imprécision sera élevé.

Théorème central limite

Si $X_1, X_2, ...X_n$ est une suite de variables aléatoires appartenant à des lois de distribution identiques avec une espérance μ et un écart type σ, et que ces variables sont indépendantes, alors la somme $S_n = X_1 + X_2 + ... + X_n$ a une espérance $n\mu$ et l'écart type de cette somme vaut $\sigma\sqrt{n}$. La loi de S_n tend vers la loi normale $\sim \mathcal{N}(n\mu, n\sigma^2)$. Ce théorème affirme donc qu'une somme de variables aléatoires identiques en loi suit une loi normale.
On peut en déduire que:
$S_n/n \sim \mathcal{N}(\mu, \sigma^2/n)$

Théorème central limite

En d'autres termes, si on réalise plusieurs échantillonages dans une population et qu'on regarde la distribution des moyennes de ces échantillons, cette distribution pourra être être connue à l'avance (sa moyenne et son écart type également).



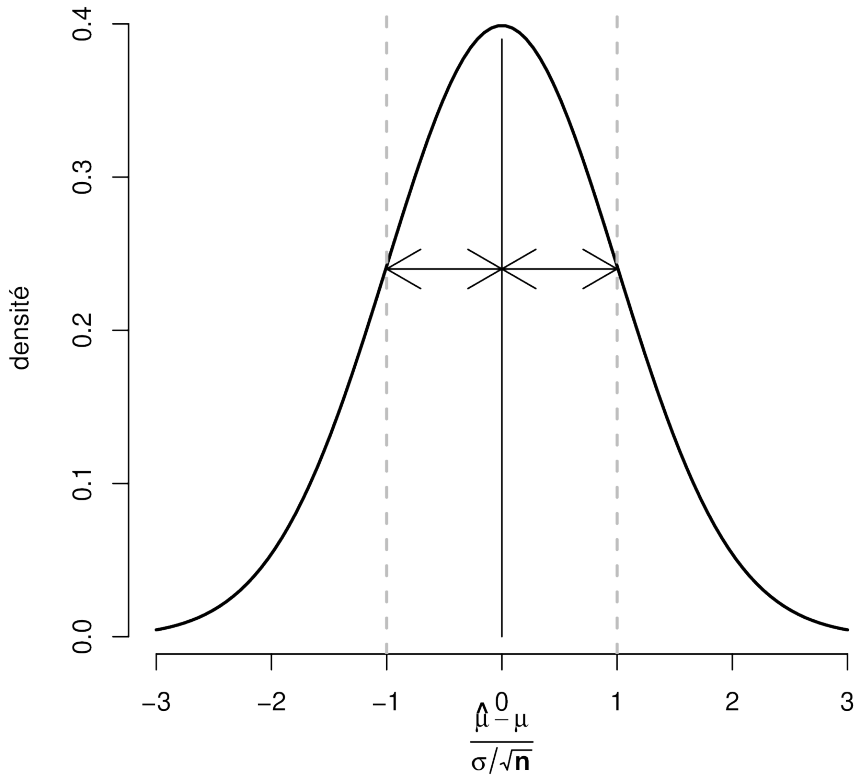
Théorème central limite

On peut aussi écrire d'après ce théorème que:

$$\frac{\sum_{i=1}^{i=n}(X_i/n) - \mu}{\sigma/\sqrt{(n)}} \sim \mathcal{N}(0,1)$$

Ce qui est équivalent à:

$$\frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{(n)}} \sim \mathcal{N}(0,1)$$



Paramètres de dispersion

- Etendue
- Espace interquartile
- Variance et Ecart type

$$Var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu_x)^2 \quad \text{, pour une présentation par observations}$$

$$Var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i - \mu_x)^2 \quad \text{, pour une présentation par fréquences}$$

La variance peut être également calculée grâce à son développement (**développement de Koenig**).

$$Var(x) = \sigma_x^2 = (\frac{1}{n}\sum_{i=1}^{i=p} n_i x_i^2) - \mu_x^2 \quad \text{, pour une présentation par fréquences.}$$

Développement de Koenig: Démonstration

On part de la somme des carrés aux écarts plutôt que de leur moyenne.

$$\sum_{i=1}^{i=p} n_i(x_i - \mu_x)^2 = \sum_{i=1}^{i=p} n_i(x_i^2 - 2x_i\mu_x + \mu_x^2)$$

$$= \sum_{i=1}^{i=p} n_i x_i^2 - 2\mu_x \sum_{i=1}^{i=p} n_i x_i + \sum_{i=1}^{i=p} n_i \mu_x^2$$

or on sait que $\sum_{i=1}^{i=p} n_i = n$ et que $\sum_{i=1}^{i=p} n_i x_i = n\mu_x$
donc $\sum_{i=1}^{i=p} n_i(x_i - \mu_x)^2 = \sum_{i=1}^{i=p} n_i x_i^2 - n\mu_x^2$
donc $var(x) = \frac{1}{n}(\sum_{i=1}^{i=p} n_i x_i^2 - n\mu_x^2) = \frac{1}{n}(\sum_{i=1}^{i=p} n_i x_i^2) - \mu_x^2$

Estimation de la variance à partir d'un échantillon

$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - \hat{\mu}_x)^2$$

Le développement de Koenig donne:

$$\hat{\sigma}_x^2 = (\frac{1}{n-1} \sum_{i=1}^{i=n} x_i^2) - \frac{n}{n-1} \hat{\mu}_x^2$$

Correction de Bessel: démonstration

Soit la variance de la population: $\sigma_x^2 = \frac{\sum_{i=1}^{i=n}(x_i-\mu_x)^2}{n}$

Soit la quantité analogue mais calculée sur un échantillon:

$S_x^2 = \frac{\sum_{i=1}^{i=n}(x_i-\hat{\mu}_x)^2}{n}$

On va alors chercher à calculer l'espérance de leur différence (c.a.d., ce qui est attendu en moyenne de leur différence).

$$\begin{aligned}
\mathbb{E}[\sigma_x^2 - S_x^2] &= \mathbb{E}[\tfrac{1}{n}\textstyle\sum_{i=1}^{i=n}(x_i-\mu_x)^2 - \tfrac{1}{n}\sum_{i=1}^{i=n}(x_i-\hat{\mu}_x)^2] \\
&= \mathbb{E}[\tfrac{1}{n}\textstyle\sum_{i=1}^{i=n}((x_i^2 - 2\mu_x x_i + \mu_x^2) - (x_i^2 - 2\hat{\mu}_x x_i + \hat{\mu}_x^2))] \\
&= \mathbb{E}[\tfrac{1}{n}\textstyle\sum_{i=1}^{i=n}(\mu_x^2 - \hat{\mu}_x^2 + 2x_i(\hat{\mu}_x - \mu_x))] \\
&= \mathbb{E}[\mu_x^2 - \hat{\mu}_x^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^{i=n}(2x_i(\hat{\mu}_x - \mu_x))] \\
&= \mathbb{E}[\mu_x^2 - \hat{\mu}_x^2 + 2\hat{\mu}_x(\hat{\mu}_x - \mu_x)] \\
&= \mathbb{E}[\mu_x^2 + \hat{\mu}_x^2 - 2\hat{\mu}_x\mu_x] \\
&= \mathbb{E}[(\hat{\mu}_x - \mu_x)^2] \\
&= Var(\hat{\mu}_x)
\end{aligned}$$

On sait que $Var(\hat{\mu}_x) = \frac{\sigma_x^2}{n}$

donc $\mathbb{E}[\sigma_x^2 - S_x^2] = \frac{\sigma_x^2}{n} \Rightarrow \mathbb{E}[\sigma_x^2] - \mathbb{E}[S_x^2] = \frac{\sigma_x^2}{n}$

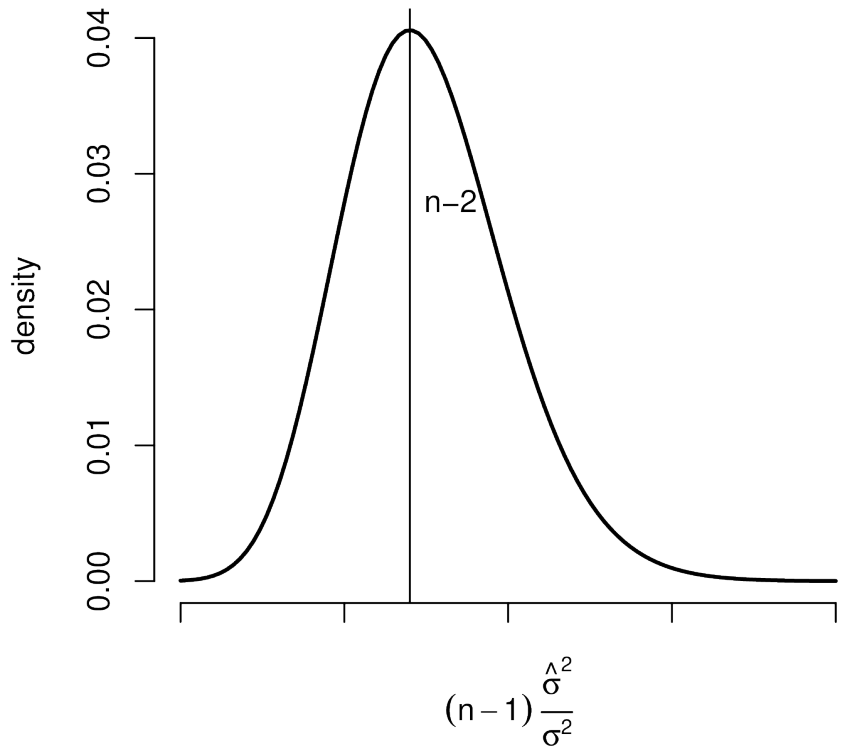
donc $\mathbb{E}[S_x^2] = \sigma_x^2 - \frac{\sigma_x^2}{n} = \frac{n-1}{n}\sigma_x^2$.

Estimation de la variance

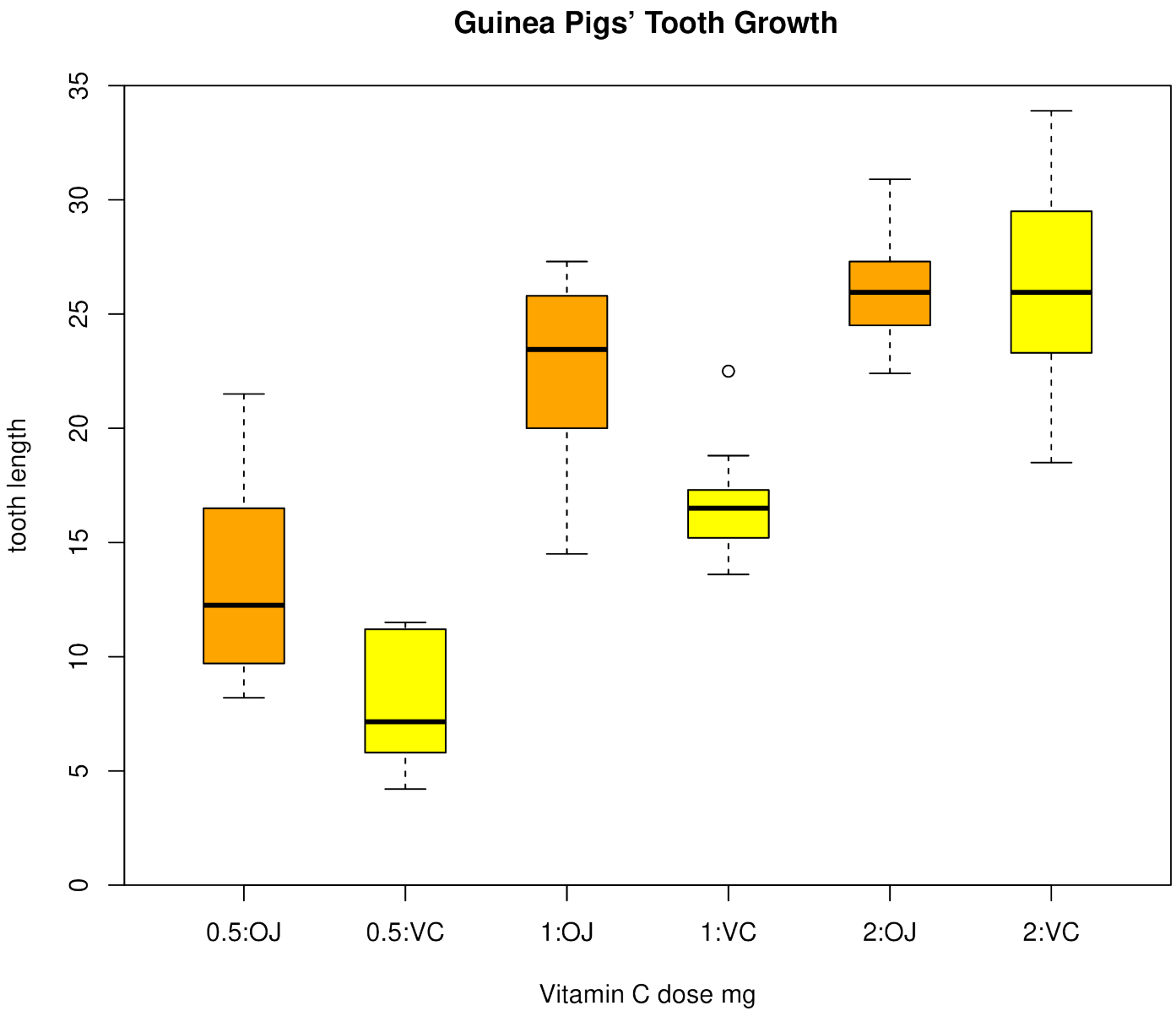
Le comportement de la variance pour un échantillon compte tenu de la variabilité de la population est connue.

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2$$

à $n-1$ degrés de liberté



Boîtes à moustaches et valeurs extrêmes



Loi de probabilité

C'est ce qui permet d'associer les valeurs de X et la probabilité de leur réalisation. Ces lois sont représentées par des fonctions. Dans le cas des variables discrètes, on utilise la fonction de masse qui exprime directement les probabilités $P(X = x)$ avec les valeurs de x.
Soit Ω : l'ensemble des modalités de x.

$$p(X = x) == \begin{cases} 0.5 & \text{si } x \in \Omega \\ 0 & \text{si } x \notin \Omega \end{cases}$$

Si cette loi définit un pîle ou face avec $\Omega = (\text{pile, face})$ et $P(pile) = P(face) = 0.5$, on dit alors que $P(x)$ suit une **loi de Bernouilli** de **paramètre** 0.5.

Fonction de répartition et densité de probabilité

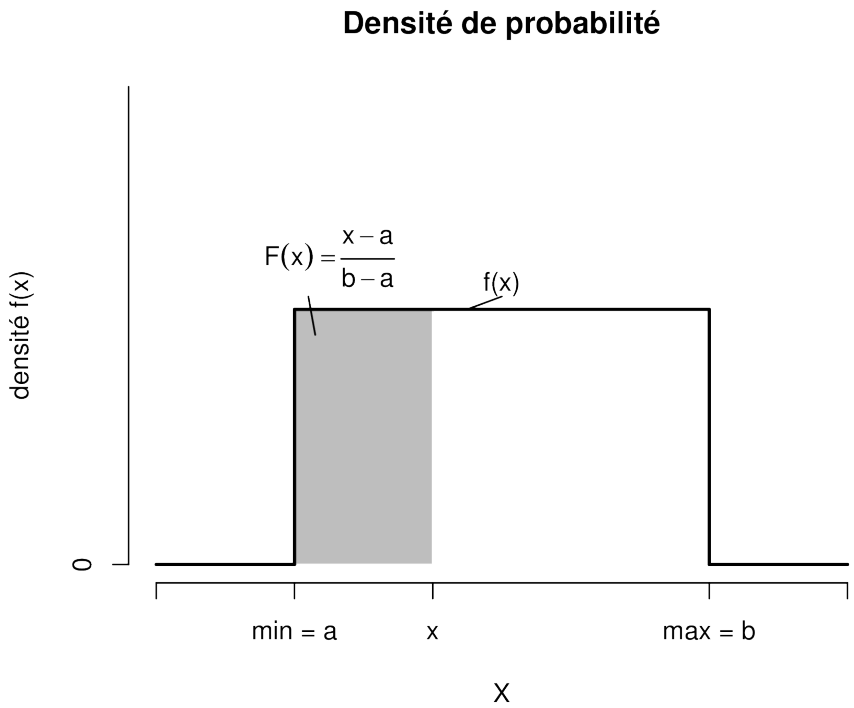
Si X est une variable ordonnée, on peut définir la loi de distribution à partir de **la fonction de répartion** (ou fonction de distribution cumulative) de X comme $F(X) = P(X < x)$. Cela exprime les fréquences cumulées de X en fonction des valeurs prises par X. Le polygone des fréquences cumulées est une représentation de la fonction de répartition. De plus, si X est une variable aléatoire continue définie sur un espace de probabilités tel que $a \leq x \leq b$, on peut définir la loi de probabilité de X par sa **fonction de densité** (ou densité de probabilité). La densité $f(x)$ est la dérivée de a fonction de répartition $F(X)$ et on peut alors écrire:

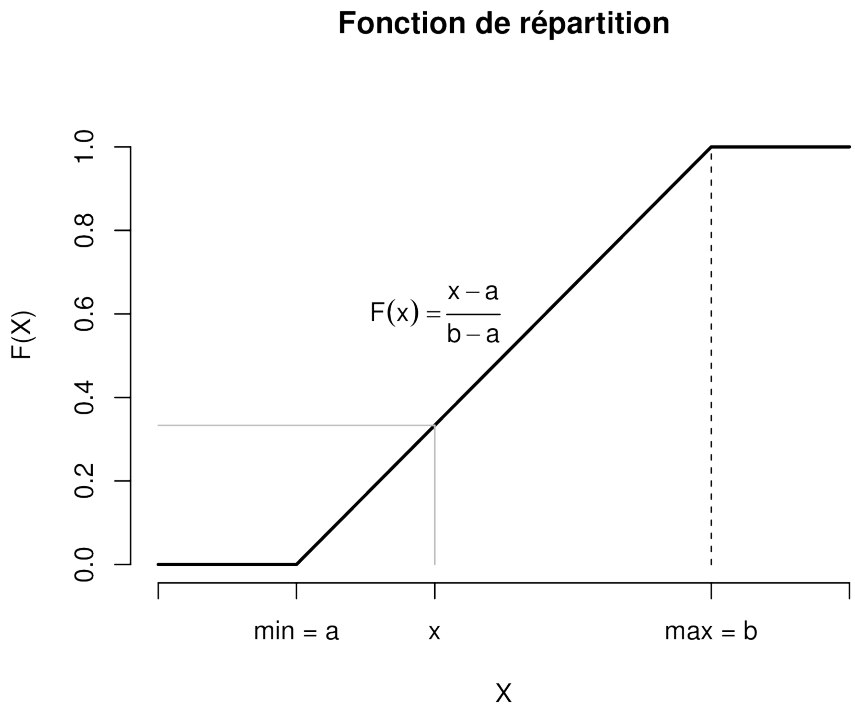
$$P(a \leq x \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

L'aire totale contenue sous la courbe $f(x)$ vaut 1, de sorte que la valeur maximale de $F(X)$ augmente vers 1 quand X augmente.

Exemple: Loi uniforme pour une variable continue

Cette loi est telle que P(X=x) est constante sur l'intervalle [a;b] et qu'elle vaut 0 en dehors de cet intevalle.

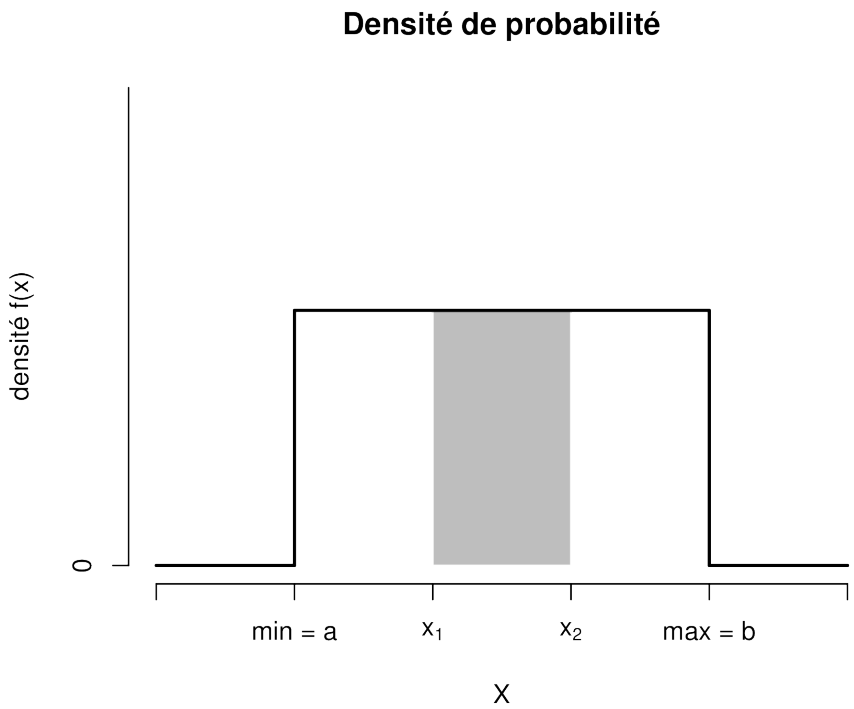


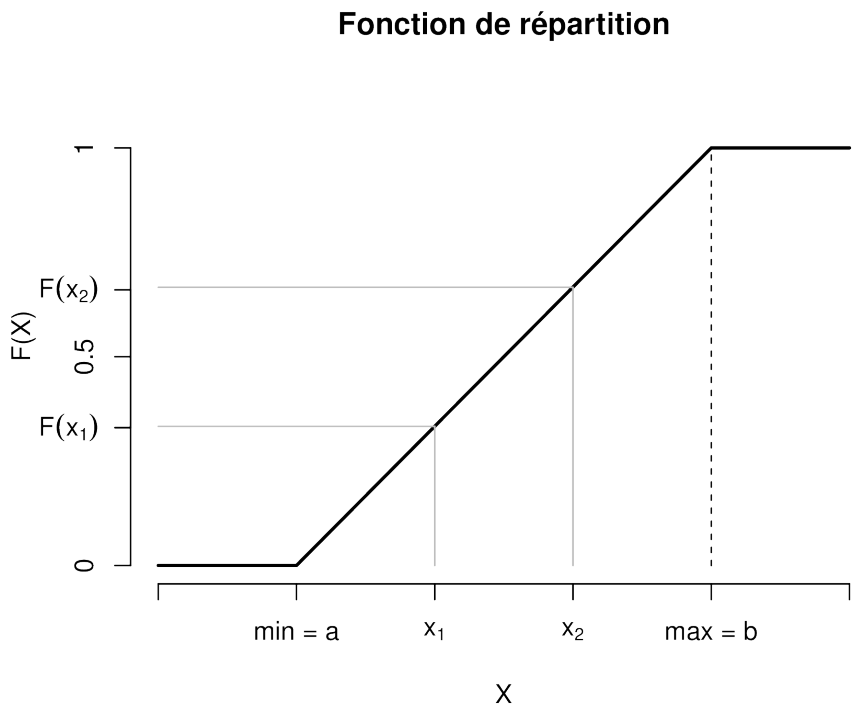


Exemple: Loi uniforme pour une variable continue

A partir de la loi, on peut établir la probabilité pour x d'appartenir à un intervalle donné entre x_1 et x_2.

$$P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$$





Lois utiles en biologie: loi ou événement de Bernouilli

Cette loi de probabilité discrète décrit la probabilité d'une variable binaire prenant les valeurs 0 (échec) ou 1 (succès).

$$P(X = 1) = p; P(X = 0) = q = 1 - p$$

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

$$Var(X) = pq$$

Loi Binomiale

Cette loi discrète correspond au nombre de succès à l'issue de n épreuves de Bernouilli de paramètre p. La fonction de masse est donnée par:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Le coefficient binomial $\binom{n}{k}$ ou C_n^k correspond au nombre de combinaisons de k parmi n.

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

Le triangle de Pascal donne les premières combinaisons.

Pour $n = 0$, $C_0^0 =$ 1

Pour $n = 1$, C_1^0 et $C_1^1 =$ 1 1

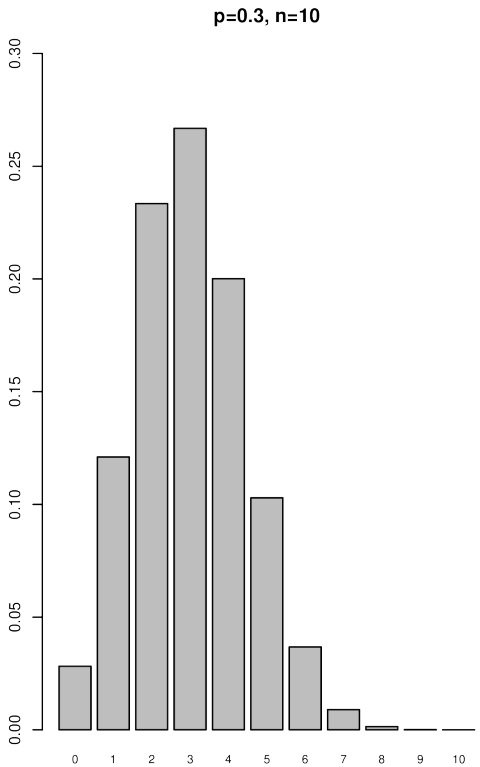
Pour $n = 2$, C_2^0, C_2^1 et $C_2^2 =$ 1 2 1

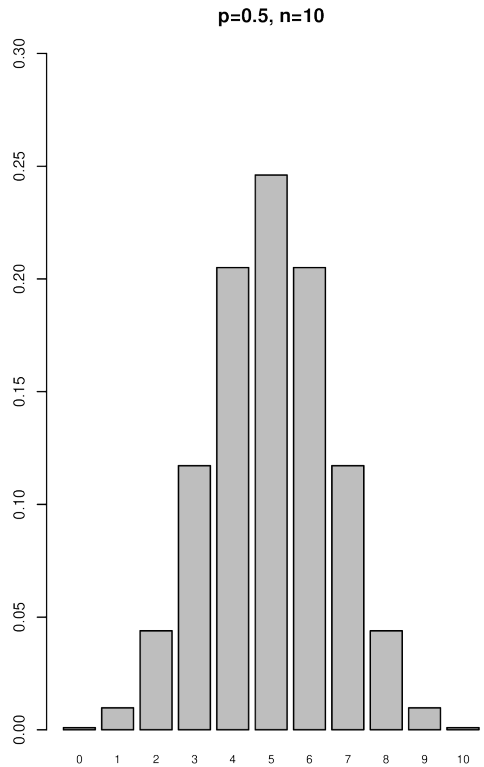
Pour $n = 3$, C_3^0, C_3^1, C_3^2 et $C_3^3 =$ 1 3 3 1

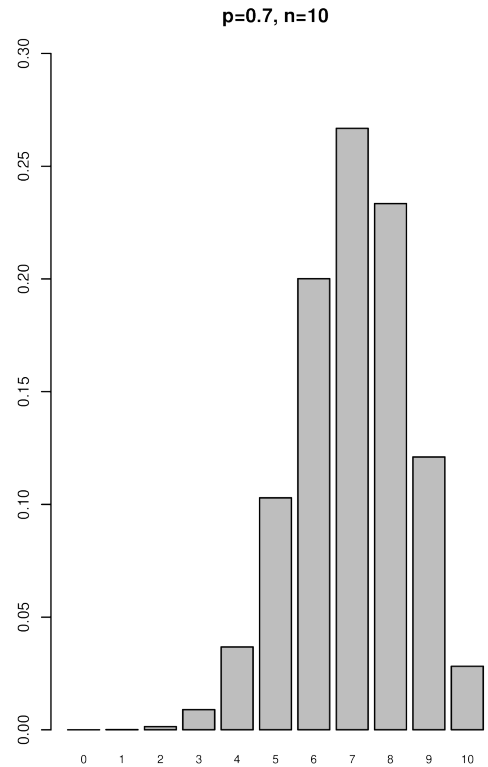
Pour $n = 4$, C_4^0, C_4^1, C_4^2, C_4^3 et $C_4^4 =$ 1 4 6 4 1

etc...

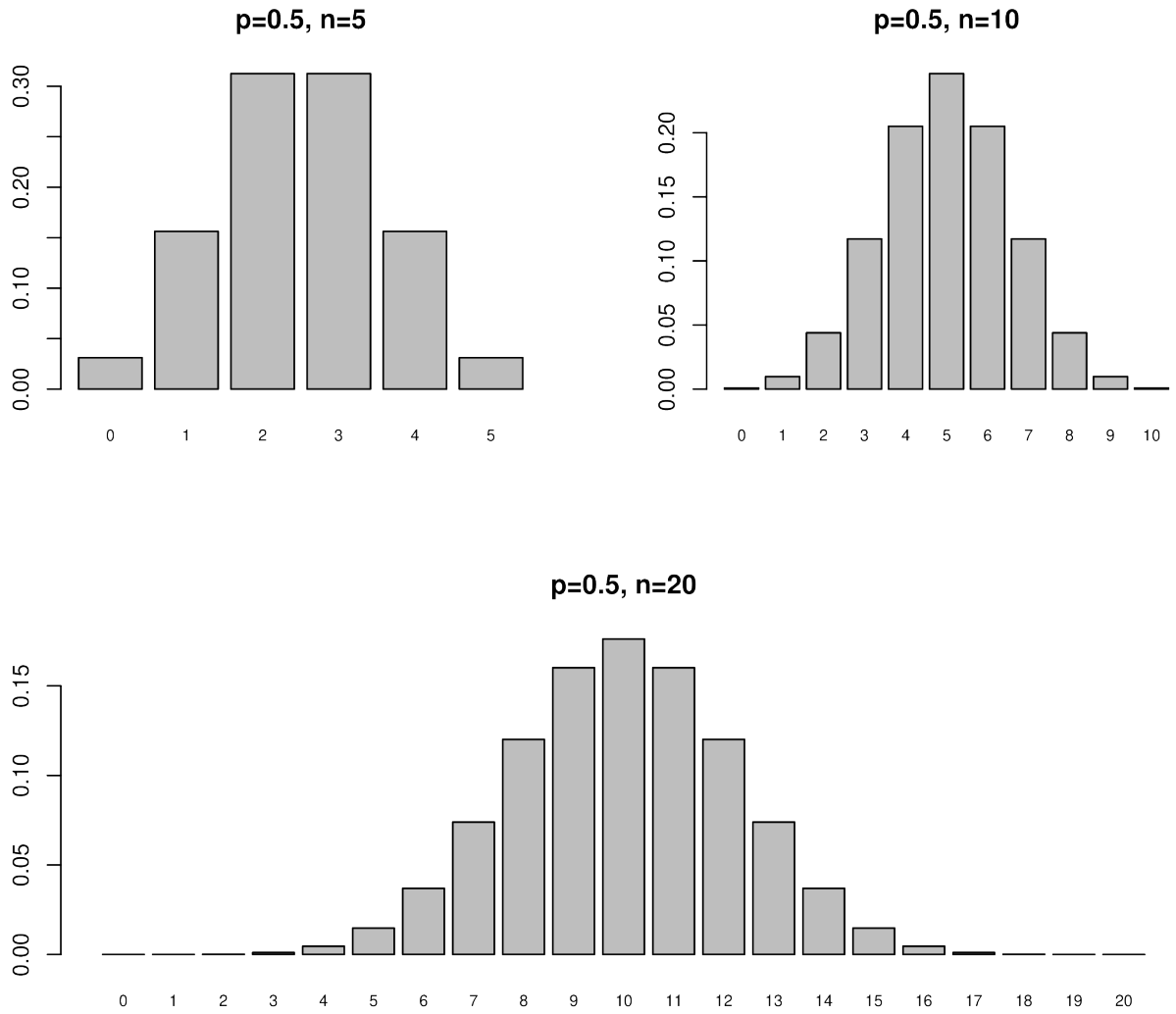
Loi Binomiale: exemple de fonctions de masse



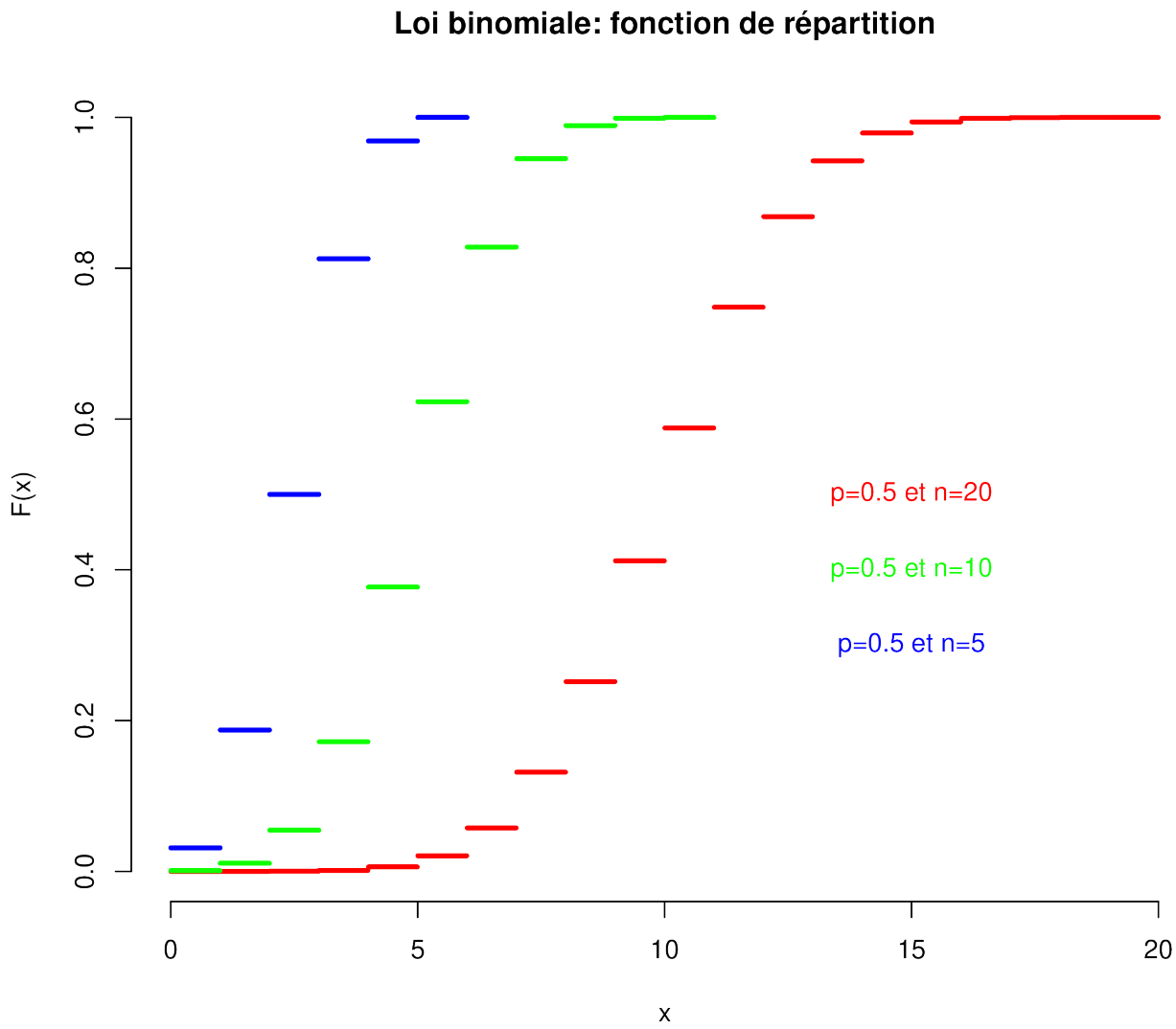




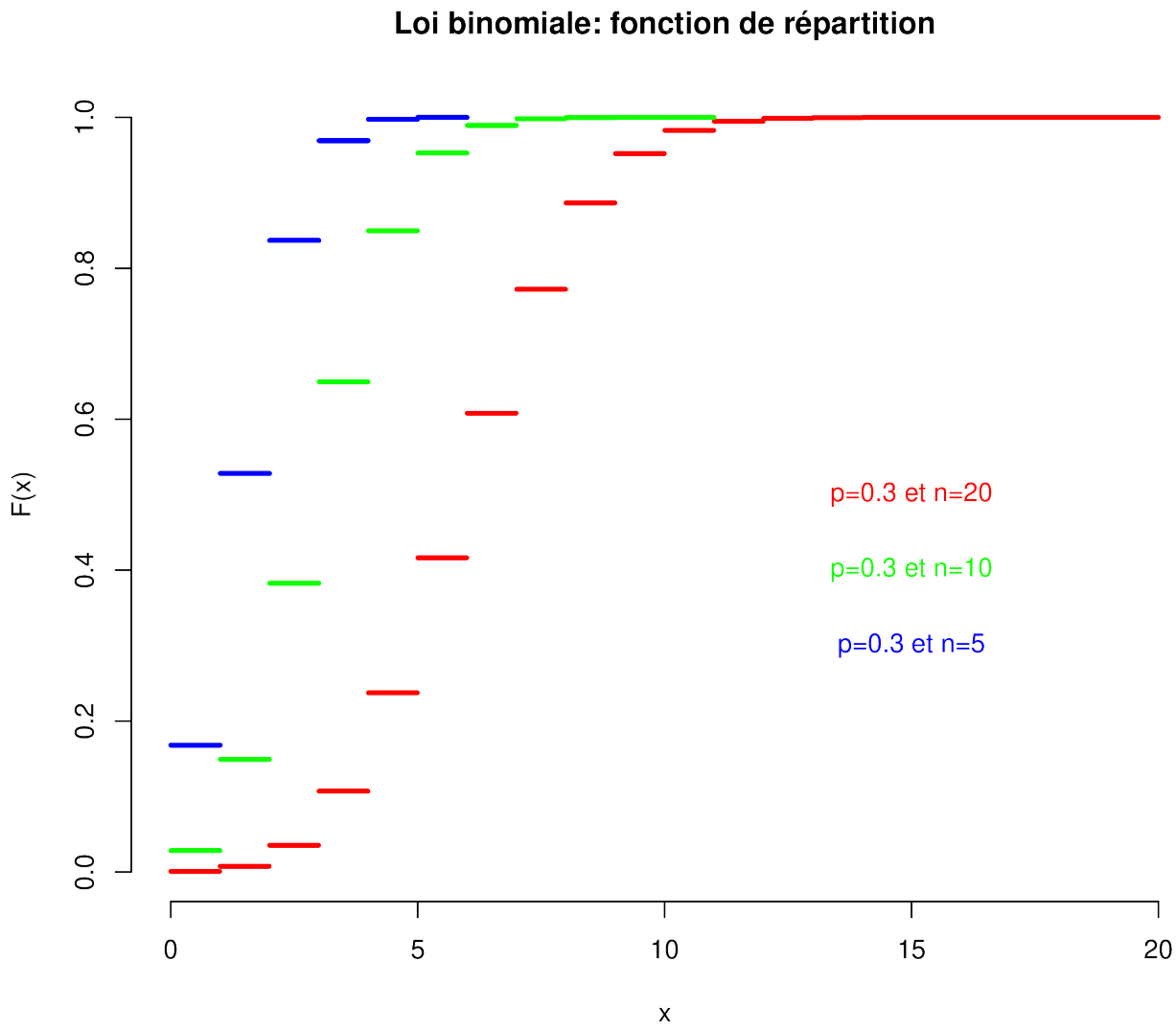
Loi Binomiale: exemple de fonctions de masse



Loi Binomiale: exemple de fonctions de répartition



Loi Binomiale: exemple de fonctions de répartition



Loi Binomiale: exemple d'application

On cherche à construire ici un intervalle de confiance concernant un nombre de succès et d'échecs (identique à construction d'un intervalle de confiance pour une proportion). Dans cet exemple, on cherche à évaluer si un poison X ou Y est plus efficace. Pour cela on mesure la survie de 10 individus avec X, et 10 avec Y. On trouve 5 individus mort avec X, 7 avec Y.

La question est alors de savoir si les valeurs de survie avec ces échantillons peuvent être représentatives de la population. On va donc calculer les probabilités associées à l'ensemble des issues possibles de cette expérience si $p = 0.5$ pour le poison X, et $p = 0.7$ pour le poison Y.

Loi Binomiale: exemple d'application

Calcul des Coefficients binomiaux:

n morts	0	1	2	3	4	5	6	7	8	9	10
Comb	C_{10}^0	C_{10}^1	C_{10}^2	C_{10}^3	C_{10}^4	C_{10}^5	C_{10}^6	C_{10}^7	C_{10}^8	C_{10}^9	C_{10}^{10}
C_n^k	1	10	45	120	210	252	210	120	45	10	1

Probabilité (Fonctions de masse) pour $\hat{p} = 0.5$ et $\hat{p} = 0.7$

n morts	0	1	2	3	4	5	6	7	8	9	10
$P(X = k\|\hat{p} = 0.5)$	0.001	0.01	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.01	0.001
$P(X = k\|\hat{p} = 0.7)$	0.000	0.000	0.001	0.01	0.07	0.103	0.200	0.267	0.233	0.121	0.001

Pour que la valeur prise par un échantillon soit représentative, on cherche à l'encadrer de sorte de contenir 95% de sa distribution attendue. Si $p = 0.5$, pour un échantillon de $n = 10$, alors au moins 95% de la distribution de X devrait être comprise entre 2 et 8 morts. Si $p = 0.7$, pour un échantillon de $n = 10$, alors au moins 95% de la distribution de X devrait être comprise entre 4 et 9 morts. On prendrait donc un risque d'au moins 5% de se tromper en déclarant que l'efficacité des poisons diffère car les échantillons ne sont peut être pas représentatifs.

Loi Binomiale: Espérance et variance

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

L'espérance et la variance variable X suivant la loi Binomiale avec les paramètres p et q sont:

$$\mathbb{E}(X) = np$$

$$Var(X) = npq$$

Loi Géométrique

C'est la loi qui permet de savoir quand dans k tirages doit advenir le premier succès. La loi de masse est donnée par

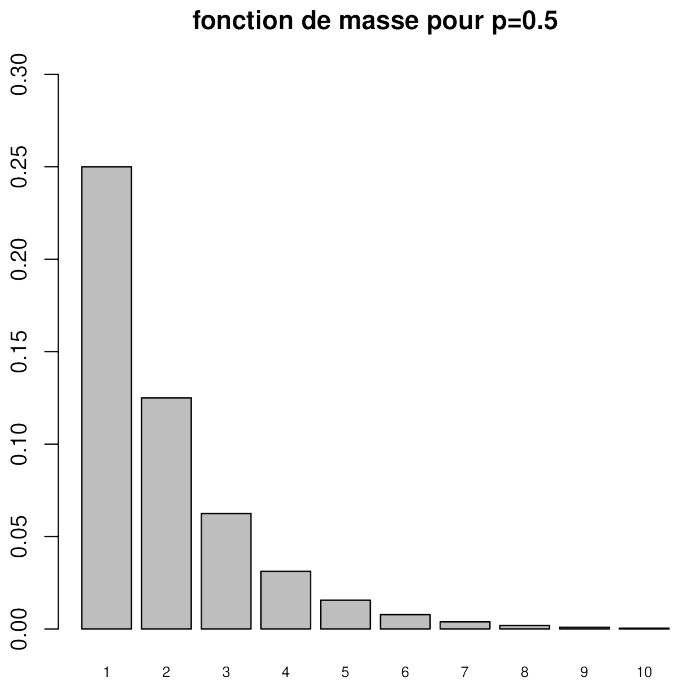
$$P(X = k) = p(1-p)^{k-1}$$

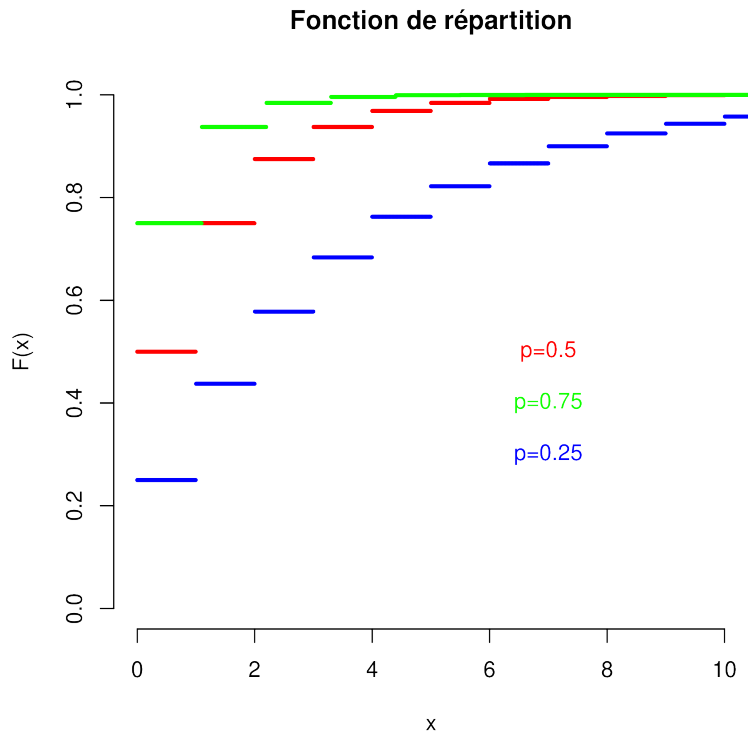
L'espérance et la variance de cette loi dépendent de p.

$$\mathbb{E}(X) = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2} = \frac{q}{p^2}$$

Loi Géométrique: fonction de masse et de répartion





Loi de Poisson

Loi discrète qui caractérise la probalité d'un nombre k d'événements qui se produisent dans un temps ou un espace fixé et de paramètres λ; ce paramètre étant le nombre moyen d'évènements dans le temps ou l'espace. La loi de masse est donnée par :

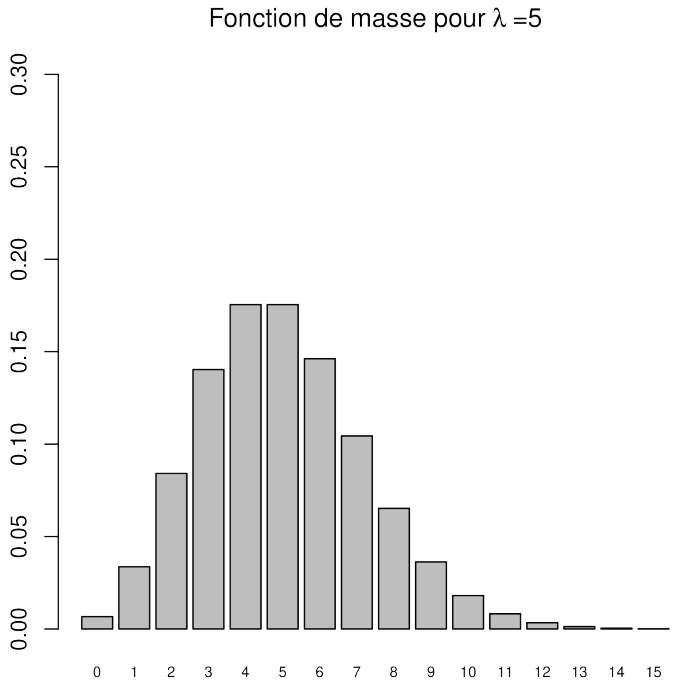
$$P(X = k) = \frac{\lambda^k}{k!} \exp^{-\lambda}$$

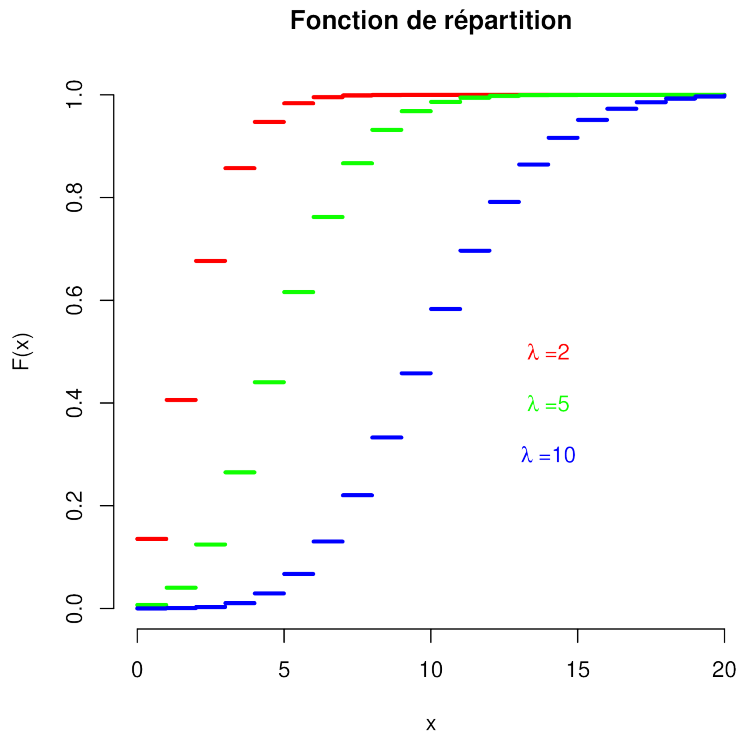
L'espérance et la variance de cette loi dépendent de l'unique paramètre λ.

$$\mathbb{E}(X) = \lambda$$

$$Var(X) = \lambda$$

Loi de Poisson: fonction de masse et de répartion





Loi Binomiale négative

Cette loi discrète traduit le nombre k d'échecs nécessaires jusqu'à ce que n succès se produisent sachant que p est la probabilité d'un succès et $q = 1 - p$ est la probabilité d'un échec.
La loi de masse est donnée par

$$P(X = k) = \binom{k + n - 1}{n - 1}(p)^n(q)^k$$

L'espérance et la variance de cette loi dépendent de p.

$$\mathbb{E}(X) = \frac{nq}{p}$$

$$Var(X) = \frac{nq}{p^2}$$

Retour sur le TCL: Loi normale

D'après ce théorème:

$$\frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{(n)}} \sim \mathcal{N}(0, 1)$$

La différence entre les moyennes des échantillons de taille n et la moyenne de la population, divisée par l'erreur type tend vers la Loi normale centrée réduite quand n est grand. Cela est très utile pour prédire le comportement des moyennes que l'on calculerait sur un échantillon. Cette propriété peut être utilisée pour avoir une idée de la variation du calcul de la moyenne à partir d'un échantillon de taille donnée. La loi normale permet aussi de modéliser le comportement d'une suite d'évènements lorsque le nombre d'essais est grand.
Quand un paramètre mesuré en biologie est déterminé par de très nombreux déterminismes, sa distribution dans une population a très souvent une distribution normale (e.g. la taille corporelle).

Loi normale: fonction de densité

Cas général:

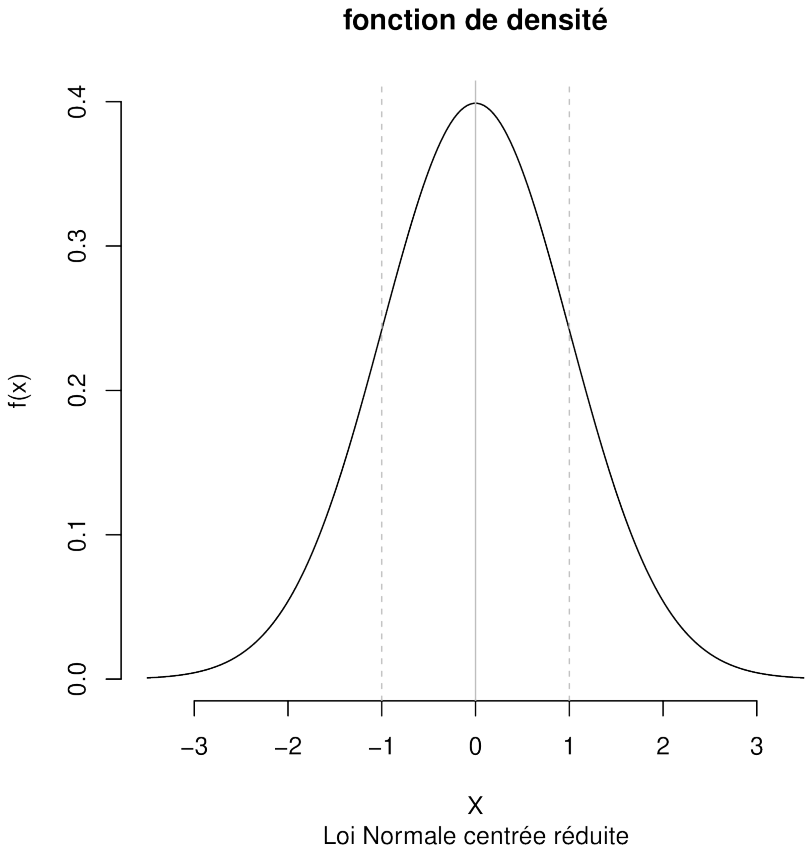
$$P(x) = f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

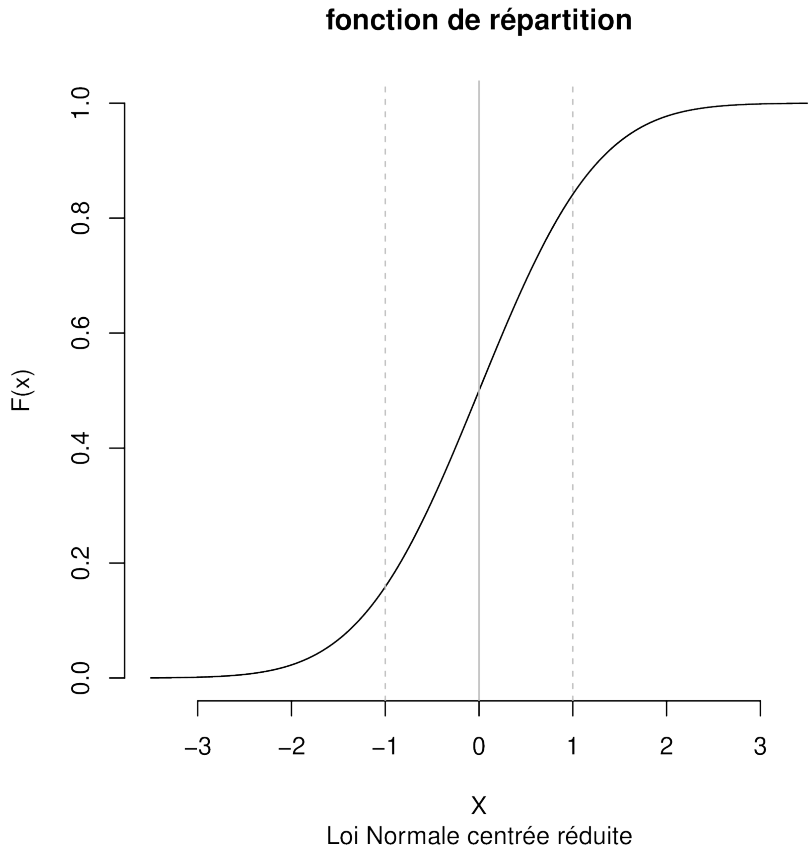
Loi Normale centrée réduite: dans ce cas $\mu = 0$ et $\sigma^2 = 1$

$$P(x) = f(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x)^2}{2}}$$

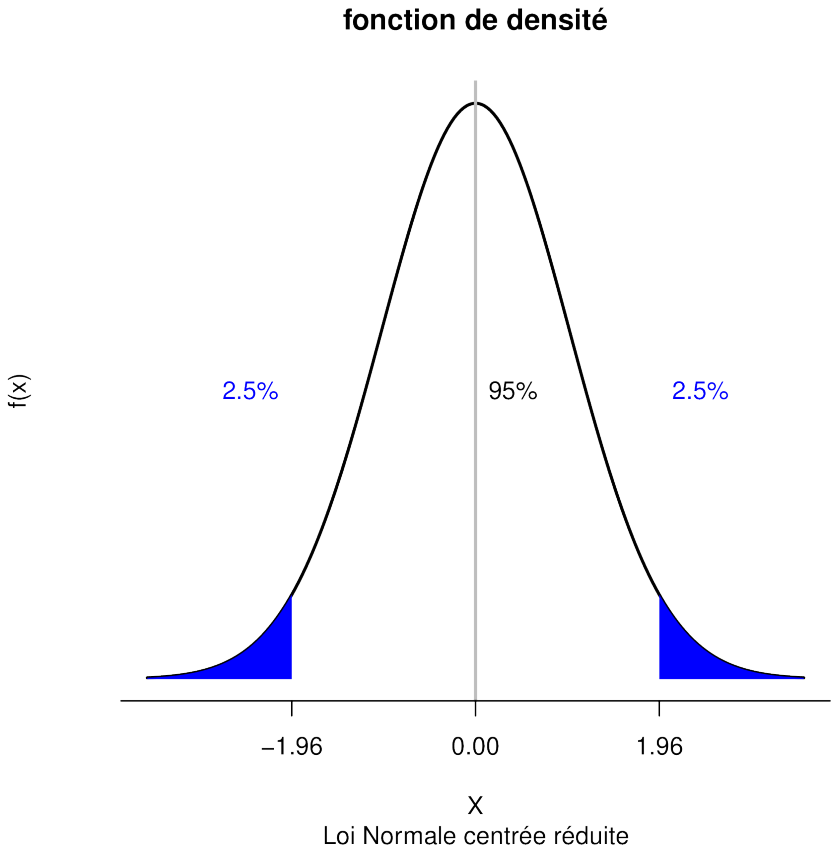
Pour passer de $\mathcal{N}(0, 1)$ à $\mathcal{N}(\mu, \sigma)$, on multiplie les quantiles de la loi normale centrée réduite (valeurs X dans $\mathcal{N}(0, 1)$) par l'écart type σ et on ajoute à cette quantité la moyenne μ.

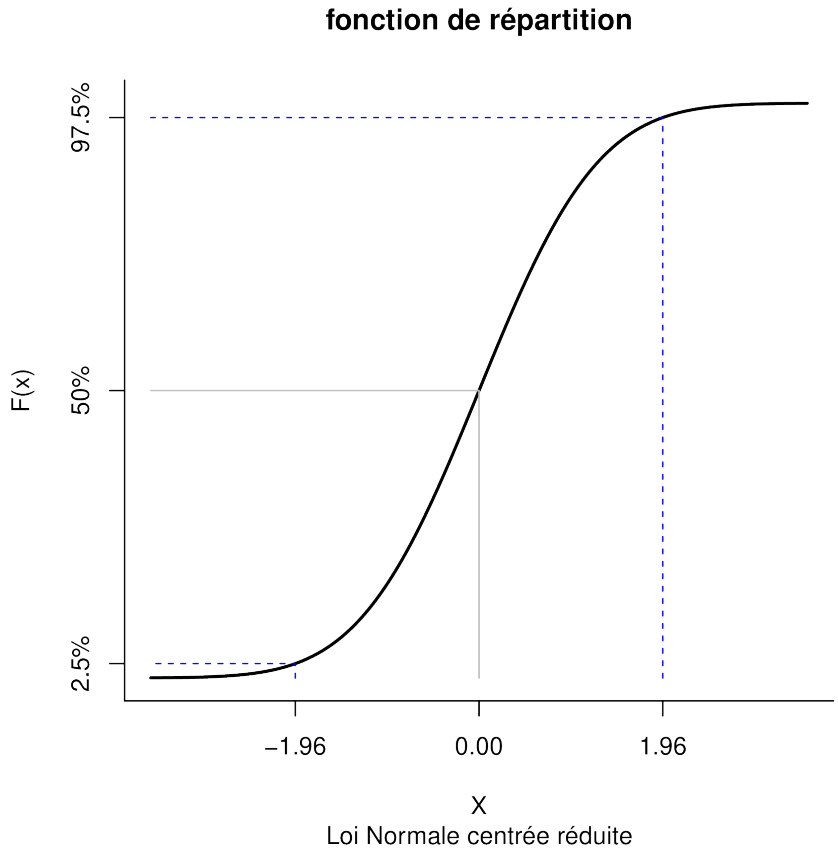
Loi normale centrée réduite



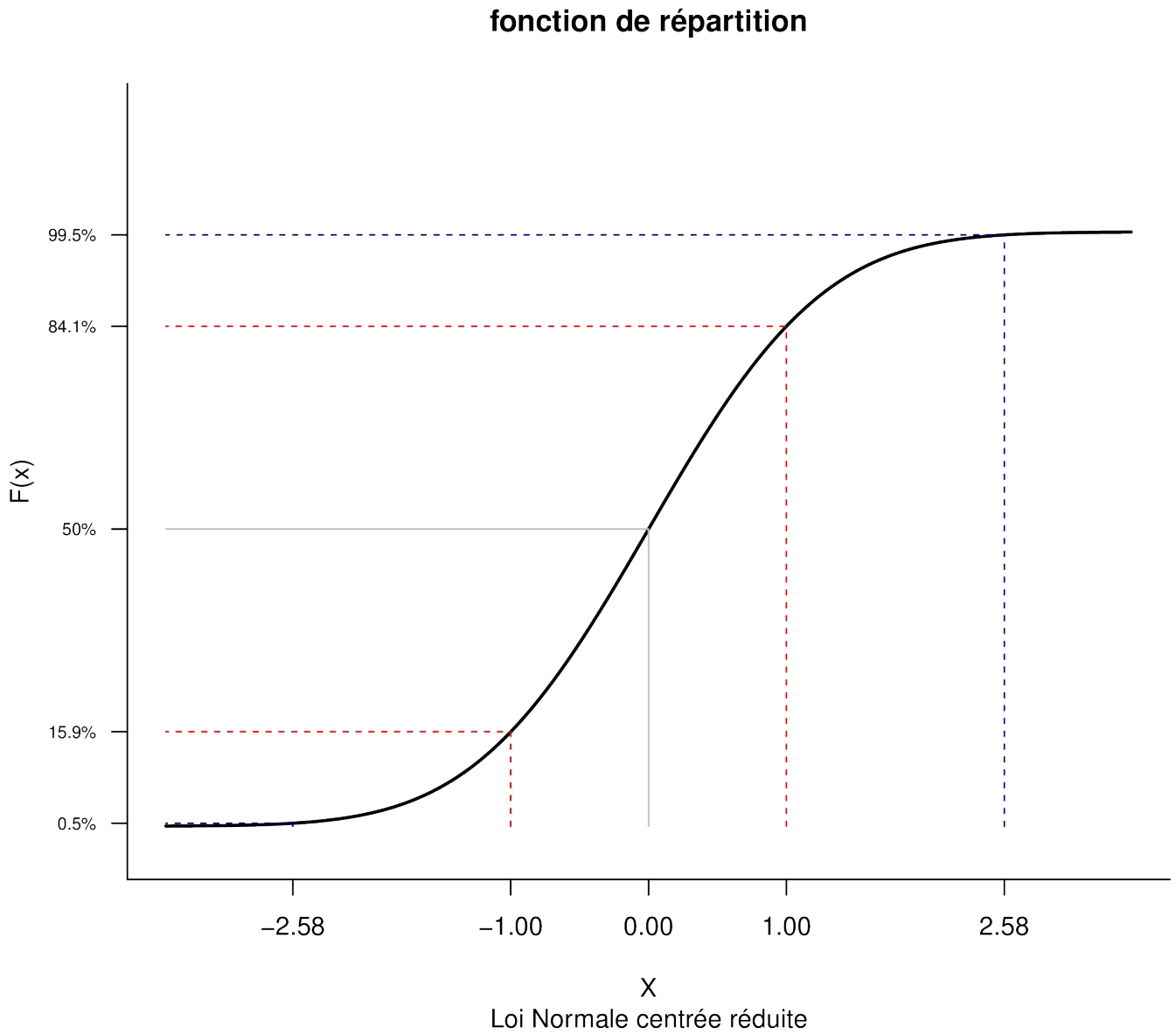


Loi normale centrée réduite

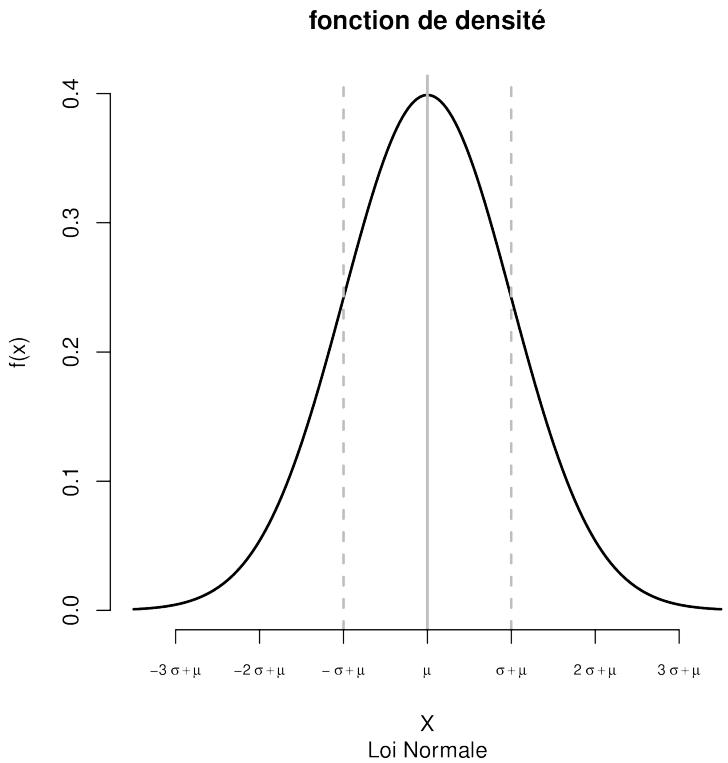


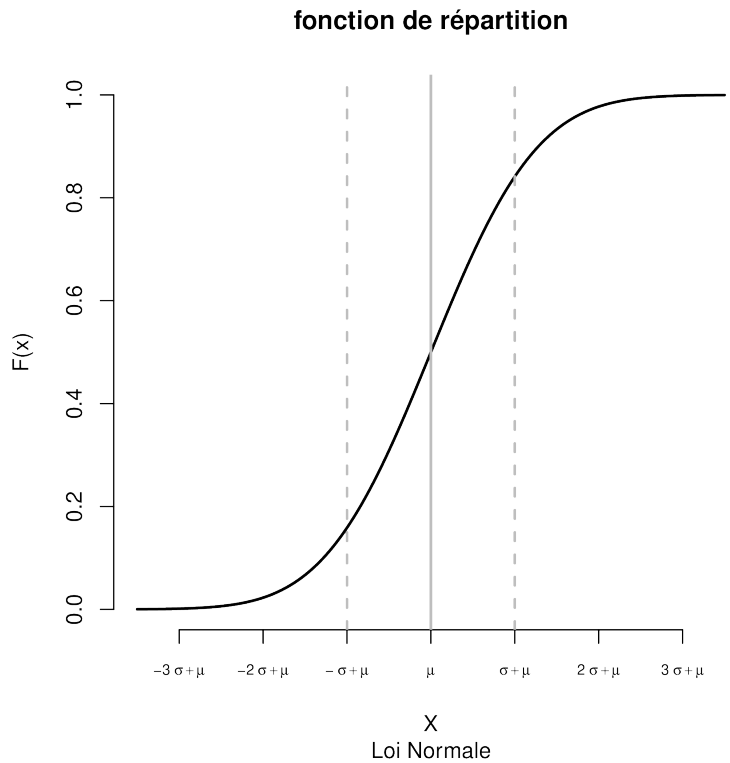


Loi normale centrée réduite



Loi Normale: recherche de quantile et de probabilité





Pour passer de $\mathcal{N}(0,1)$ à $\mathcal{N}(\mu,\sigma)$, si on connait les probabilités associées, on multiplie les quantiles de la loi normale centrée réduite (valeurs X dans $\mathcal{N}(0,1)$) par l'écart type σ et on ajoute à cette quantité la moyenne μ.

Application: intervalle de confiance sur la moyenne

D'après le théorème central limite:

$$Z = \frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{(n)}} \sim \mathcal{N}(0, 1)$$

Comme on connait les probabilités associées aux quantiles de la loi normale centrée réduite, on peut en calculant $\hat{\mu}$ et en estimant σ construire un intervalle de confiance sur la moyenne pour peu que n soit assez grand. En particulier 2.5% de Z sera inférieur à 1.96, et 97.5% de Z sera supérieur à 1.96.

Application: intervalle de confiance sur la moyenne

Ainsi 95% de la distribution de z est comprise entre -1.96, et 1.96 les quantiles à 2.5% et 97.5% de la loi normale centrée réduite. :

$$-1.96 < \frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{(n)}} < 1.96 \quad \text{ce qui équivaut à} \quad -1.96 < \frac{(\mu - \hat{\mu})}{\sigma/\sqrt{(n)}} < 1.96$$

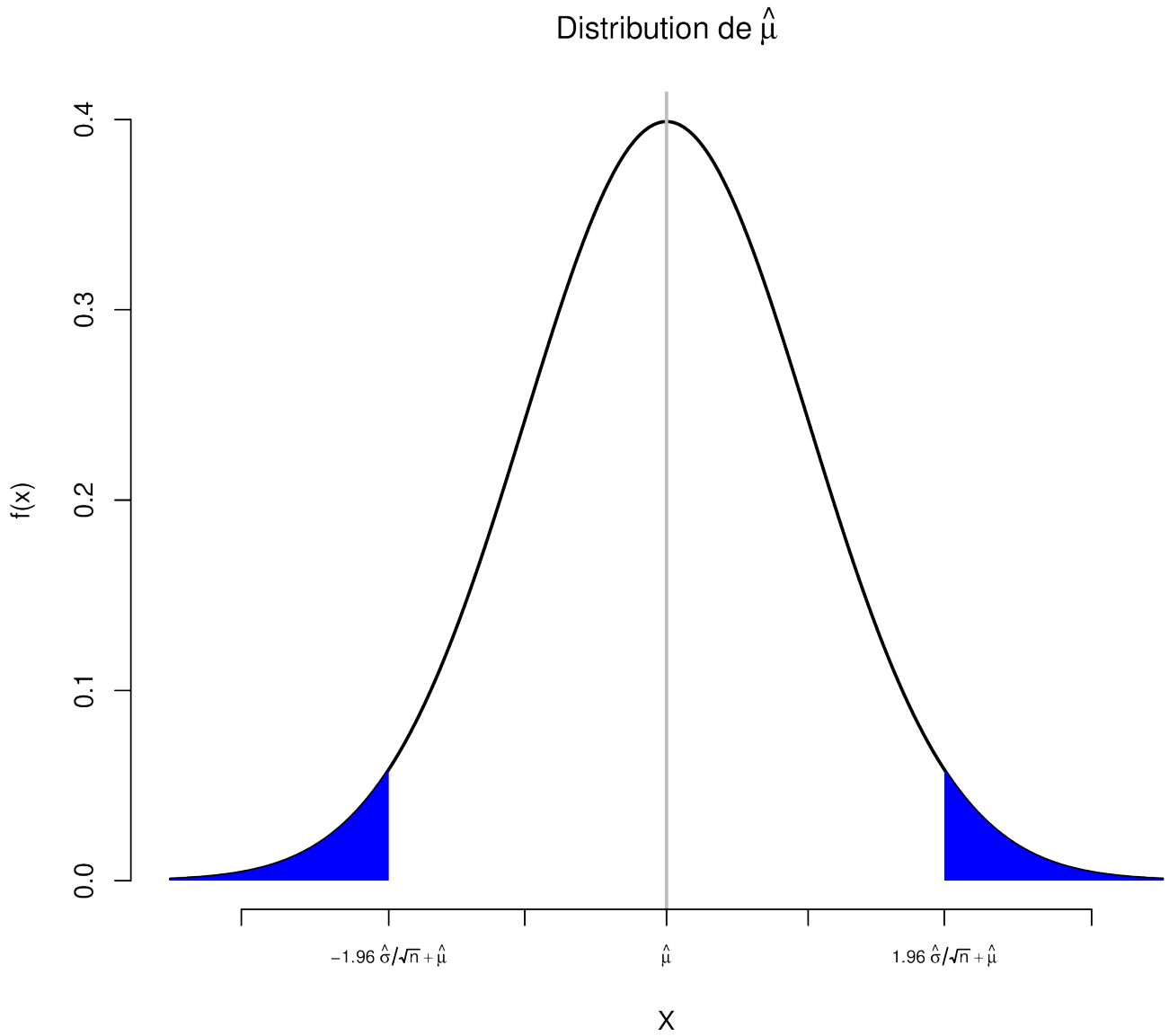
L'intervalle de confiance à 95% est alors donné par

$$[\hat{\mu} - 1.96 * \hat{\sigma}/\sqrt{(n)}; \hat{\mu} + 1.96 * \hat{\sigma}/\sqrt{(n)}]$$

De manière générale, si on veut faire un intervalle de conviance à $1 - \alpha \times 100\%$, alors on recherchera les quantiles z_2 et z_1 $\alpha/2$ et $1 - \alpha/2 \times 100\%$. La Loi normale centrée réduite est symétrique et $z_1 = -z_2$. On peut alors écrire:

$$[\hat{\mu} - z_1 * \hat{\sigma}/\sqrt{(n)}; \hat{\mu} + z_1 * \hat{\sigma}/\sqrt{(n)}]$$

Application: intervalle de confiance sur la moyenne

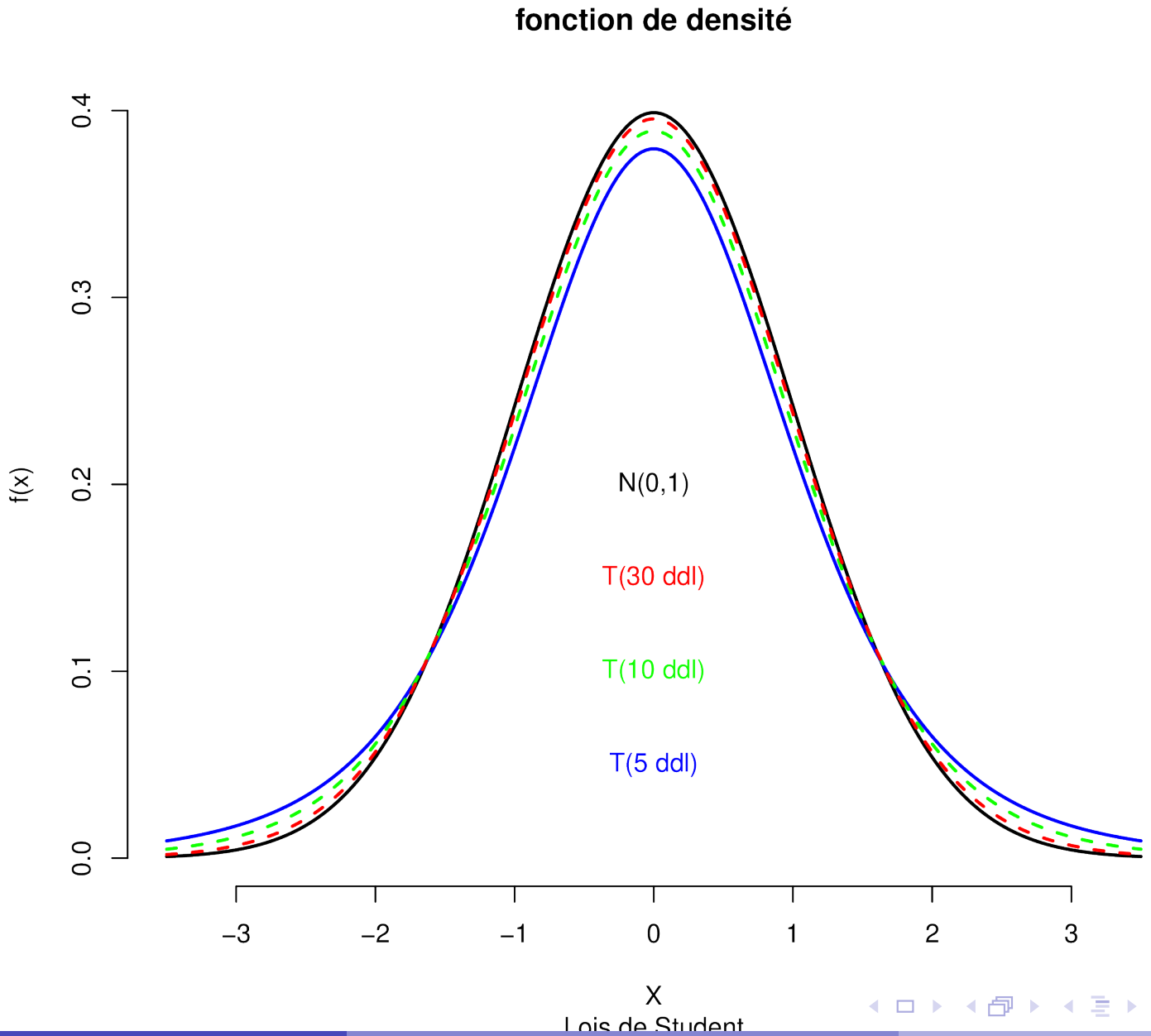


La Loi de student et les petits échantillons

Si les échantillons sont de plus petite taille alors, l'imprécision sur l'estimation de la moyenne peut augmenter et on peut alors s'écarter du théorème centrale limite. La Loi de Student donne l'association des quantiles et des probabilités pour des petits échantillons de moyenne théorique 0, et d'écart type 1. Elle est symétrique et converge vers la loi normale pour les grands échantillons. Elle n'est définie que par un paramètre portant le nom de nombre de degré de liberté et dépendant de l'effectif. Pour un estimateur comme la moyenne le nombre de degrés de liberté vaut $n-1$. On préfèrera alors les quantiles t de cette Loi à $n-1$ degrés de liberté pour construire un intervalle de confiance.

$$[\hat{\mu} - t_1 * \hat{\sigma}/\sqrt{(n)}; \hat{\mu} + t_1 * \hat{\sigma}/\sqrt{(n)}]$$

La Loi de student: fonctions de densité



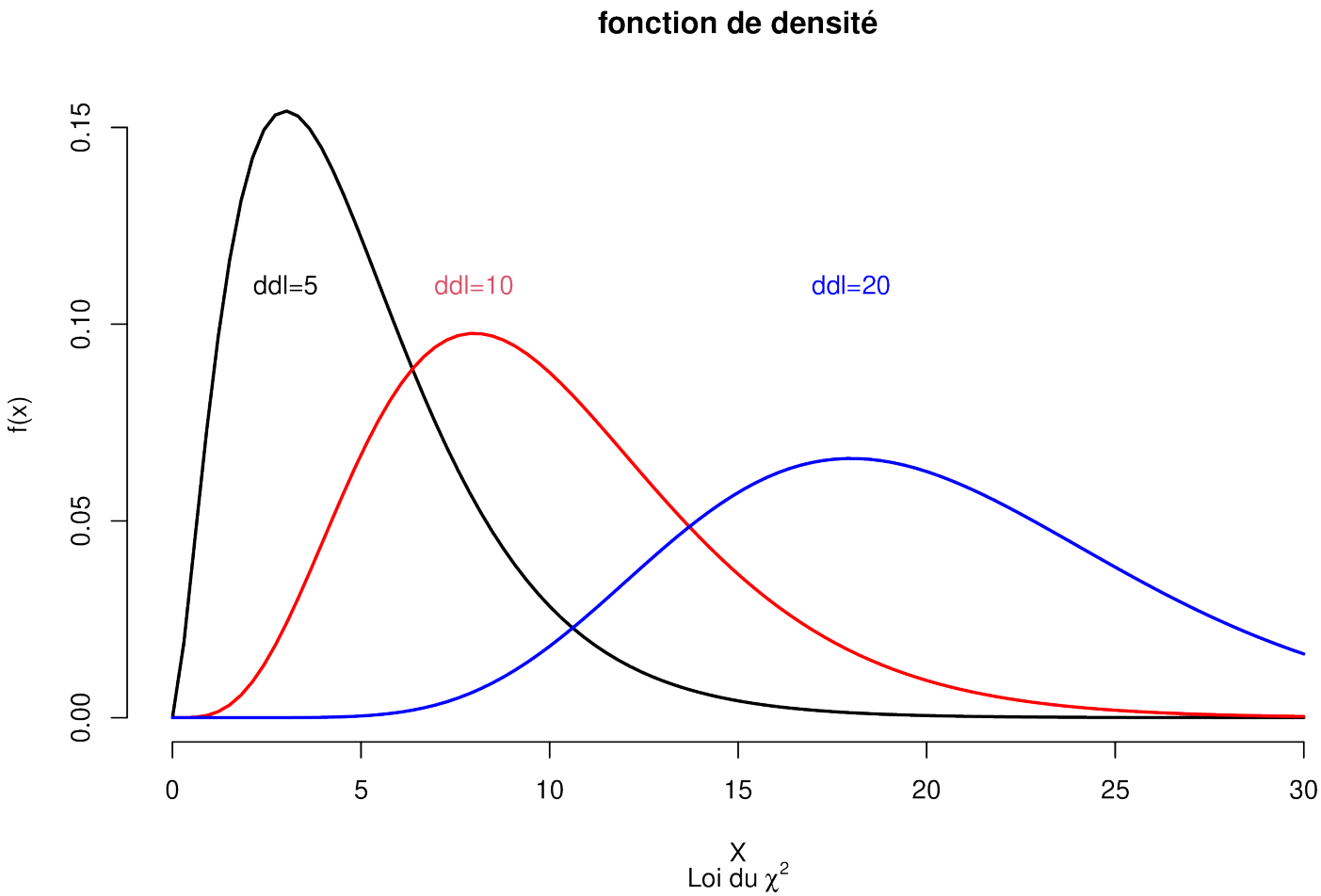
La Loi de student: Quelques quantiles remarquables

Tableau des correspondances entre quantiles x et $P(X < x)$

Loi	0.005	0.01	0.025	0.05	0
N(0,1)	-2.58	-2.33	-1.96	-1.64	0
T(50 ddl)	-2.68	-2.40	-2.00	-1.68	0
T(30 ddl)	-2.74	-2.48	-2.04	-1.70	0
T(10 ddl)	-3.16	-2.76	-2.23	-1.81	0
T(5 ddl)	-4.03	-3.36	-2.57	-2.02	0

La Loi du Chi2 et intervalle de confiance sur la variance

La quantité $(n-1)\frac{\hat{\sigma}^2}{\sigma^2}$ suit une loi de distribution du χ^2 à $n-1$ degrés de liberté.



La Loi du Chi2 et intervalle de confiance sur la variance

Comme la quantité $(n-1)\frac{\hat{\sigma}^2}{\sigma^2}$ suit une loi de distribution du χ^2 à $n-1$ degrés de liberté, on va chercher à encadrer σ^2 à l'aide des quantiles à $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ qu'on définira respectivement par χ_1^2 et χ_2^2 à $n-1$ degrés de liberté. L'intervalle est donné par:

$$[(n-1)\frac{\hat{\sigma}^2}{\chi_2^2}, (n-1)\frac{\hat{\sigma}^2}{\chi_1^2}]$$

L'intervalle de confiance étant donné à $1-\alpha \times 100\%$. Attention, la loi n'est pas symétrique, il faut aller chercher les deux quantiles extrêmes.

Intervalle de confiance pour une proportion

D'après le théorème centrale limite, la plupart des lois peuvent être approchées par une loi normale sous condition d'indépendance. Par exemple, sachant que la variable binomiale est bien une somme de variables indépendantes (de Bernoulli). On sait qu'une loi $\mathcal{B}(n; p)$ a pour espérance np et pour variance $np \times (1 - p)$. Donc

$$\mathcal{B}(n, p) \sim \mathcal{N}(np, \sqrt{np(1-p)})$$

Le demi intervalle de confiance est approximé par $z_1 \times \frac{\sigma}{\sqrt{n}}$ quand n était grand. Pour une proportion, le demi intervalle devient $z_1 \times \sqrt{\frac{pq}{n}} = z_1 \times \sqrt{\frac{p(1-p)}{n}}$, et l'intervalle est donné par:

$$[\hat{p} - z_1 \times \sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_1 \times \sqrt{\frac{p(1-p)}{n}}]$$

avec z_1 correspondant au quantile à $1 - \frac{\alpha}{2} \times 100$ pourcents de $\mathcal{N}(0, 1)$.

Taille nécessaire d'un échantillon pour une précision donnée

Cas d'une proportion

On a pour objectif que le demi écart a_α de confiance pour une proportion soit inférieur à une valeur λ. On sait que le demi intervalle de confiance vaut: $z_\alpha \sqrt{\frac{p(1-p)}{n}}$.

$\Rightarrow z_\alpha \sqrt{\frac{p(1-p)}{n}} \leq \lambda$

$\Rightarrow z_\alpha^2 \frac{p(1-p)}{n} \leq \lambda^2$

$\Rightarrow n \geq z_\alpha^2 \frac{p(1-p)}{\lambda^2}$

avec z_α le quantile de la loi normale centrée réduite correspondant à la probabilité associée à un intervalle de confiance à $1 - \alpha \times 100$ pourcents.

Taille nécessaire d'un échantillon pour une précision donnée

Cas de la moyenne

Soit a_α la demi longueur de l'intervalle de confiance:

On veut $a_\alpha < \lambda$ donc

$t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} < \lambda$

$\Rightarrow t_\alpha^2 \frac{\hat{\sigma}^2}{\lambda^2} < n$

avec t_α, le quantile de la loi de Student correspondant à la probabilité associée à un intervalle de confiance à $1 - \alpha \times 100$ pourcents.

Principe d'un test en statistique: un exemple

Test de comparaison de moyenne

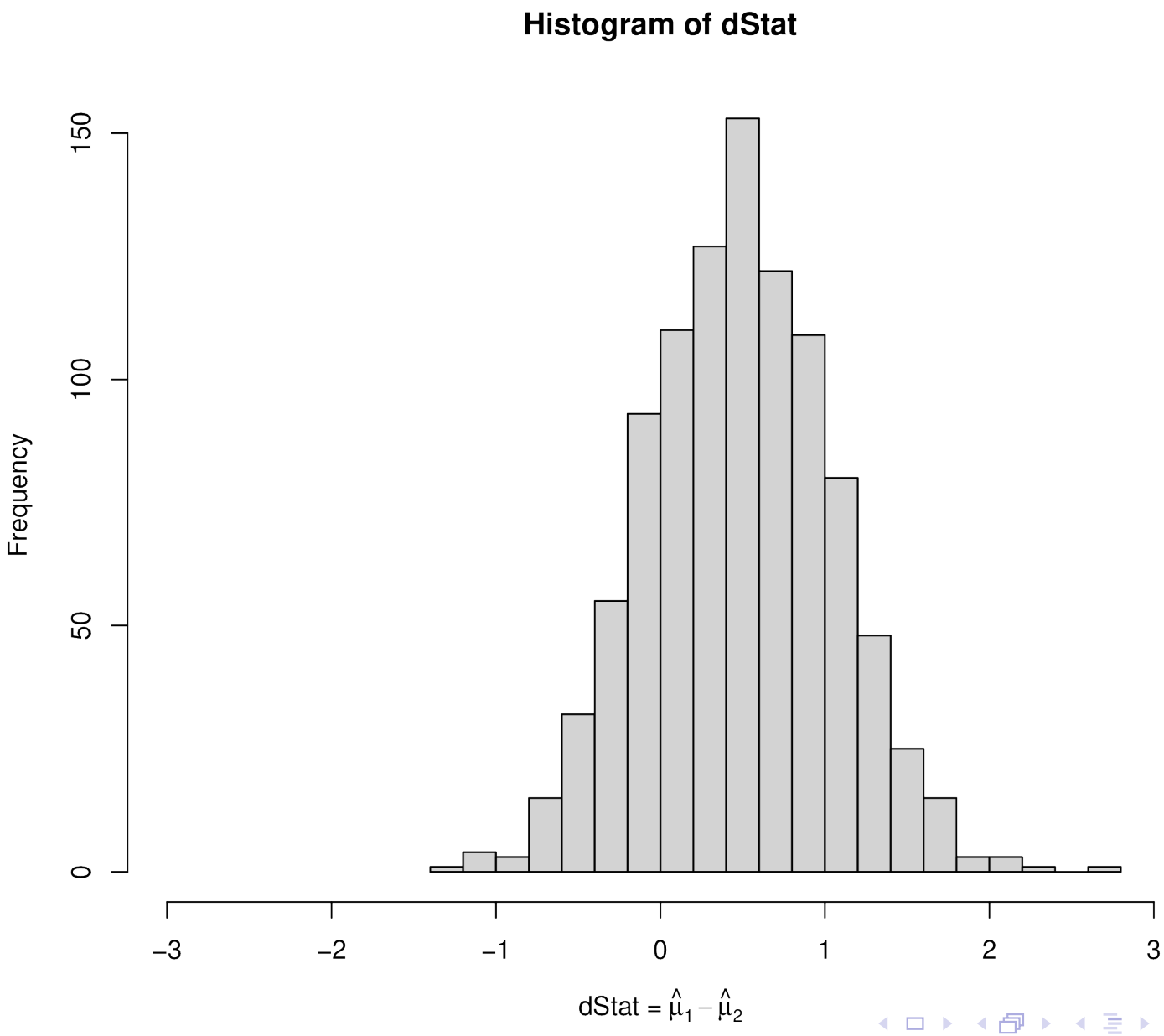
Soient deux échantillons s_1 et s_2 d'effectifs n_1 et n_2, on veut savoir si il existe une différence de moyenne entre les deux populations pop_1 et pop_1 qu'ils représentent.

On sait que du fait d'erreurs liés à l'échantillonage, il est peu probable que $\hat{\mu}_1 = \mu_1$. On sait par exemple que plus l'échantillon est petit, plus les erreurs d'échantillonages sont grandes, et plus la dispersion de $\hat{\mu}$ est importante.

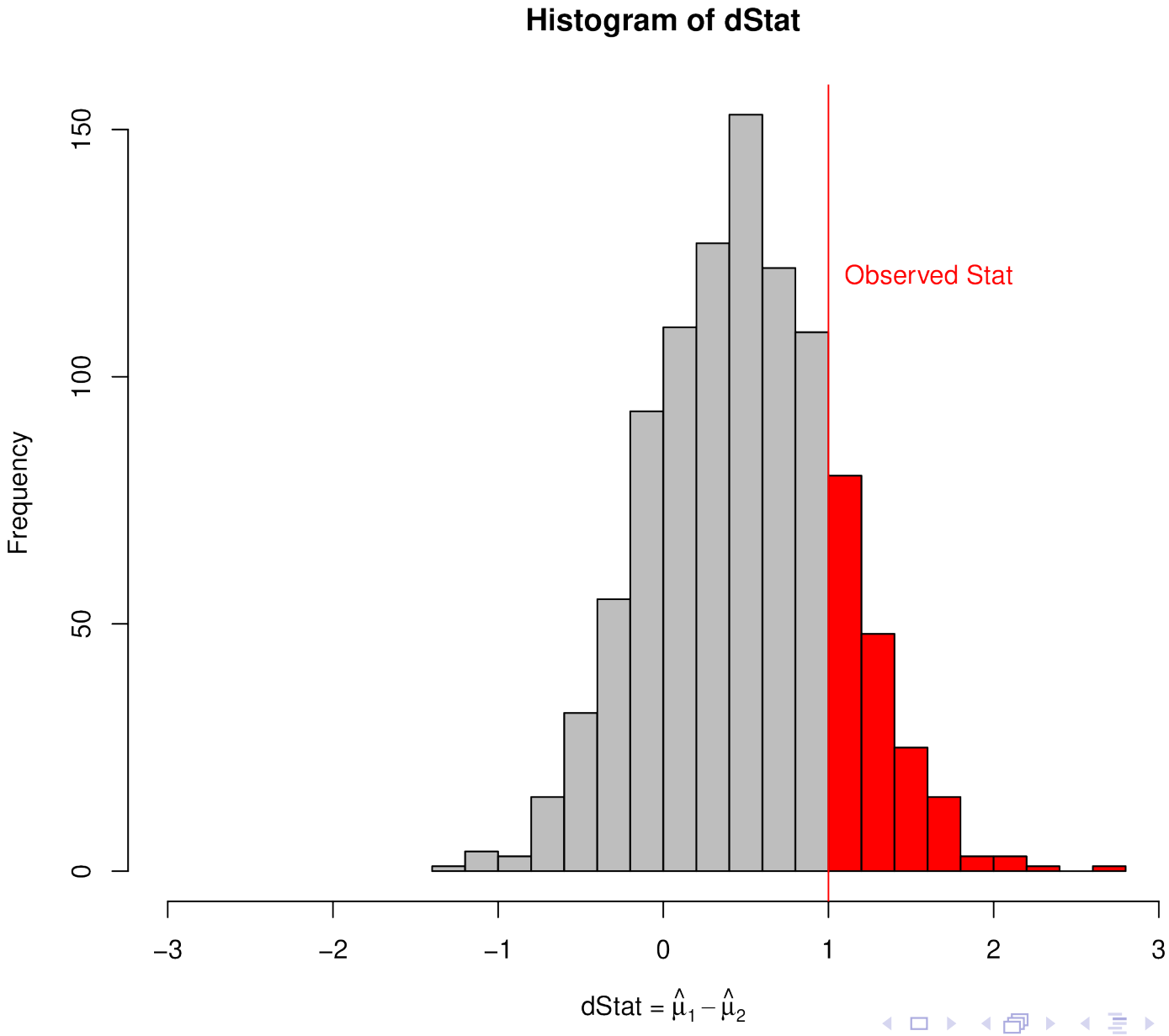
En d'autres termes on veut savoir si la différence qu'il existe entre $\hat{\mu}_1$ et $\hat{\mu}_2$ est suffisemment grande pour pouvoir conclure que ces deux échantillons sont issus de populations qui auraient la même moyenne.

Pour savoir si cette différence est suffisemment grande, si on avait accès aux populations pop_1 et pop_2, on pourrait tirer de manière aléatoire des échantillons de même taille que s_1 et s_2 et calculer la distribution de ces différences.

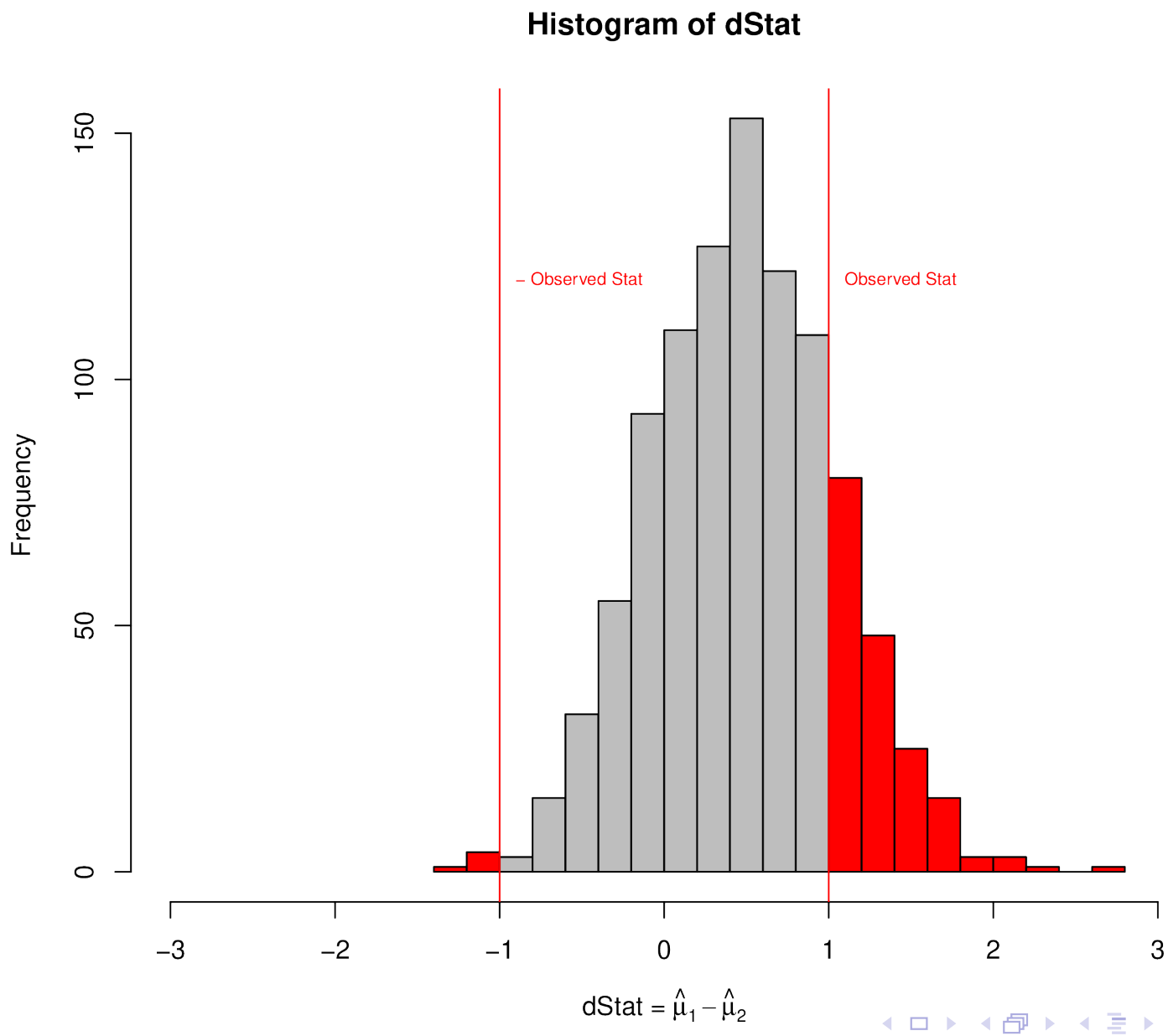
Principe d'un test en statistique: un exemple



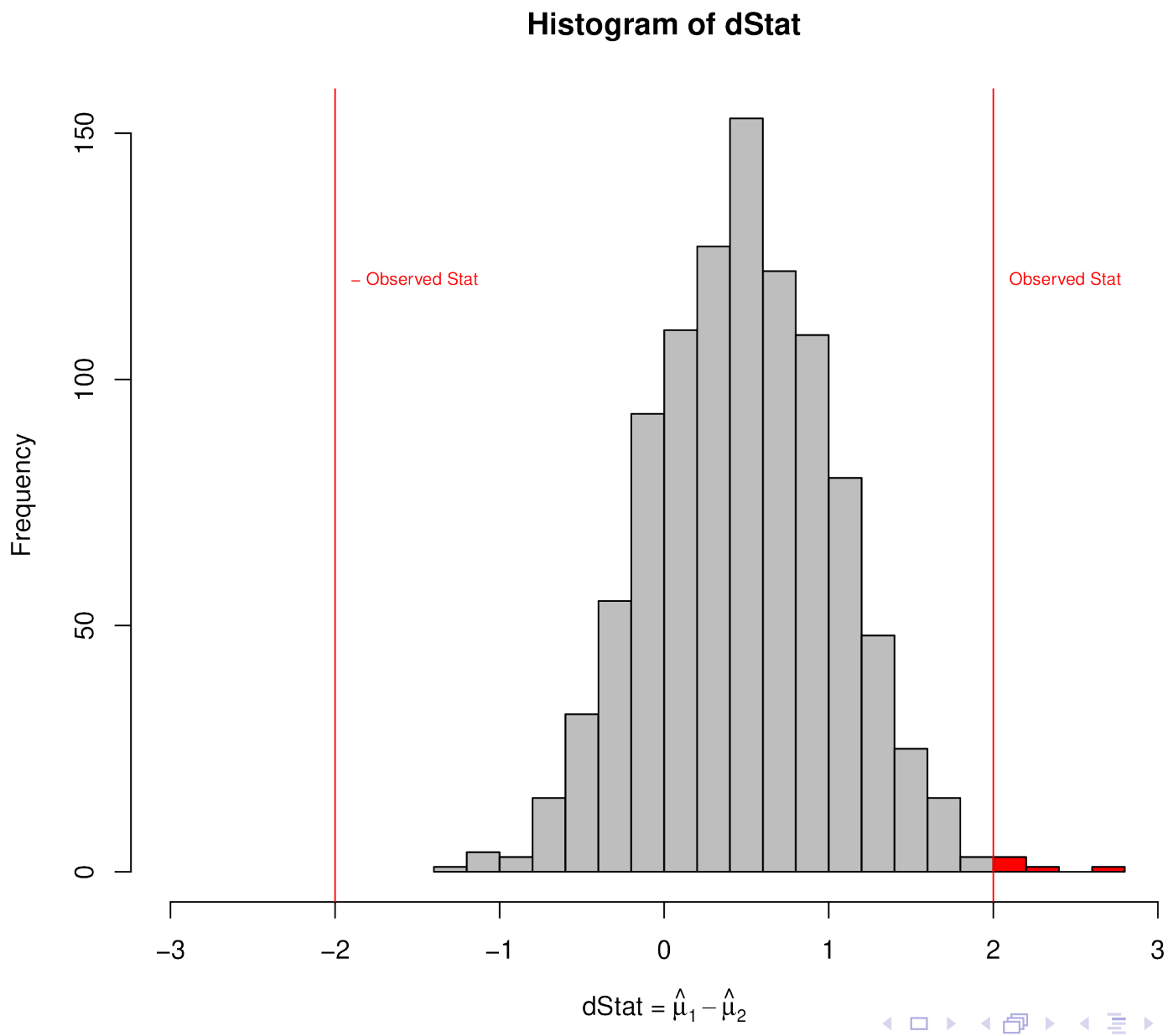
Principe d'un test en statistique: un exemple



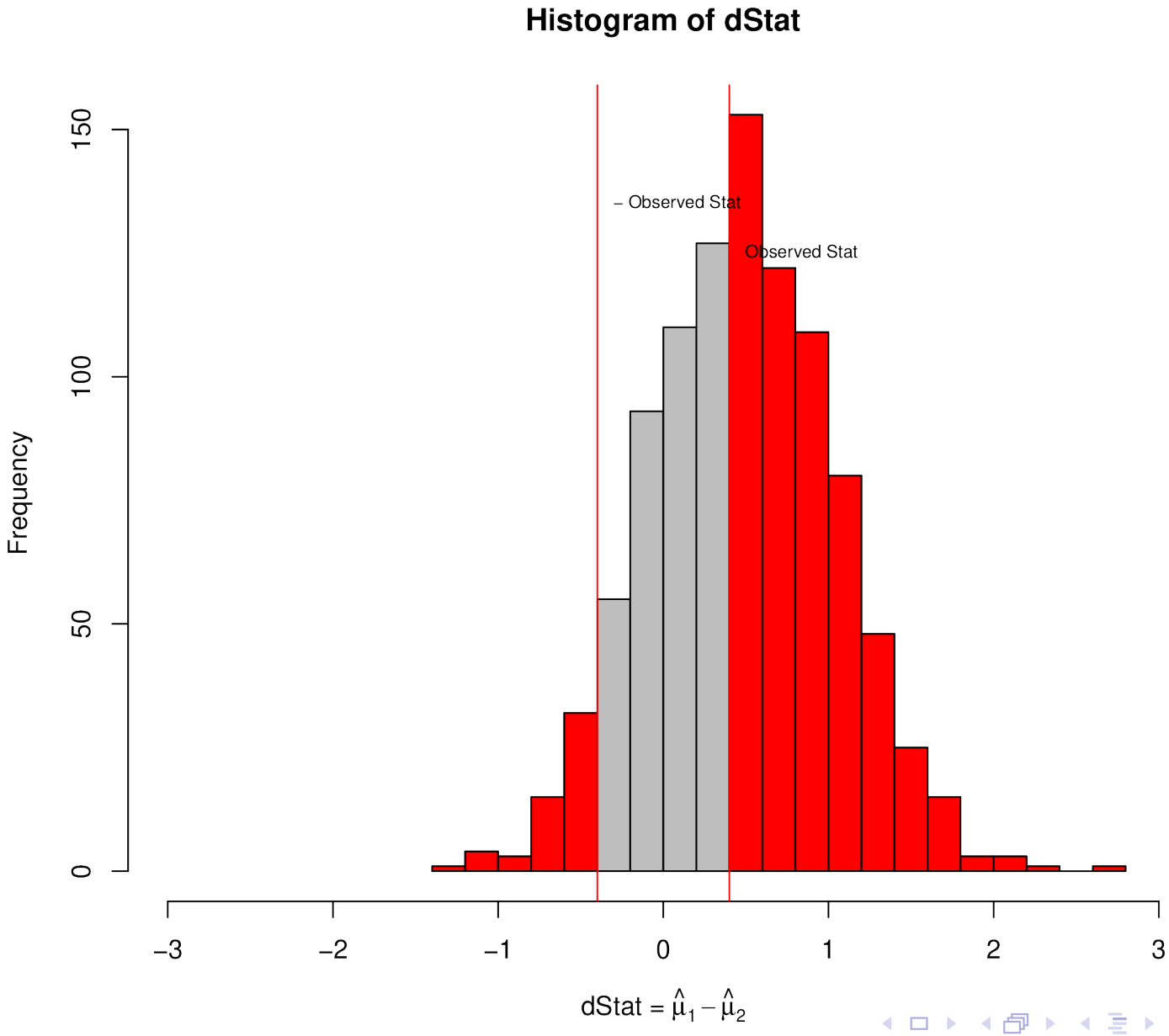
Principe d'un test en statistique: un exemple



Principe d'un test en statistique: un exemple



Principe d'un test en statistique: un exemple



Principe d'un test en statistique: un exemple

Test de comparaison de moyennes

Le problème est que la plupart du temps, nous n'avons pas accès aux populations originales pour faire ces tirages aléatoires

Deux approches sont néanmoins possibles.

approche 1

On peut tout d'abord modéliser le comportement de notre statistique en faisant le pari que les distribution des valeurs dans pop_1 et pop_2 sont conformes à des distribution connues, et alors, soit tirer aléatoirement dans ces distributions connues, soit établir de manière scientifique quel sera le comportement de la statistique.

Principe d'un test en statistique: un exemple

Test de comparaison de moyennes
approche 2

On peut utiliser les valeurs de nos deux échantillons et les regrouper. On peut ensuite tirer ensuite n_1 valeurs de manière aléatoire les assigner à un échantillon, et réserver les valeurs restantes (n_2) à un deuxième échantillon et calculer la différences entre ces deux échantillons aléatoires. En répétant la procédure plusieurs fois, on obtiendra la distribution de la différence entre deux échantillons aléatoires issus de la distribution de l'ensemble des valeurs de l'échantillon.

Principe d'un test en statistique: un exemple

Hypothèses: Ho et H1

En statistique, en général on cherche à savoir si on conserve l'hypothèse nulle ou on la rejette. Dans notre cas, on veut connaitre le risque de se tromper en déclarant que les deux échantillons ne sont pas égaux en moyenne alors qu'ils pourraient l'être. L'hypothèse d'égalité est l'hypothèse nulle (Ho), c'est sur la base de l'égalité qu'on a calculé la distribution de la statistique (ie que vaudrait la différences entre deux échantillons aléatoire issus de la même distribution). L'hypothèse alternative est tout ce qui est différent de l'hypothèse nulle. Ici, que les moyennes des populations (ou de la la population) dont sont issus nos échantillons sont différentes. Ho et H1 sont toujours exclusives !!

Principe d'un test en statistique: un exemple

Hypothèses: Ho et H1

Via nos approches, nous pouvons obtenir la probabilité d'observer une valeur aussi extrême que la différence observée entre nos deux échantillons que la différence entre les moyennes de deux échantillons de taille identiques tirés d'une population uniqueé. Cette probabilité sera d'autant plus faible que la différence entre les deux échantillons sera grande et que la variabilité dans les échantillons sera petite. Cette probabilité est la $p - valeur$ en statistique. On peut l'écrire $P(|Stat_{obs}| > |dStat||Ho)$
Plus généralement, la $p - valeur$ est défnie comme étant la probabilité de se tromper en déclarant que Ho est fausse, alors qu'elle est vraie.

Principe d'un test en statistique: un exemple

Test de comparaison de moyennes: Approche 1 en principe

Dans le cas de l'approche 1, le plus simple est de considérer la variation de notre statistique comme si les deux échantillons était issus de la même population dont la loi de distribution est conforme au paramètres de nos populations. On sait que si on a des échantillons tirés dans une même loi normale que $\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{(n)}}$ suivra une loi de Student à $n-1$ ddl.

Si j'ai deux moyennes de deux échantillons alors: $\frac{\hat{\mu}_1}{\hat{\sigma}_1/\sqrt{(n_1)}}$ et $\frac{\hat{\mu}_2}{\hat{\sigma}_2/\sqrt{(n_2)}}$ suivent l'une et l'autre des lois noramles.

Si on considère que $\hat{\sigma}_1$ et $\hat{\sigma}_2$ représente l'écart type $\hat{\sigma}_p$ de la même population sont équivalent, alors la différence suivra une loi de student.

Principe d'un test en statistique: un exemple

Test de comparaison de moyennes: Approche 1 en principe

$\hat{\sigma}_p$ serait alors estimer par l'écart type moyen de nos deux échantillons, et serait estimé par

$\hat{\sigma}_p = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}}$

La valeur $t = \frac{\hat{\mu}_1-\hat{\mu}_2}{\hat{\sigma}_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ suit une loi de student à $n-2$ degrés de liberté.

Plus cette valeur sera extrême, et plus on pourra penser que la valeur de t sera extrême par rapport à l'attendue et donc que nos moyennes diffèrent.