L'estimation statistique

Estimation des paramètres statistiques décrivant une population à partir d'un échantillon de cette population

M. L. Delignette-Muller
VetAgro Sup

12 novembre 2018

VetAgro Sup

UNIVERSITÉ DE LYON

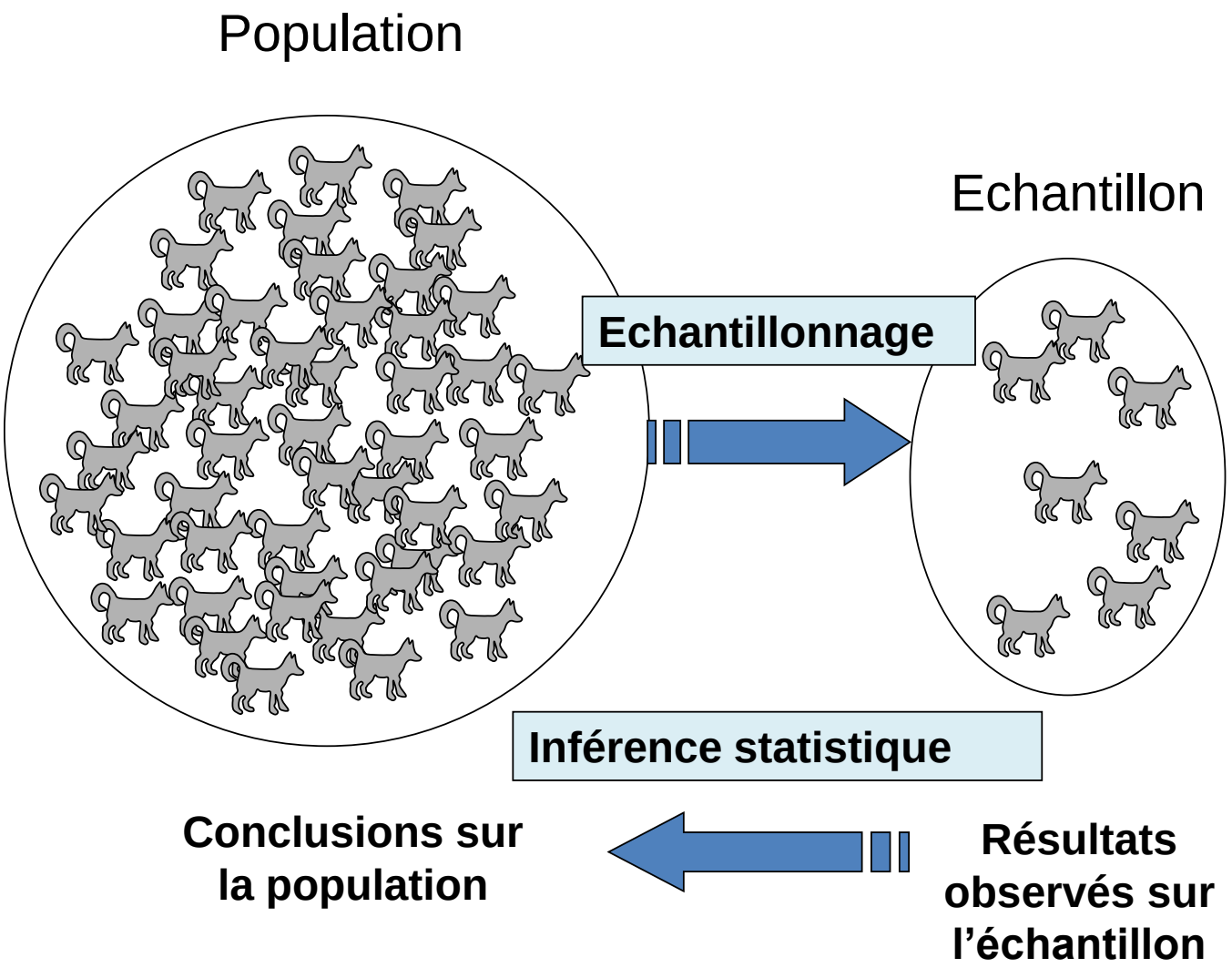
Objectifs pédagogiques

- Savoir définir les notions suivantes : inférence statistique, échantillonnage aléatoire simple, distribution d'échantillonnage, estimation et estimation sans biais.
- Savoir ce que représentent SD et SE (ou SEM).
- Avoir bien compris le théorème de l'approximation normale.
- Savoir juger de l'applicabilité de ce théorème et vérifier les conditions d'utilisation des divers intervalles de confiance.
- Savoir ce que représente un intervalle de confiance et ce qui le différencie d'un intervalle de fluctuation.
- Savoir calculer à la main (avec une calculatrice) un intervalle de confiance sur une moyenne et sur une fréquence.*

* *savoir faire évalué uniquement en S5*

Plan

Echantillonnage et inférence statistique



Principe de l'échantillonnage

- Peut-on caractériser une population en ne disposant que d'un échantillon de celle-ci ?
- Comment obtenir un échantillon représentatif de la population étudiée ?
- Comment éviter les biais d'échantillonnage ?

Un exemple historique : premiers sondages électoraux aux Etats-Unis en 1936

Revue Literacy Digest
sondage sur
2 300 000 personnes
Alfred Landon : 55%
Franklin Roosevelt : 41%

Maison Gallup
sondage sur
6 500 personnes
Alfred Landon : 35%
Franklin Roosevelt : 64%

Un exemple historique : premiers sondages électoraux aux Etats-Unis en 1936

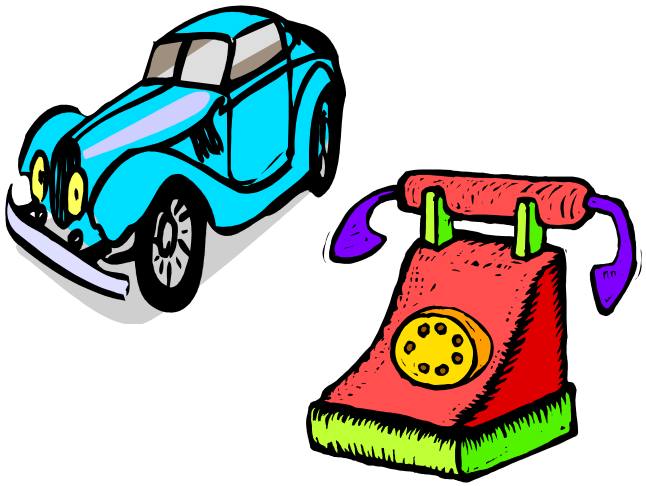
Revue Literacy Digest
sondage sur
2 300 000 personnes
Alfred Landon : 55%
Franklin Roosevelt : 41%

Maison Gallup
sondage sur
6 500 personnes
Alfred Landon : 35%
Franklin Roosevelt : 64%

Résultat des élections : réélection de Roosevelt avec 61% des voix.
Disparition de la revue Literacy Digest suite à sa terrible erreur.
Pourquoi une telle erreur ?

Bel exemple de biais d'échantillonnage

Revue Literacy Digest
sondage à partir des immatriculations et des listes des annuaires téléphoniques



Maison Gallup
sondage aléatoire

Quelques autres exemples de biais d'échantillonnage ?

- **Un biais facile à éviter dans un cadre expérimental** :
  tirages d'animaux dans une cage en prenant le premier qui vient (ce nest pas un choix au hasard et il convient d'identifier les animaux pour faire un tirage aléatoire).
- **Un biais plus difficile à éviter dans un cadre observationnel** :
  estimation de la prévalence d'une maladie dans une population sauvage à partir d'animaux capturés ou d'animaux retrouvés morts (chacune de ses 2 catégories n'est pas représentative de la population des animaux vivants).

Comment éviter les biais d'échantillonnage ?

Une méthode simple et classique :

l'échantillonnage aléatoire simple

- tirages aléatoires et indépendants des individus de l'échantillon (plus souvent sans remise)
- tous les individus ont la même probabilité d'être tiré

TABLE DE « NOMBRES AU HASARD »

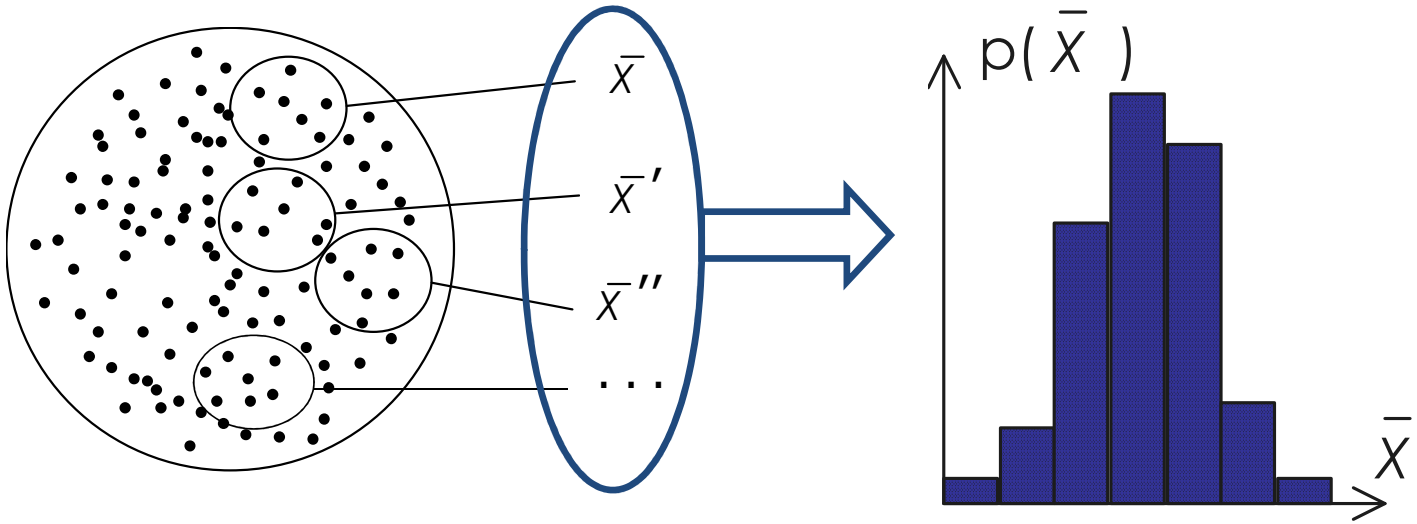
Extrait de la table de Kendall et Babington Smith

02	22	85	19	48	74	55	24	89	69	15	53	00	20	88	48	95	08	00	47
85	76	34	51	40	44	62	93	65	99	72	64	09	34	01	13	09	74	90	65
00	88	96	79	38	24	77	00	70	91	47	43	43	82	71	67	49	90	37	09
64	29	81	85	50	47	36	50	91	19	09	15	98	75	60	58	33	15	51	44
94	03	80	04	21	49	54	91	77	85	00	45	68	23	12	94	23	44	36	88

Notion de distribution d'échantillonnage

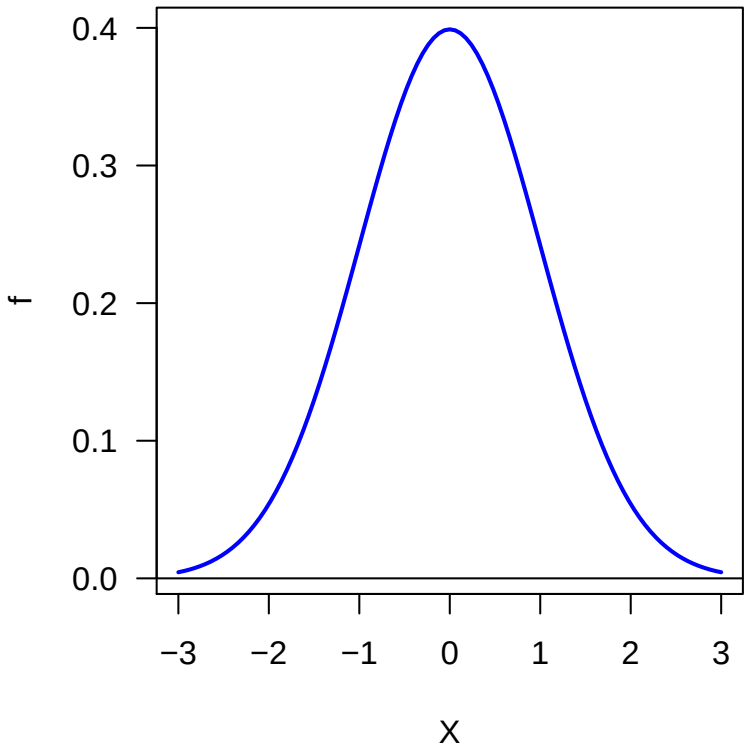
Exemple de l'étude d'une variable quantitative X.
La **densité de probabilité de** $\overline{X}$
(obtenue en supposant une répétition de l'échantillonnage)
est appelée la **distribution d'échantillonnage de** $\overline{X}$.

Plusieurs échantillons ⇨ plusieurs moyennes



Distribution d'échantillonnage pour des tirages dans une loi normale

Distribution de X dans la population : loi normale $N(0, 1)$

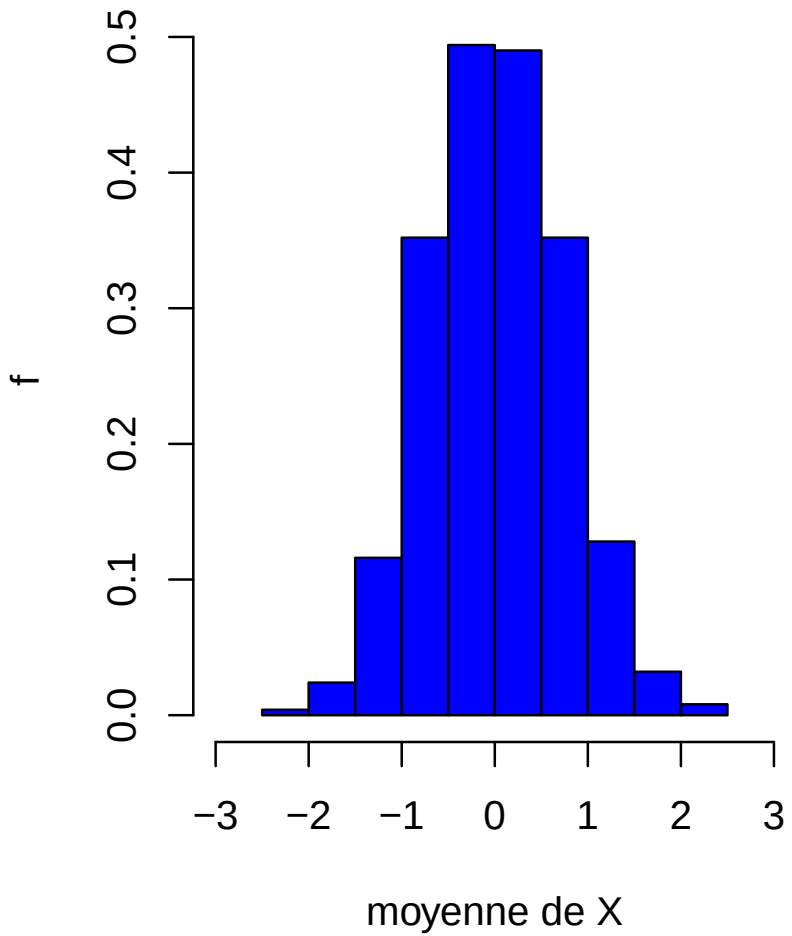


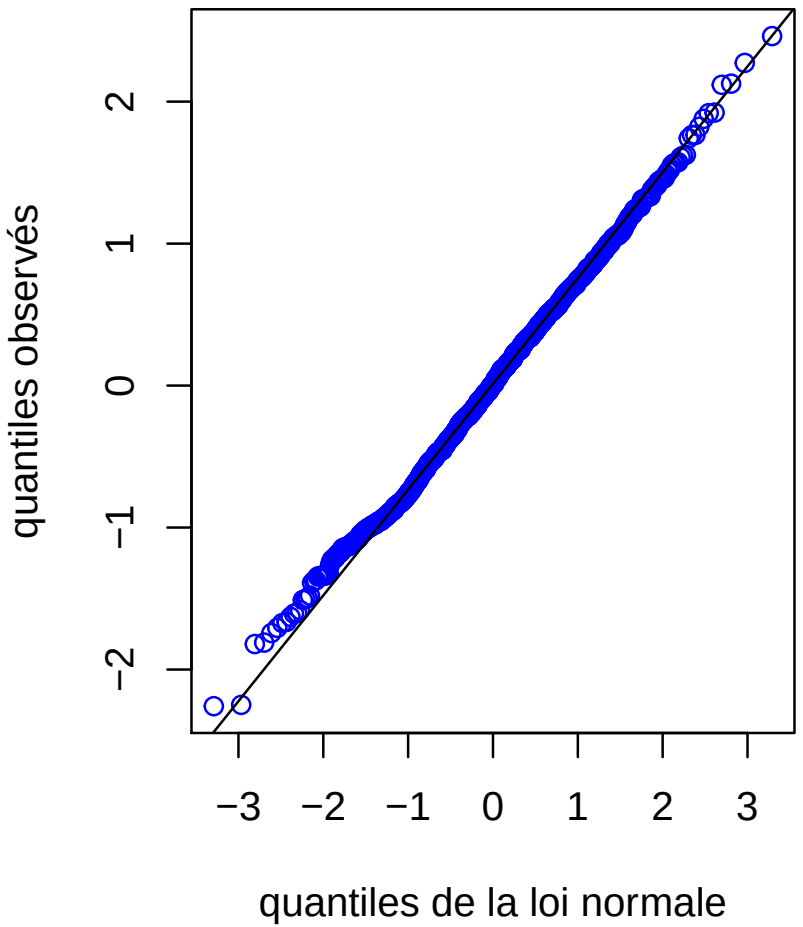
Pour i allant de 1 à 1000

- On tire N observations dans la loi
- On calcule la moyenne observée des N observations : m_i

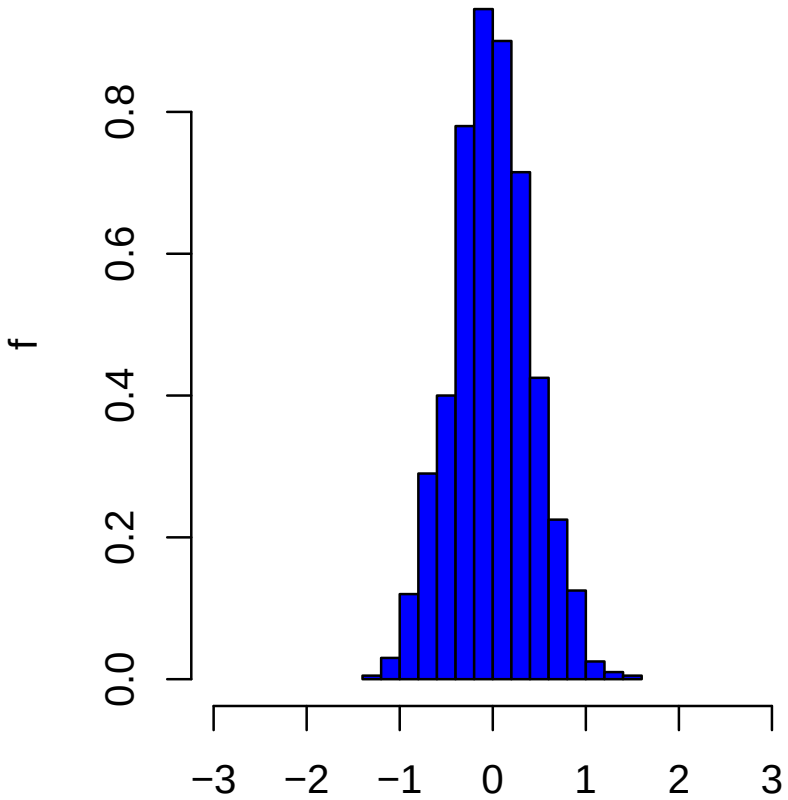
On visualise ensuite la distribution d'échantillonnage de la moyenne (distribution des m_i)

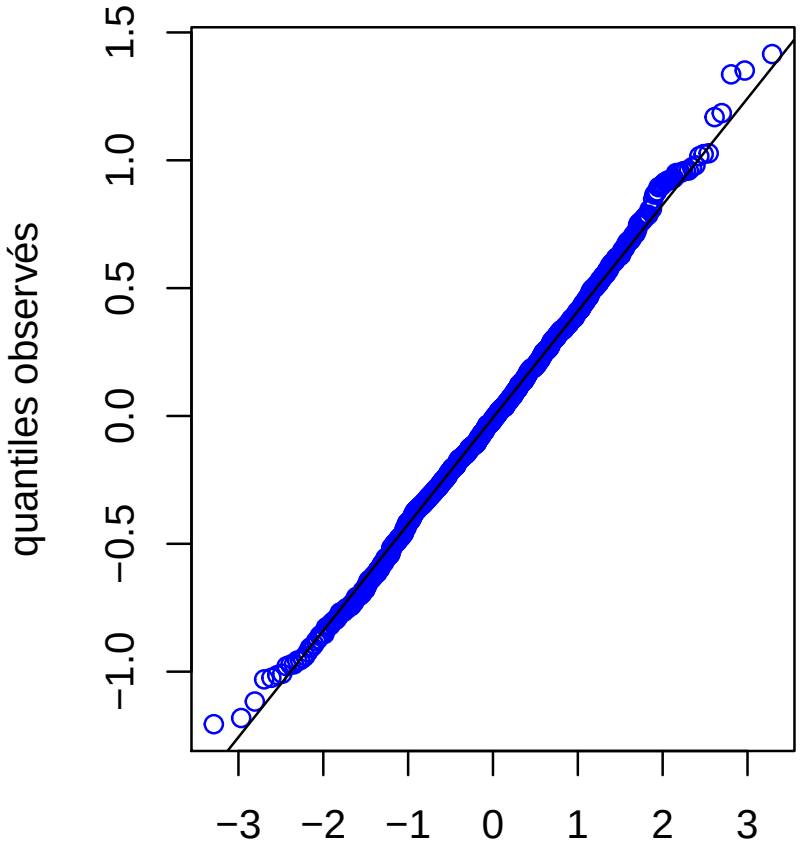
Distribution d'échantillonnage pour $N = 2$



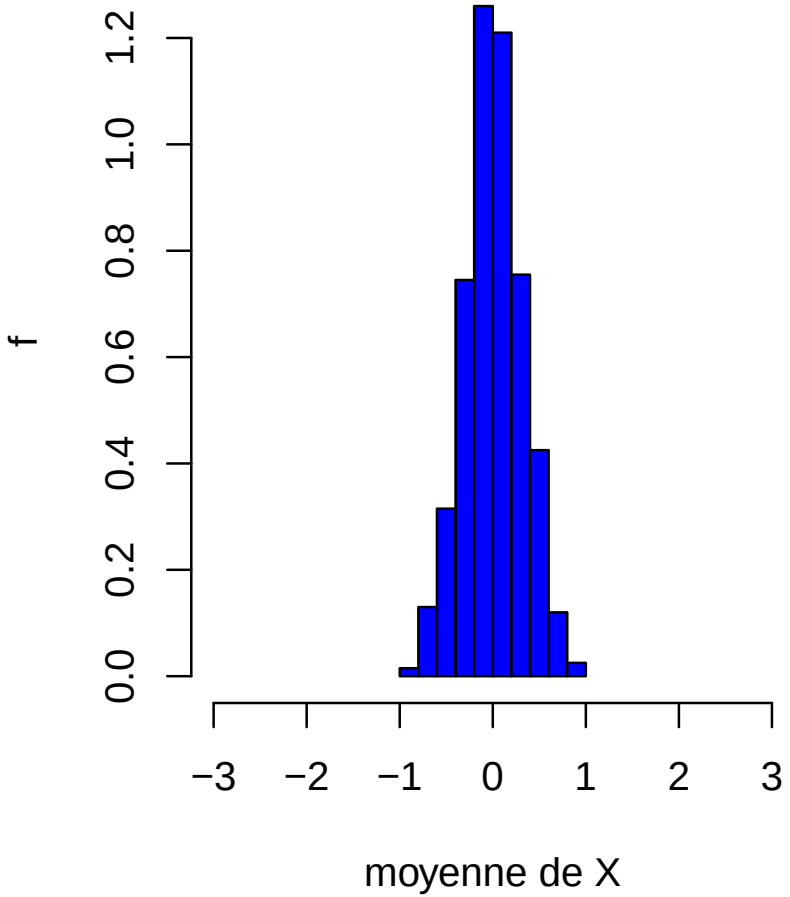


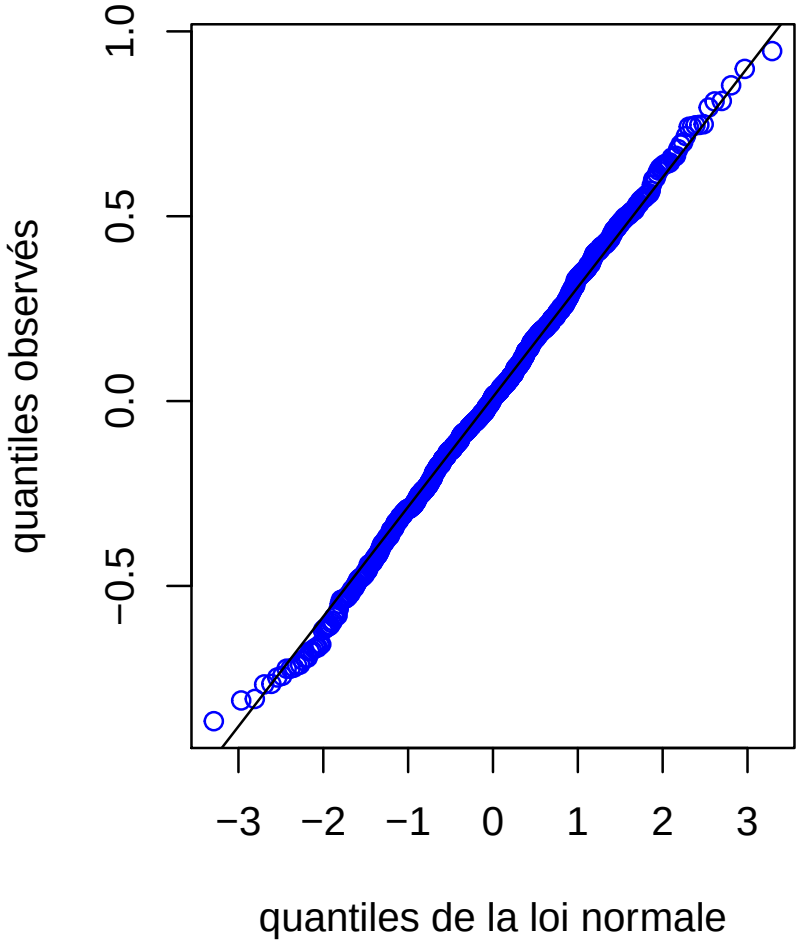
Distribution d'échantillonnage pour $N = 5$



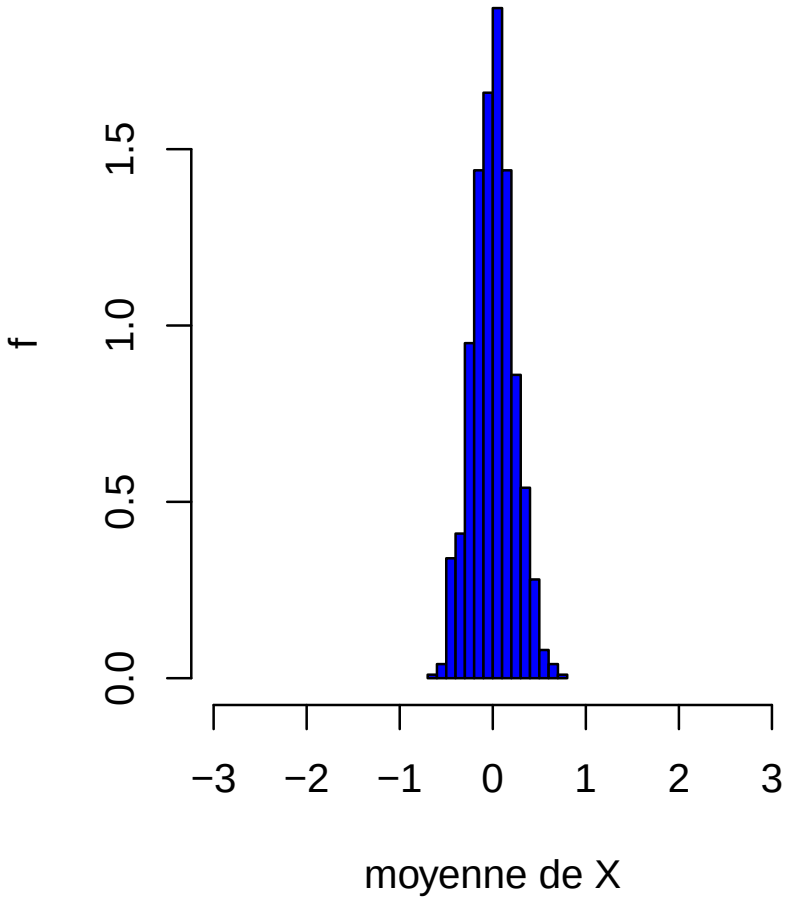


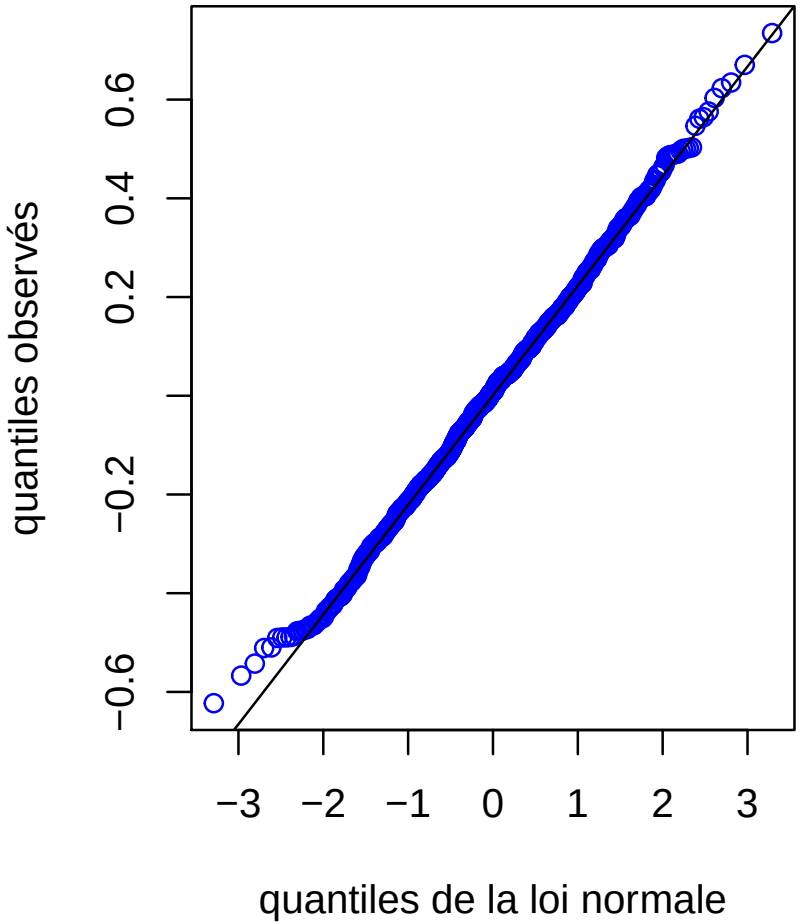
Distribution d'échantillonnage pour $N = 10$





Distribution d'échantillonnage pour $N = 20$





Début du théorème de l'approximation normale pour une variable quantitative X (à compléter)

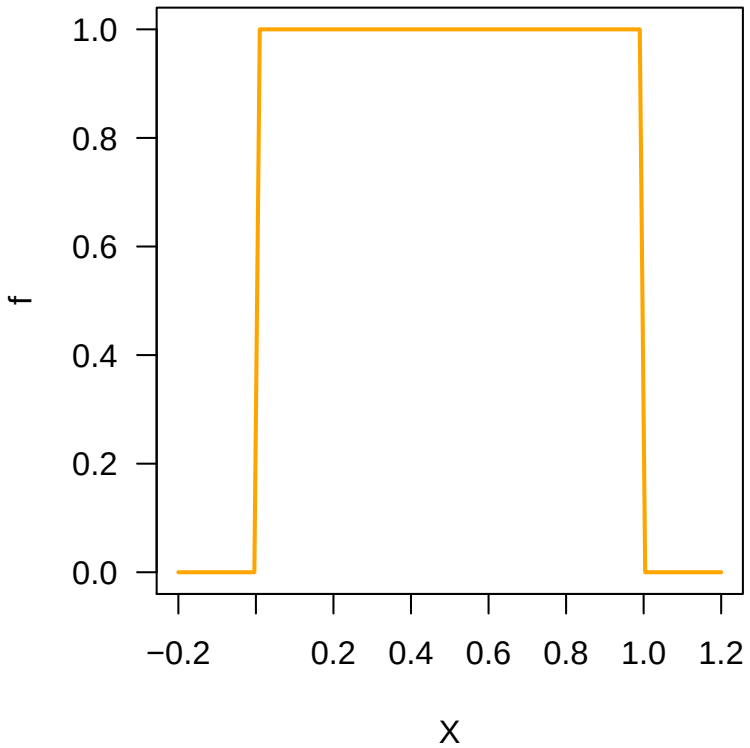
Théorème de l'approximation normale
Cas d'une variable quantitative suivant une loi normale

Pour des échantillons aléatoires simples de taille N, la moyenne $\overline{X}$ de l'échantillon varie autour de la moyenne μ de la population avec une erreur standard $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$ noté SE ou SEM ("Standard Error of the Mean"), σ étant l'écart type de la population noté SD ("Standard Deviation").

- Lorsque la distribution de X dans la population est normale, $\overline{X}$ suit la loi $N(\mu, \frac{\sigma}{\sqrt{N}})$.

Distribution d'échantillonnage pour des tirages dans une loi uniforme

Distribution de X dans la population : loi normale $U(0,1)$

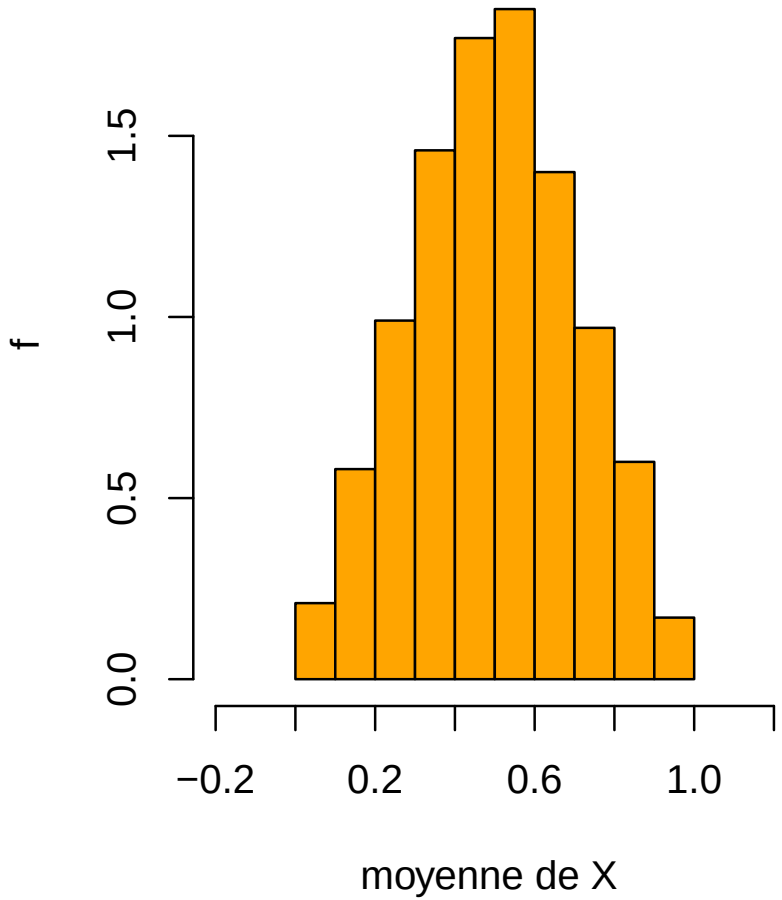


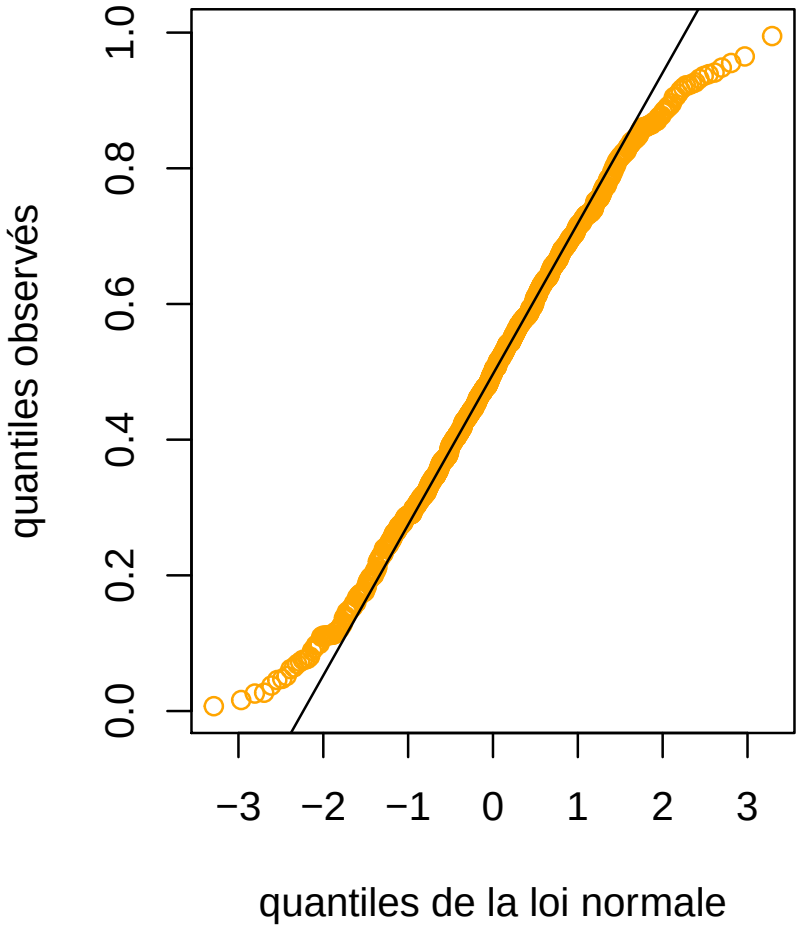
Pour i allant de 1 à 1000

- On tire N observations dans la loi $U(0,1)$
- On calcule la moyenne observée des N observations : m_i

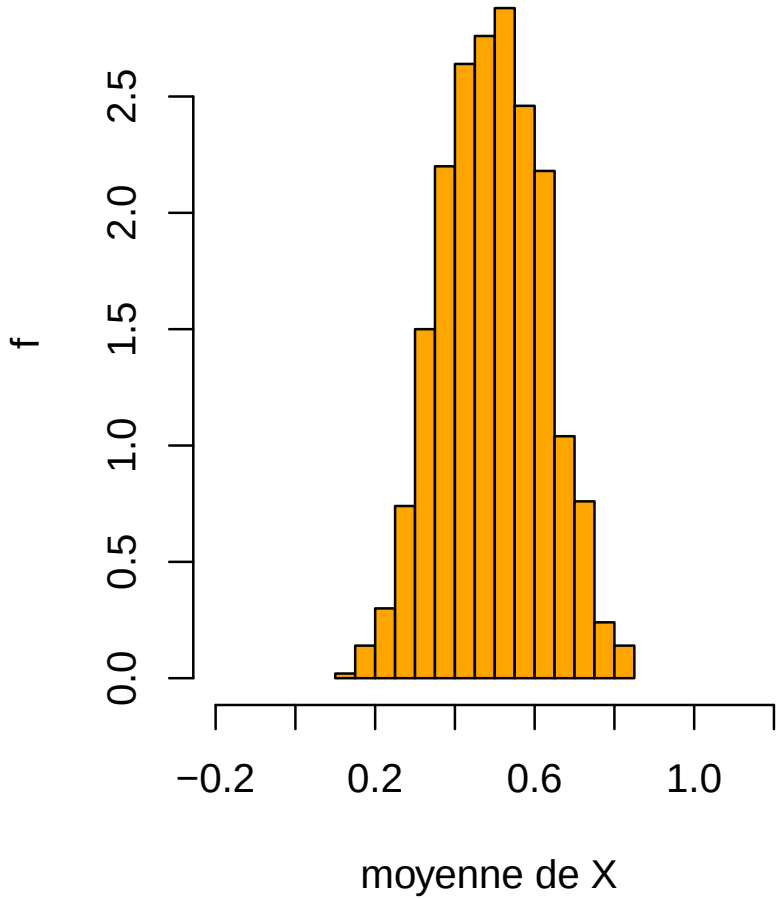
On visualise ensuite la distribution d'échantillonnage de la moyenne (distribution des m_i)

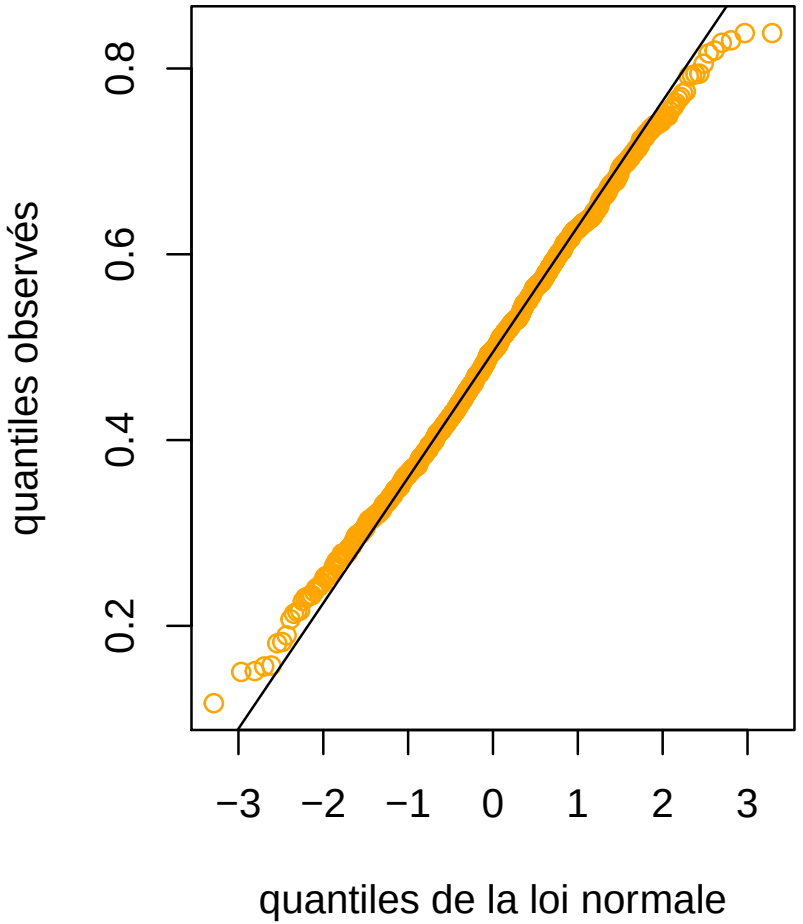
Distribution d'échantillonnage pour $N = 2$



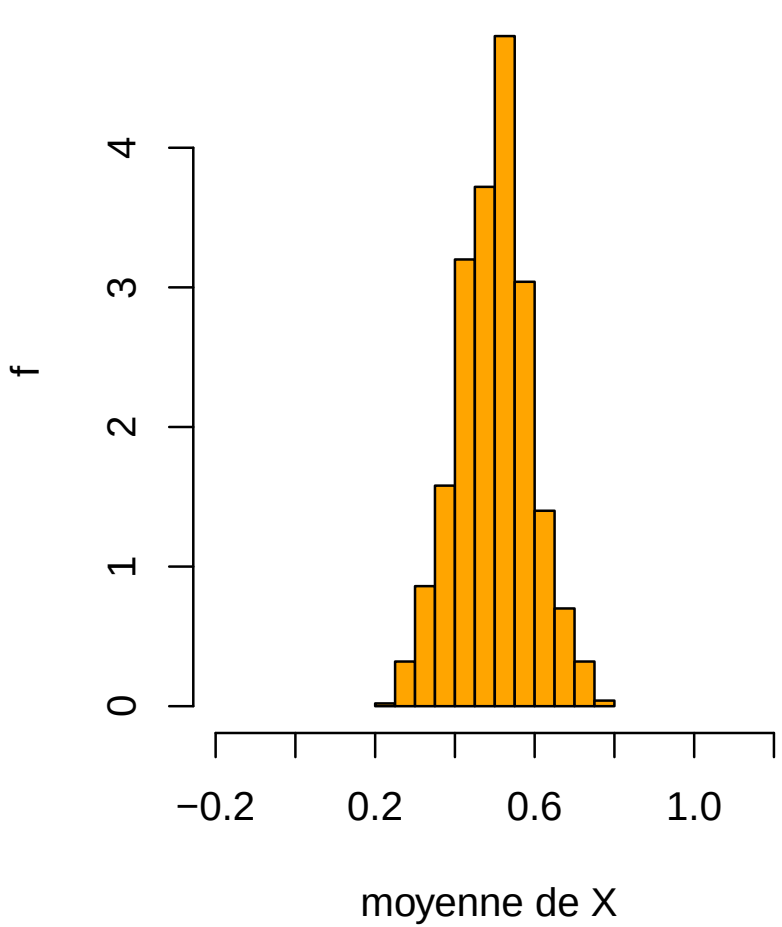


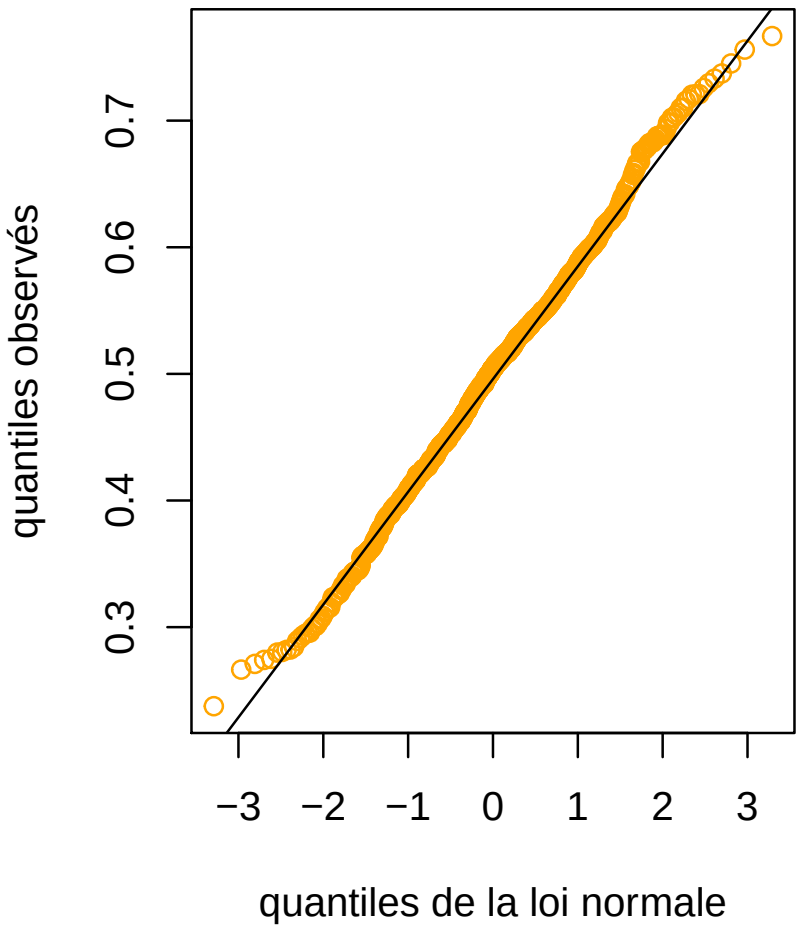
Distribution d'échantillonnage pour $N = 5$



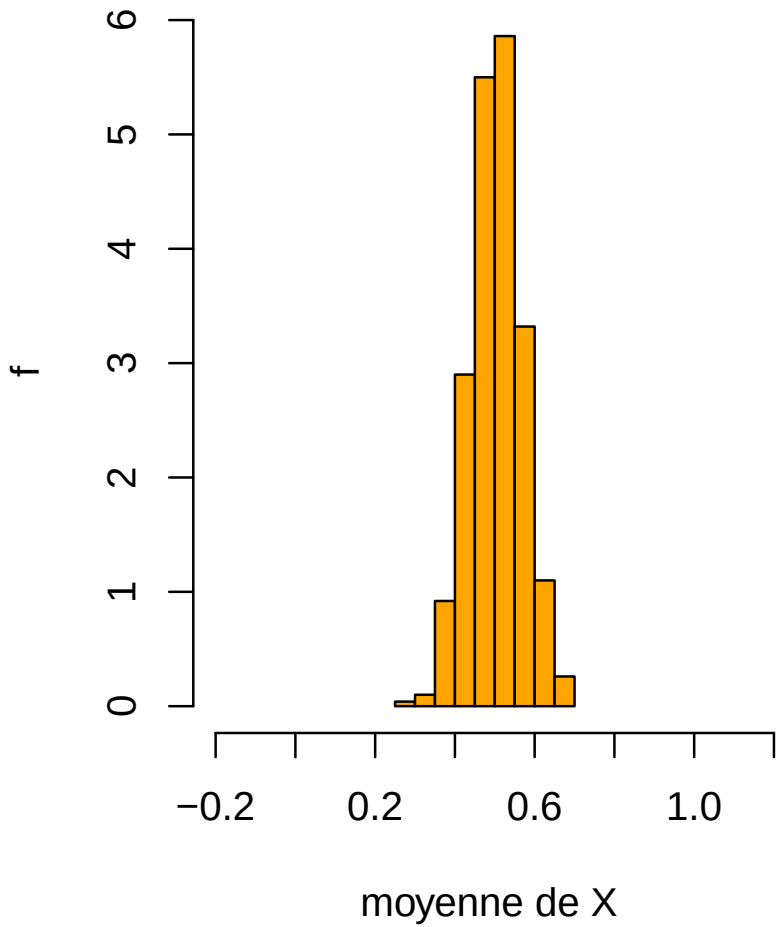


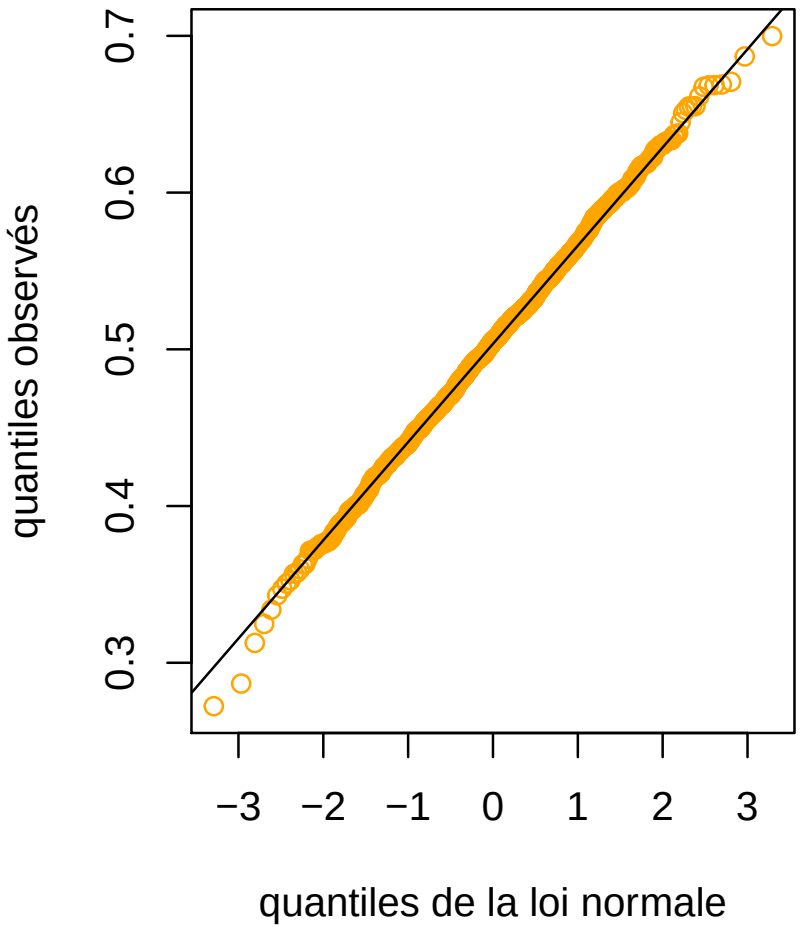
Distribution d'échantillonnage pour $N = 10$





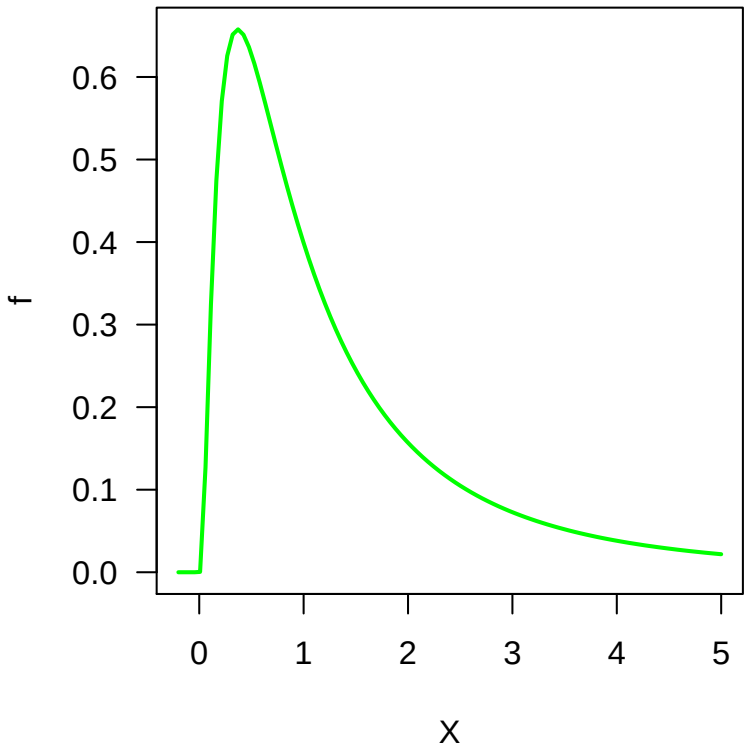
Distribution d'échantillonnage pour $N = 20$





Distribution d'échantillonnage pour des tirages dans une loi lognormale

Distribution de X dans la population : loi lognormale $LN(0, 1)$

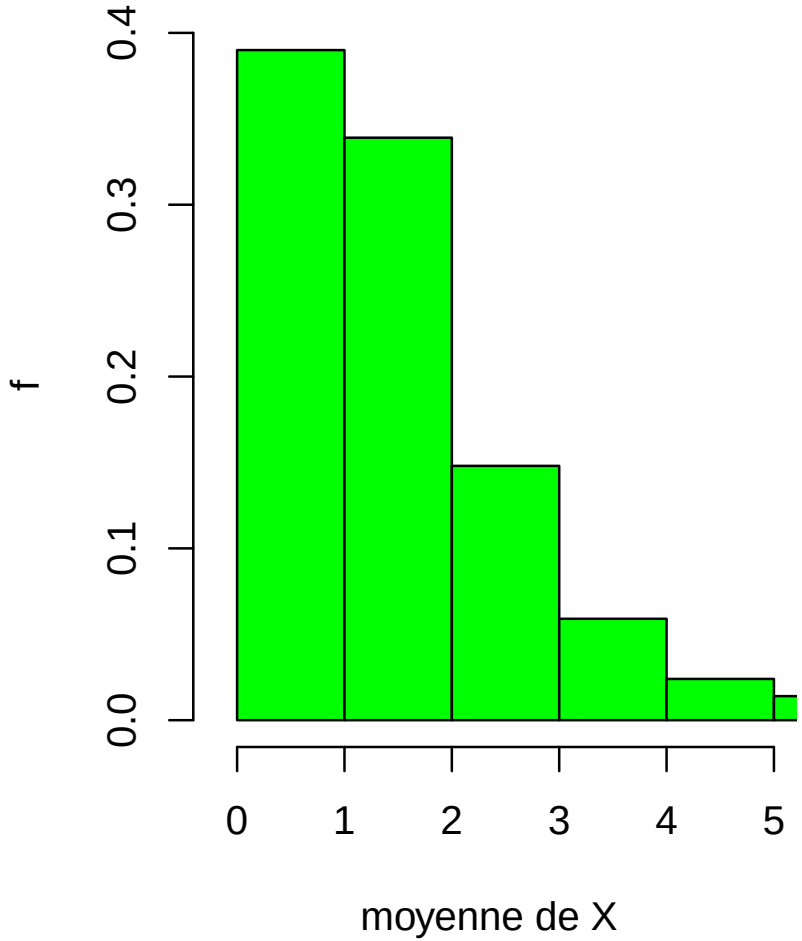


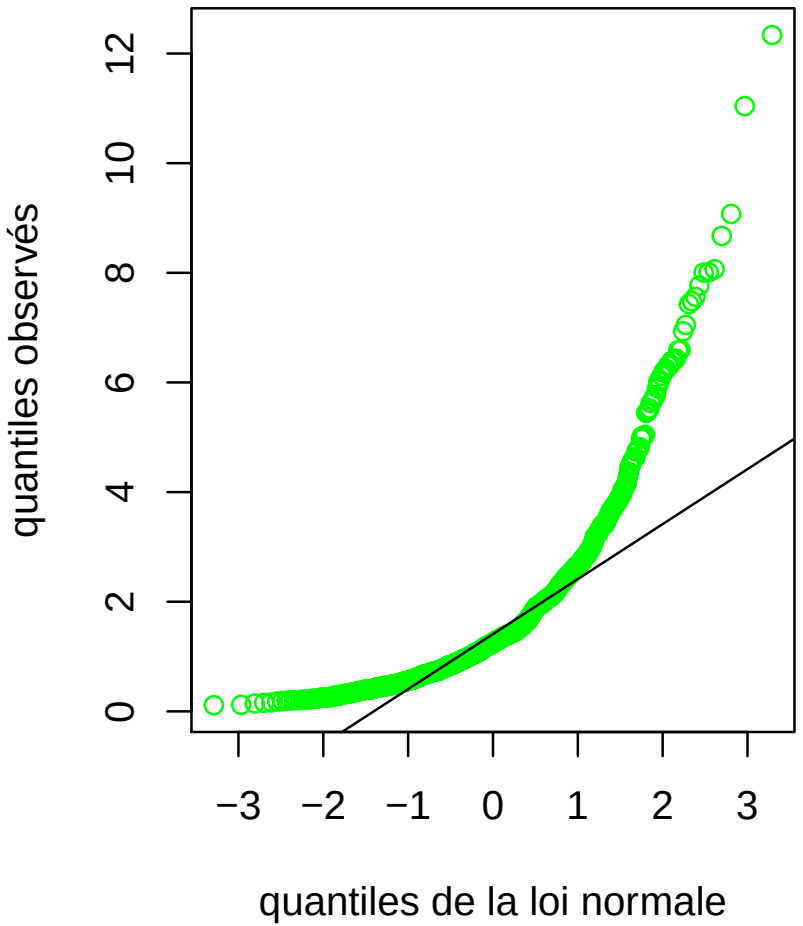
Pour i allant de 1 à 1000

- On tire N observations dans la loi $LN(0, 1)$
- On calcule la moyenne observée des N observations : m_i

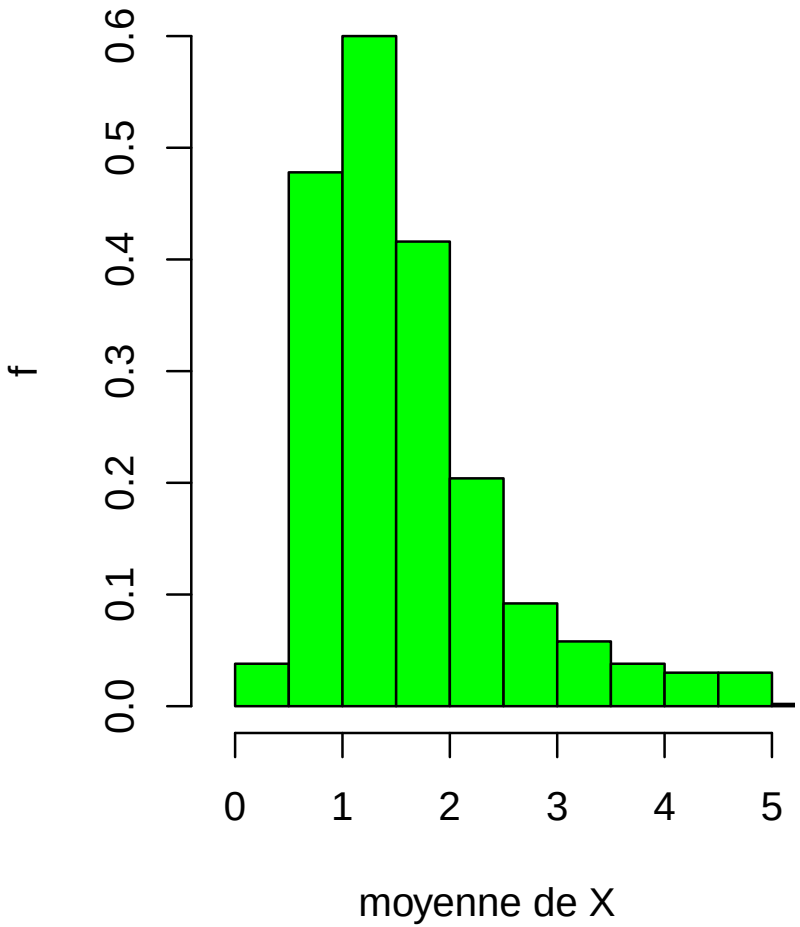
On visualise ensuite la distribution d'échantillonnage de la moyenne (distribution des m_i)

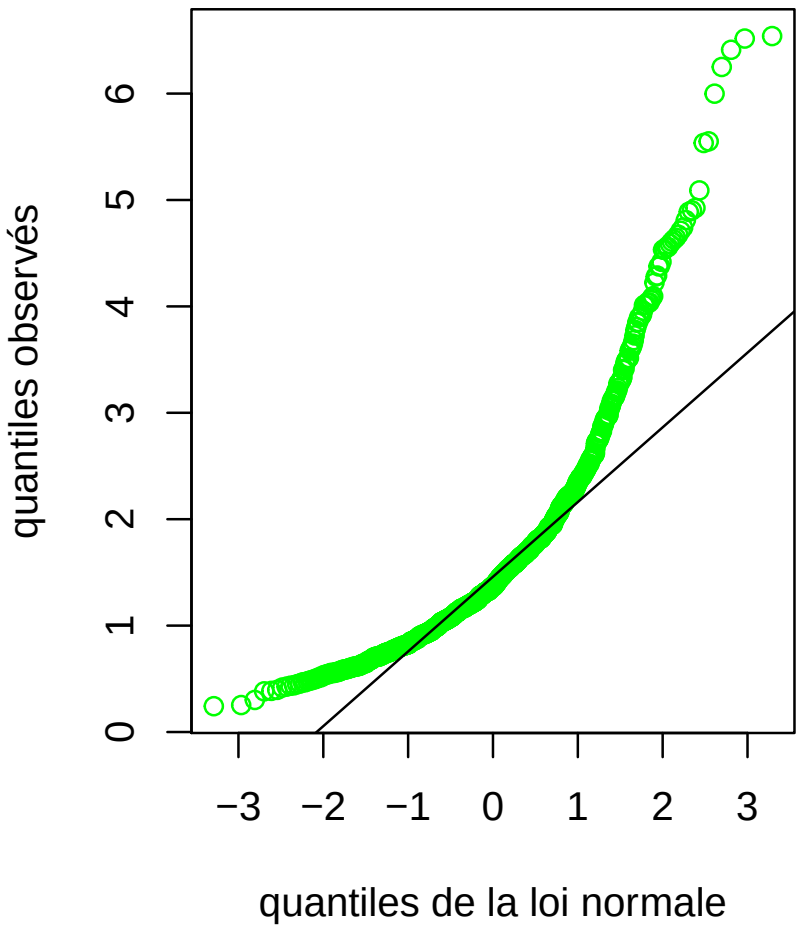
Distribution d'échantillonnage pour $N = 2$



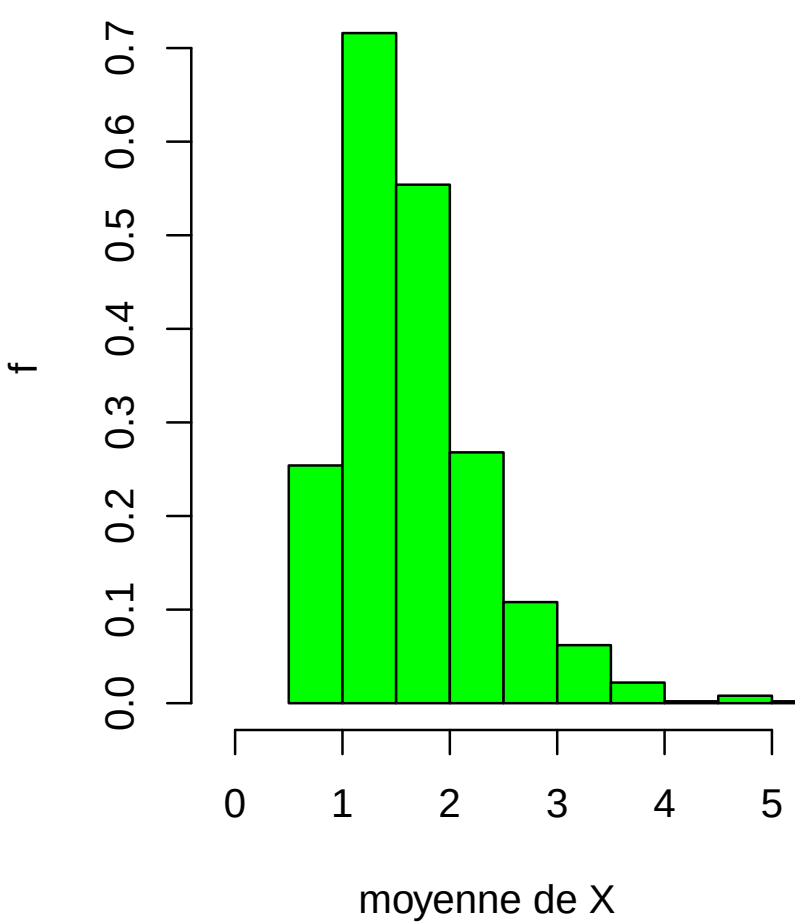


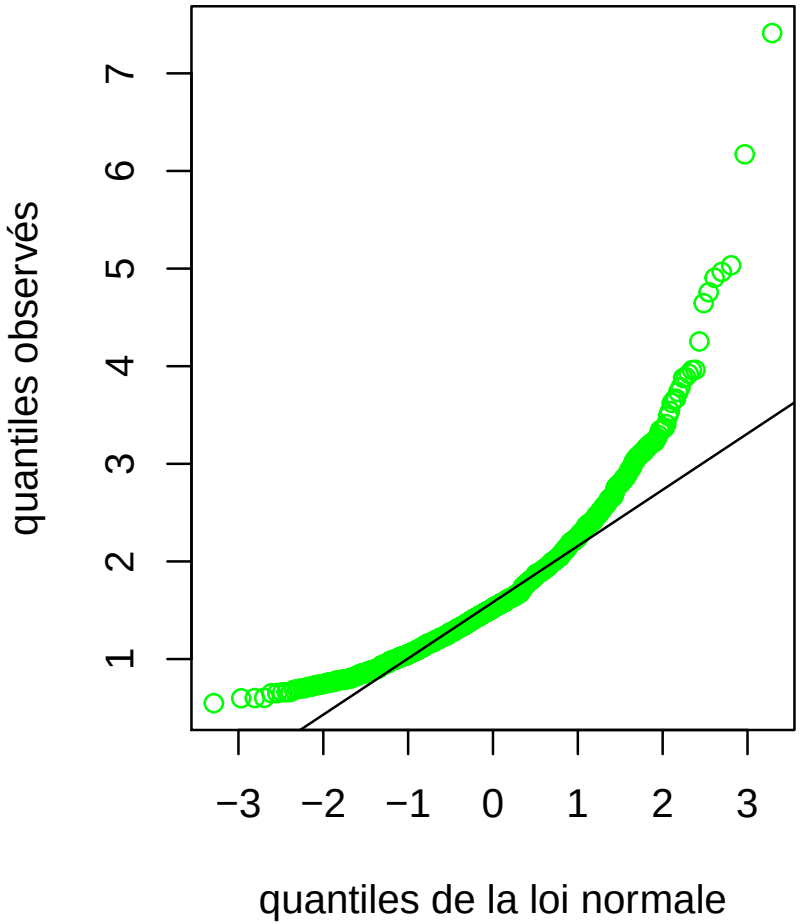
Distribution d'échantillonnage pour $N = 5$



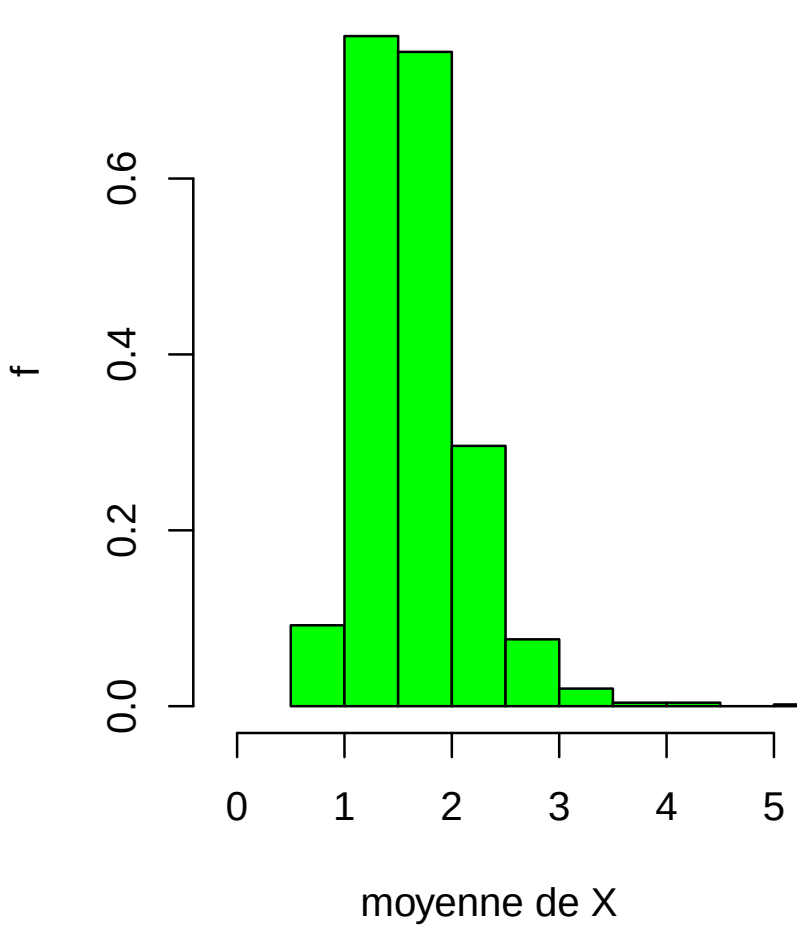


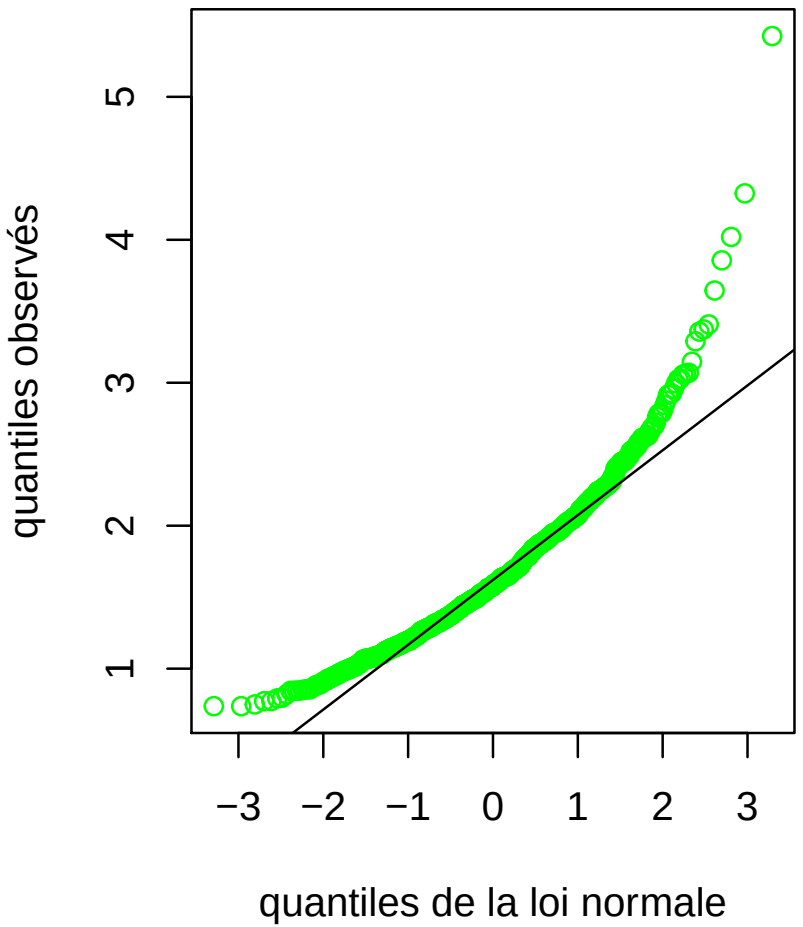
Distribution d'échantillonnage pour $N = 10$



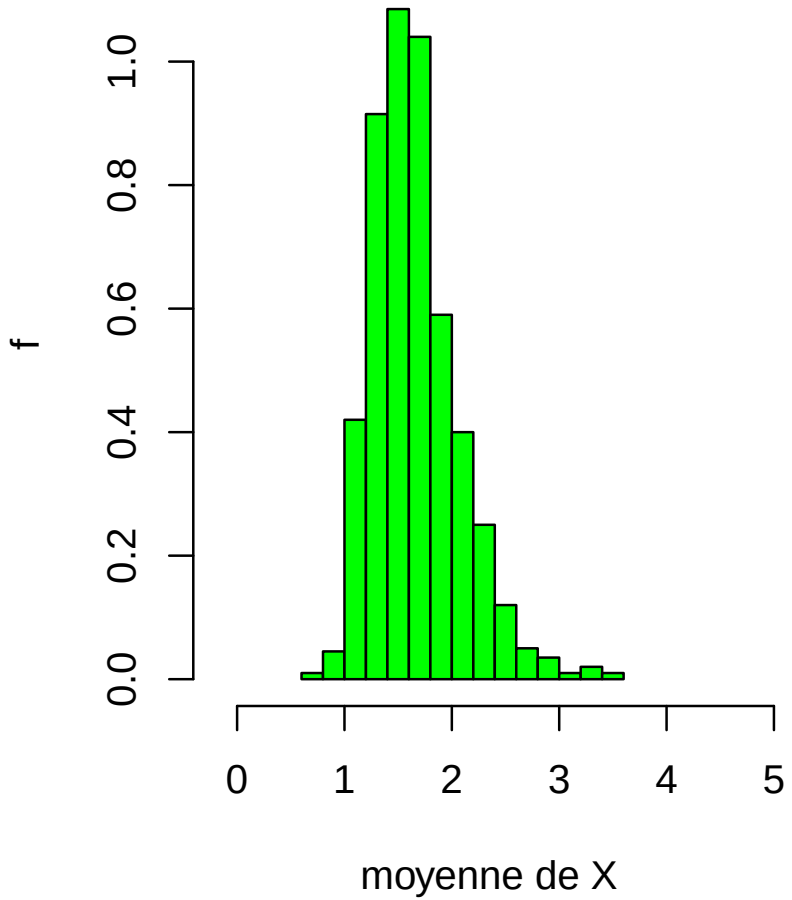


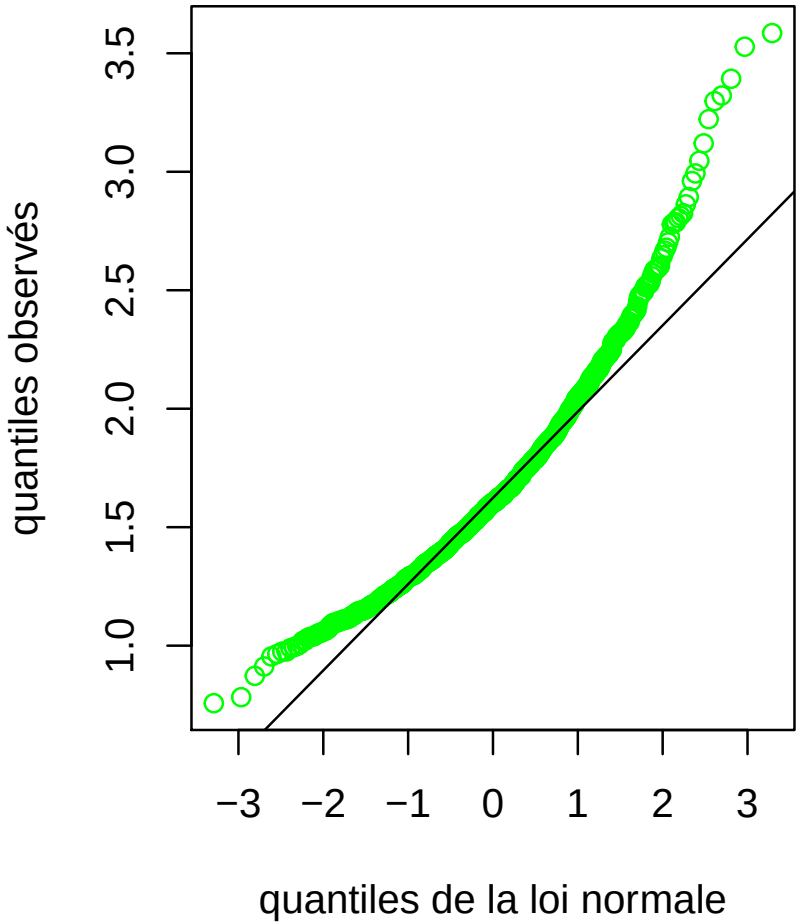
Distribution d'échantillonnage pour $N = 20$



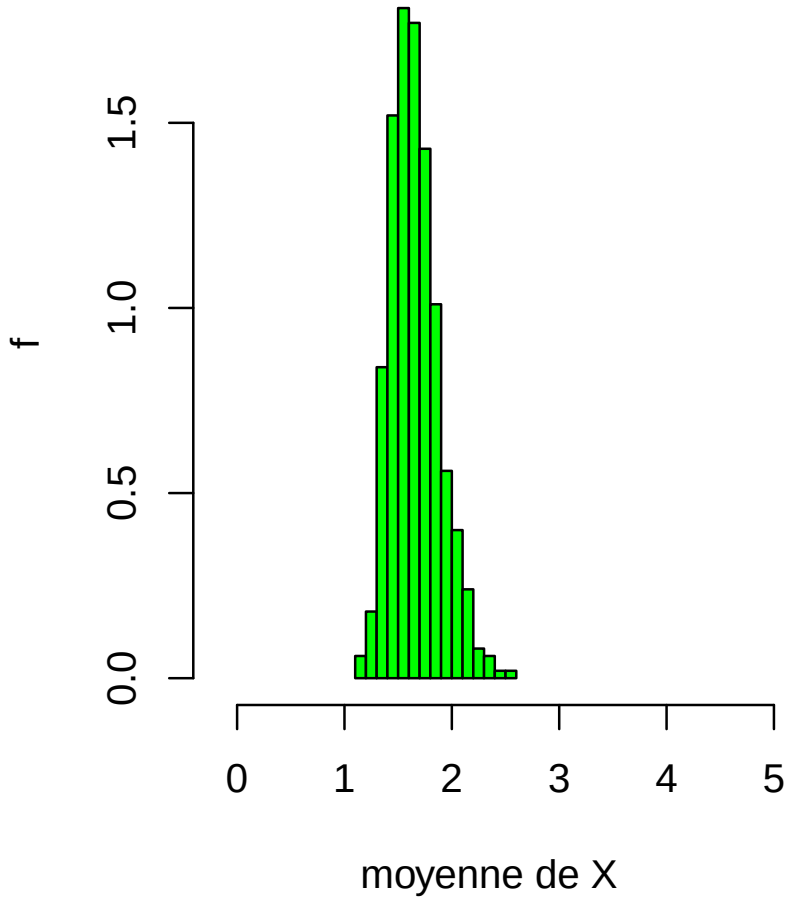


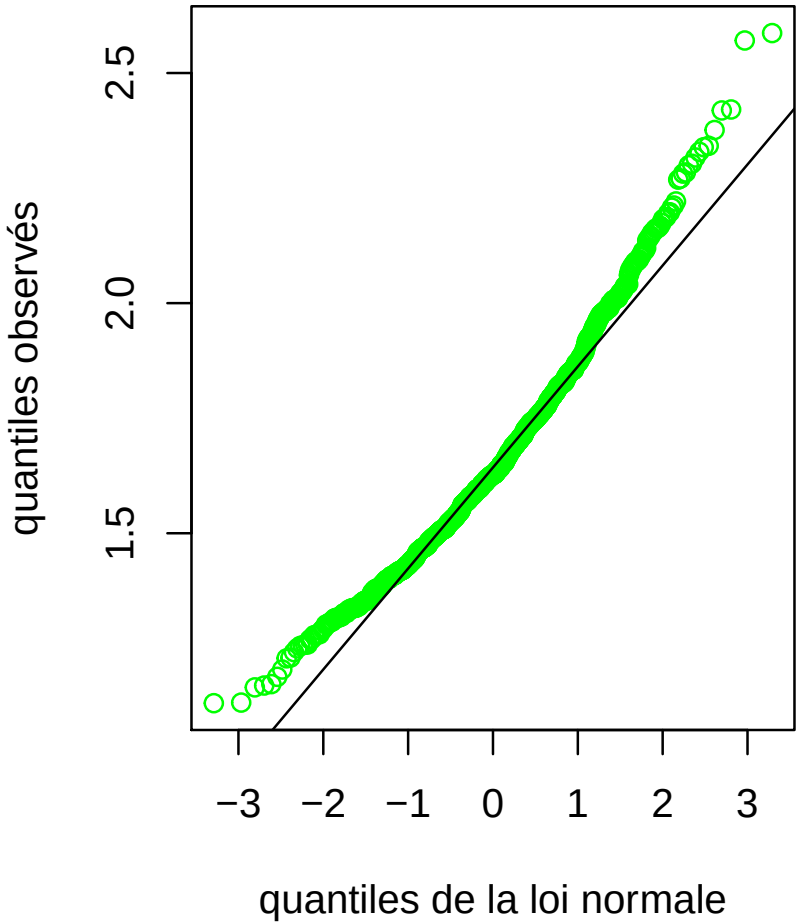
Distribution d'échantillonnage pour $N = 30$



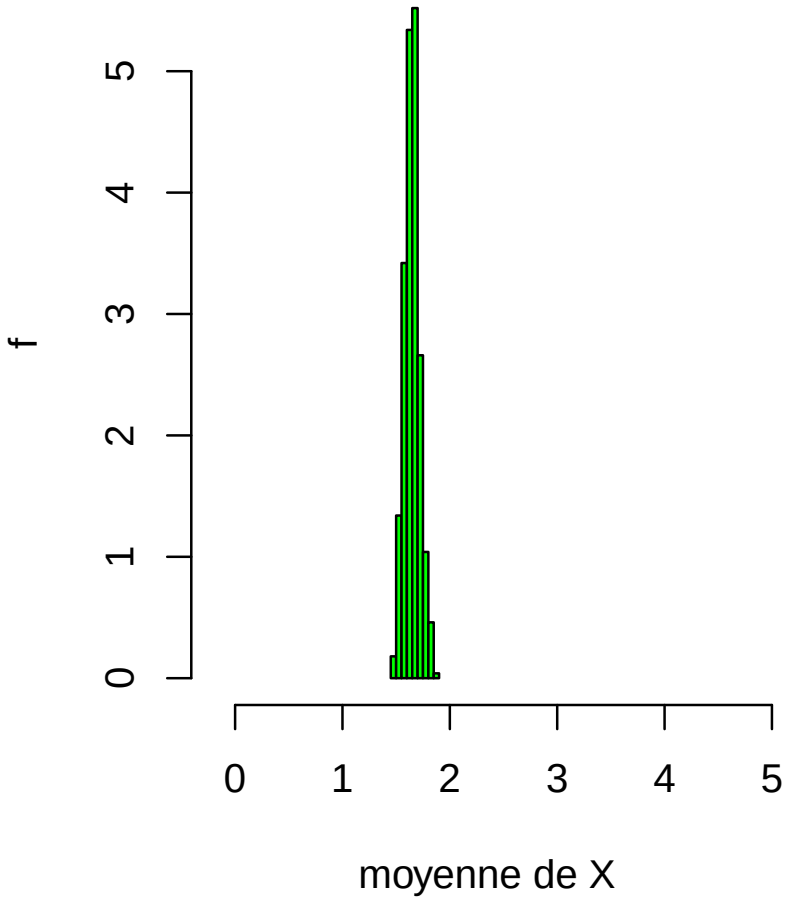


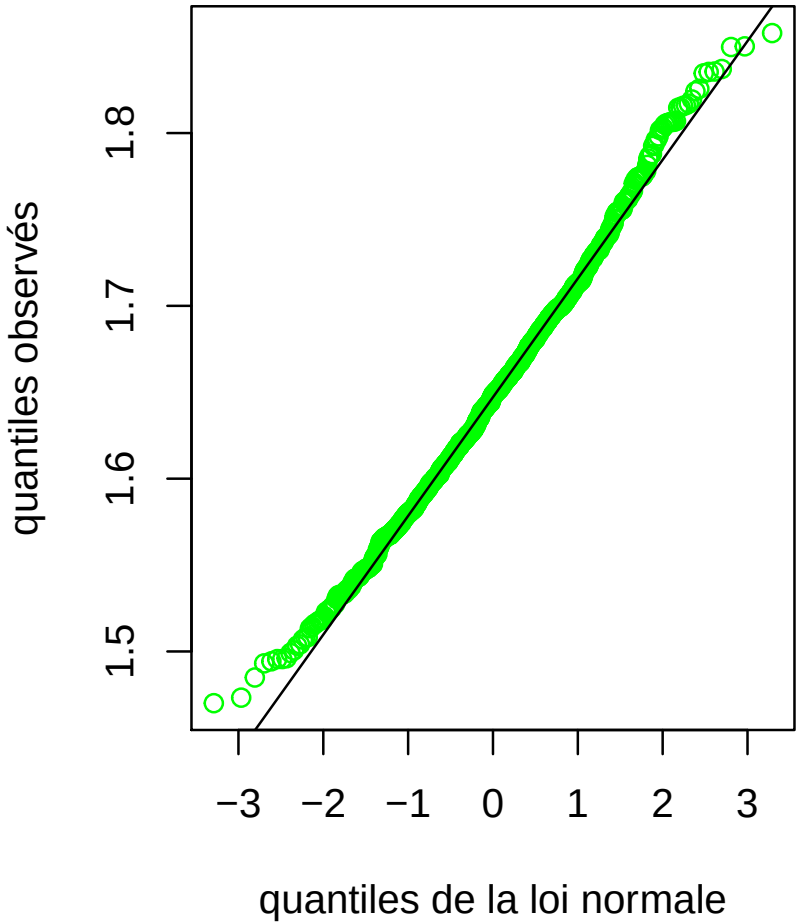
Distribution d'échantillonnage pour $N = 100$





Distribution d'échantillonnage pour $N = 1000$





Théorème de l'approximation normale pour l'estimation d'une moyenne

Théorème de l'approximation normale
Cas d'une variable quantitative

Pour des échantillons aléatoires simples de taille N, la moyenne $\overline{X}$ de l'échantillon varie autour de la moyenne μ de la population avec une erreur standard $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$ noté SE ou SEM ("Standard Error of the Mean"), σ étant l'écart type de la population.

- Lorsque la distribution de X dans la population est normale, $\overline{X}$ suit la loi $N(\mu, \frac{\sigma}{\sqrt{N}})$.
- Quelle que soit la distribution de X, lorsque l'effectif N est suffisamment grand, la loi de $\overline{X}$ s'approche de la loi normale $N(\mu, \frac{\sigma}{\sqrt{N}})$.

Conditions d'application du théorème pour l'estimation d'une moyenne

ATTENTION,
on trouve dans de nombreux ouvrages une condition d'application sous la forme d'un seuil pour N ($N > 30$). Cette condition n'a pas beaucoup de sens.
Il est impossible de juger de l'applicabilité du théorème sans regarder la forme de la distribution.
Comme vu dans les exemples précédents, plus on s'écarte de la loi normale, plus N doit être grand pour appliquer le théorème.

Distribution d'échantillonnage d'une fréquence

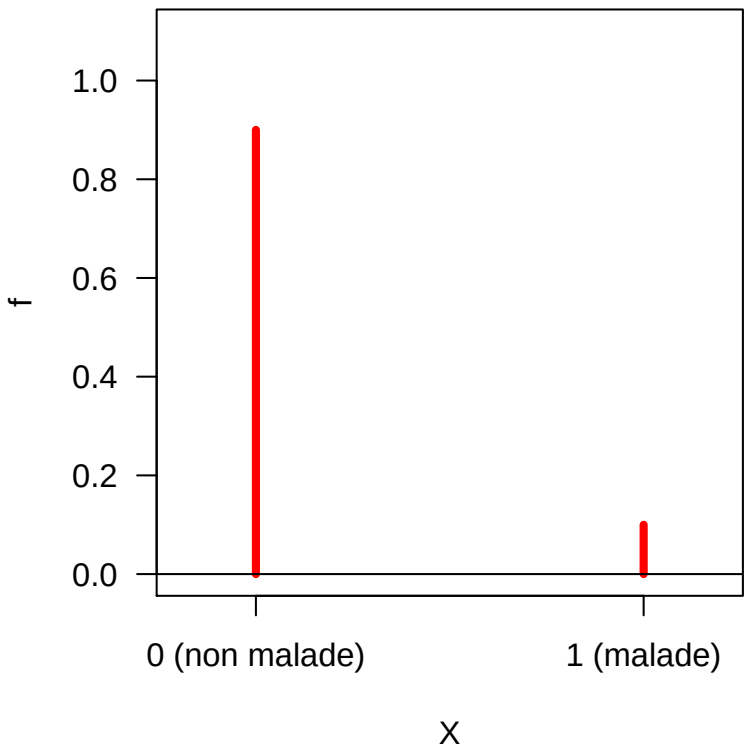
Distribution de la fréquence $F = \frac{n_{malades}}{N}$ dans le cadre de l'étude d'un caractère (ici malade ou non) **dans une population**
La fréquence F est aussi la moyenne de X codant pour la maladie (1 si malade, 0 si non malade),

$$F = \frac{n_{malades}}{N} = \frac{\sum(X_i)}{N} = \overline{X}$$

donc d'après le théorème précédent, sa loi devrait s'approcher d'une loi normale lorsque N devient grand.

Distribution d'échantillonnage pour des tirages dans une loi de Bernouilli - cas d'une variable qualitative bimodale

Distribution de X codant pour une maladie de probabilité p_0 si $p_0 = 0.1$ (10% de malades)

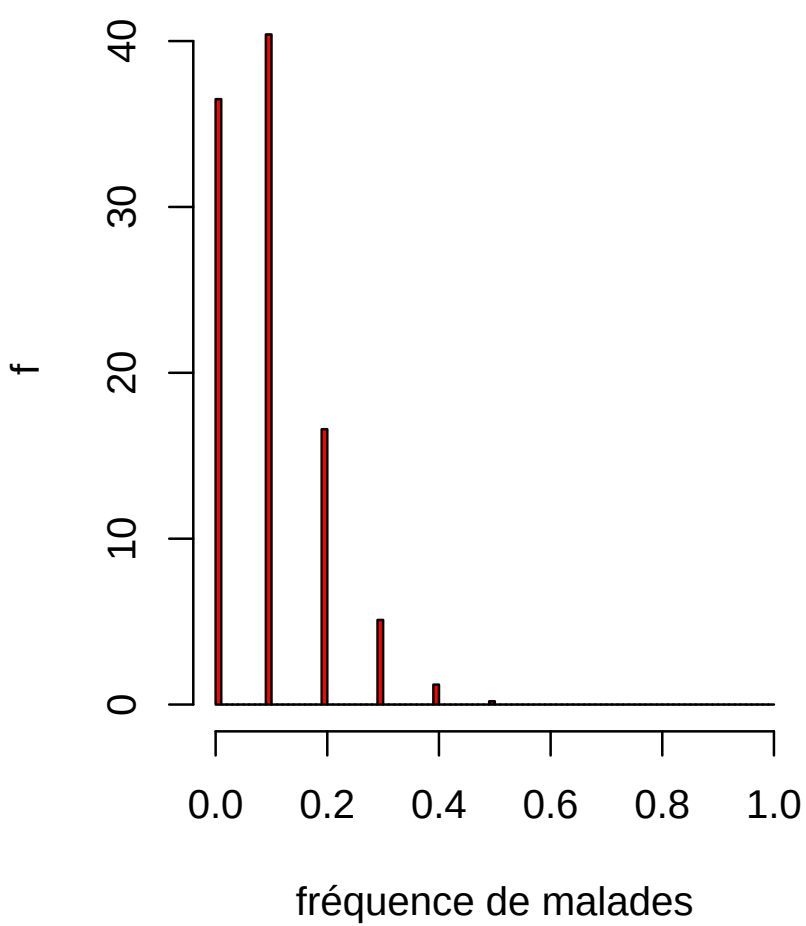


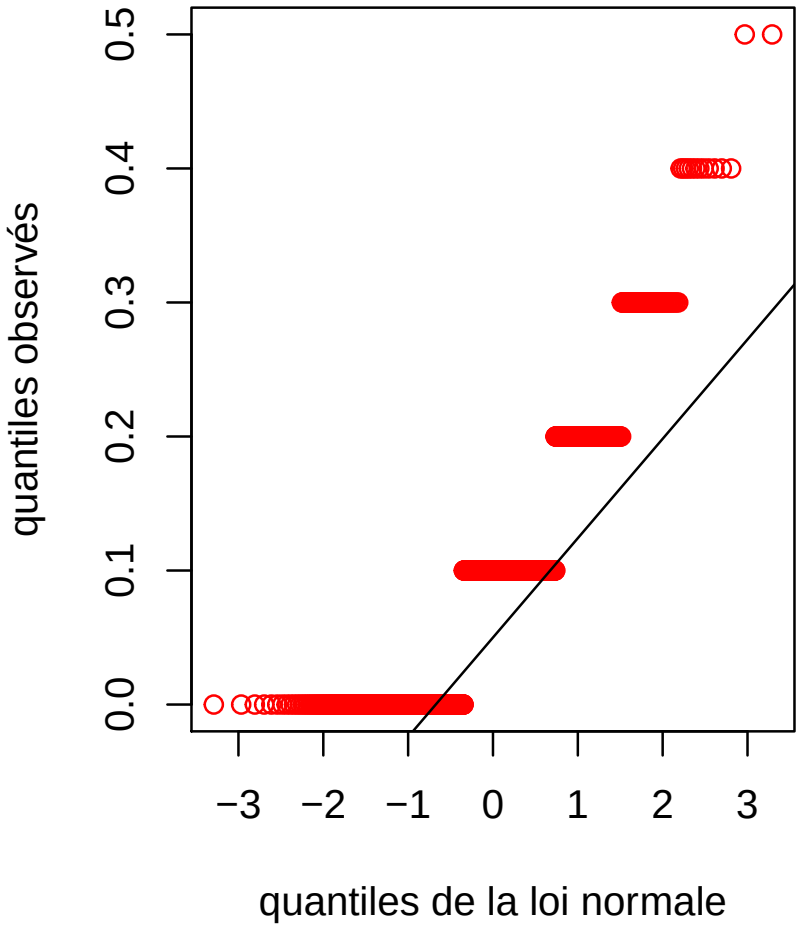
Pour i allant de 1 à 1000

- On tire N individus dans une population contenant 10% de malades.
- On calcule la fréquence observée de malades : F_i

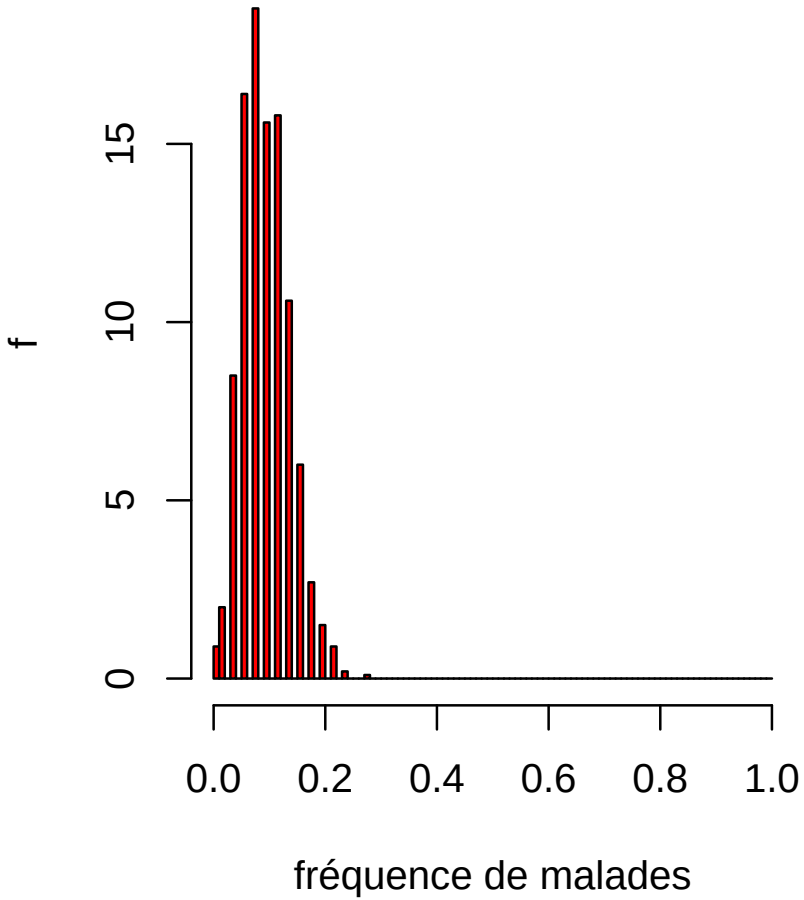
On visualise ensuite la distribution d'échantillonnage de la fréquence de malades (distribution des F_i)

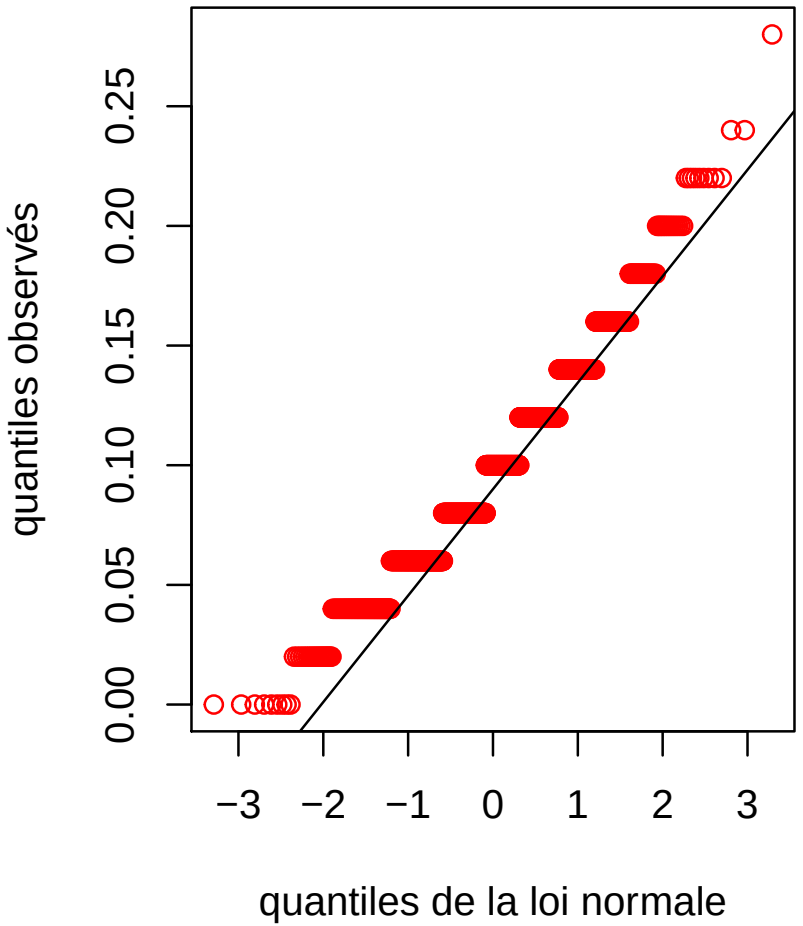
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.1$ et $N = 10$



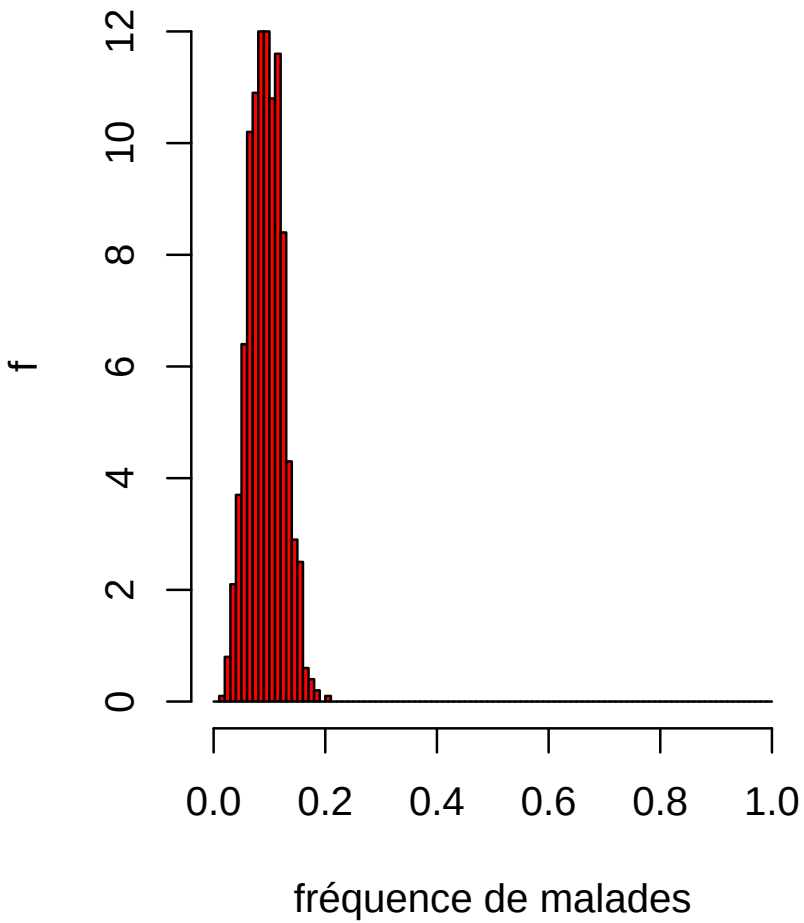


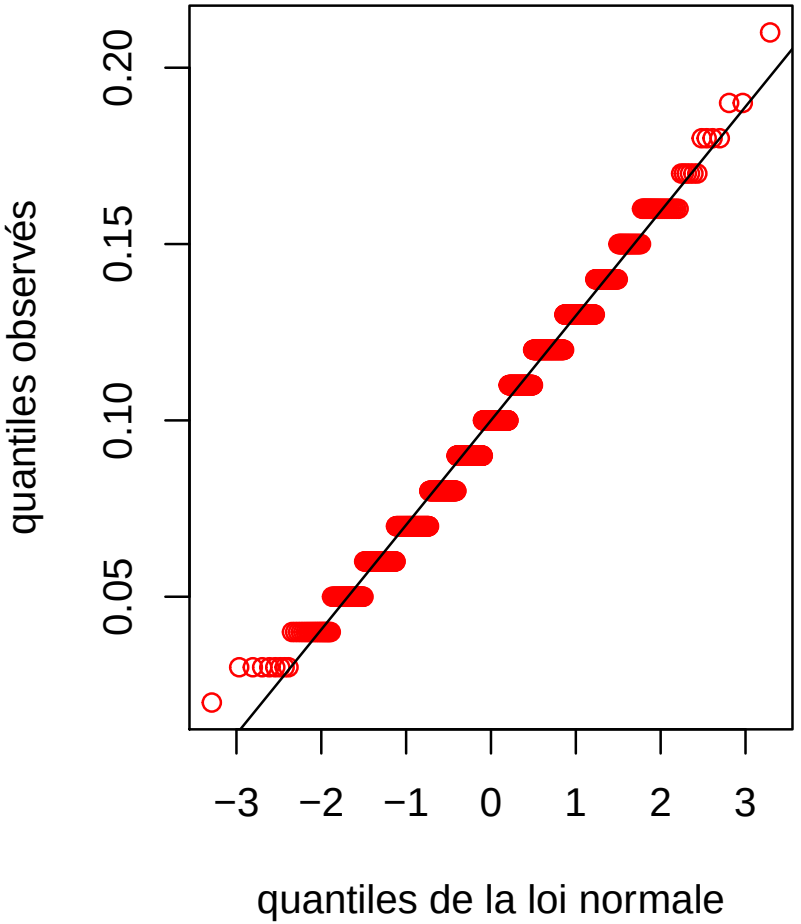
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.1$ et $N = 50$





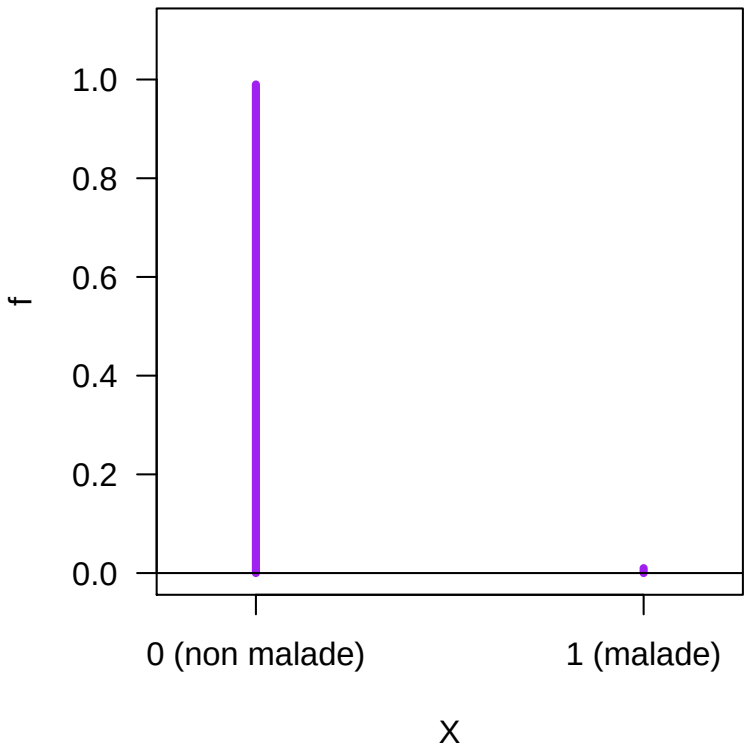
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.1$ et $N = 100$





Distribution d'échantillonnage d'une fréquence d'une maladie plus rare : $p_0 = 1\%$

Distribution de X codant pour une maladie de probabilité p_0 si $p_0 = 0.01$

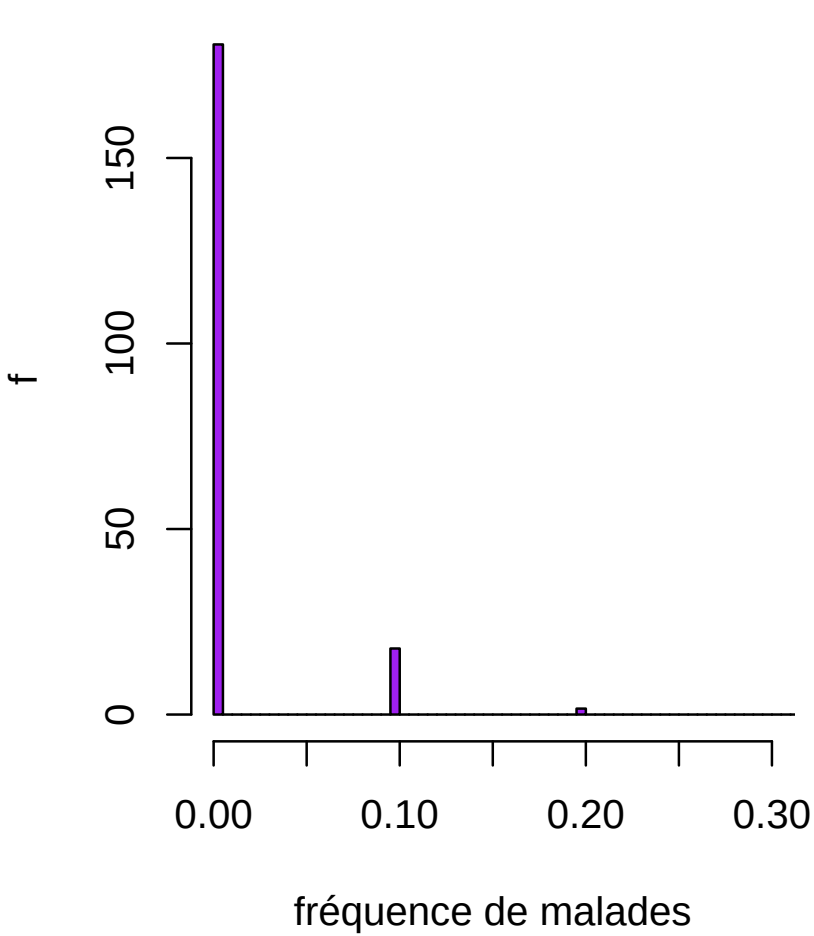


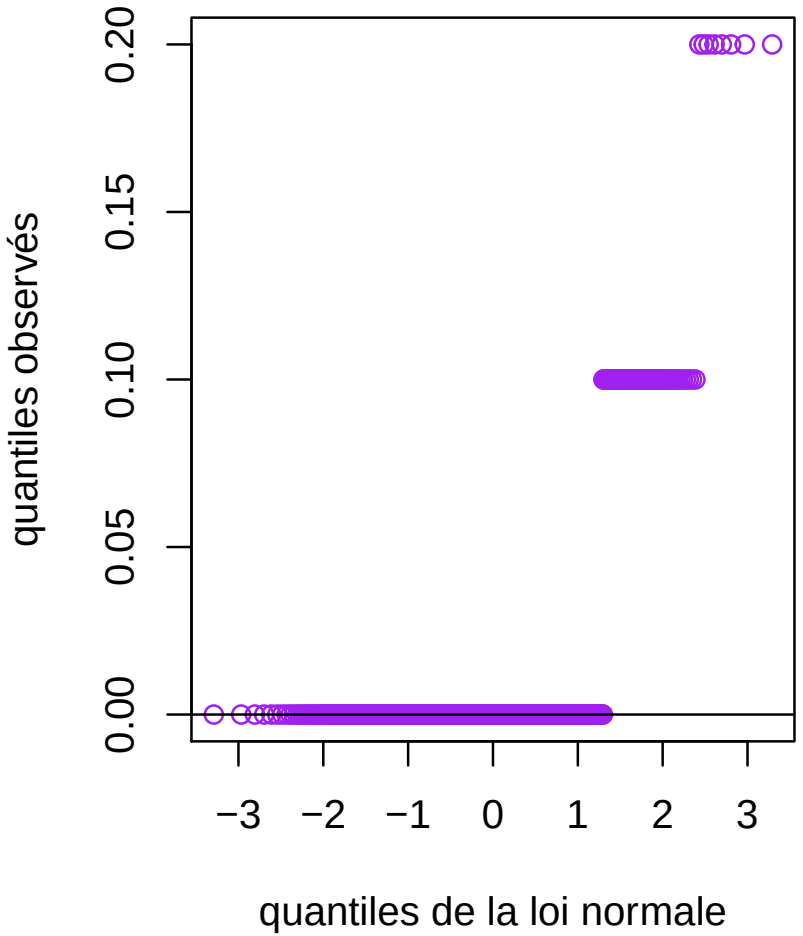
Pour i allant de 1 à 1000

- On tire N individus dans une population contenant 1% de malades.
- On calcule la fréquence observée de malades : F_i

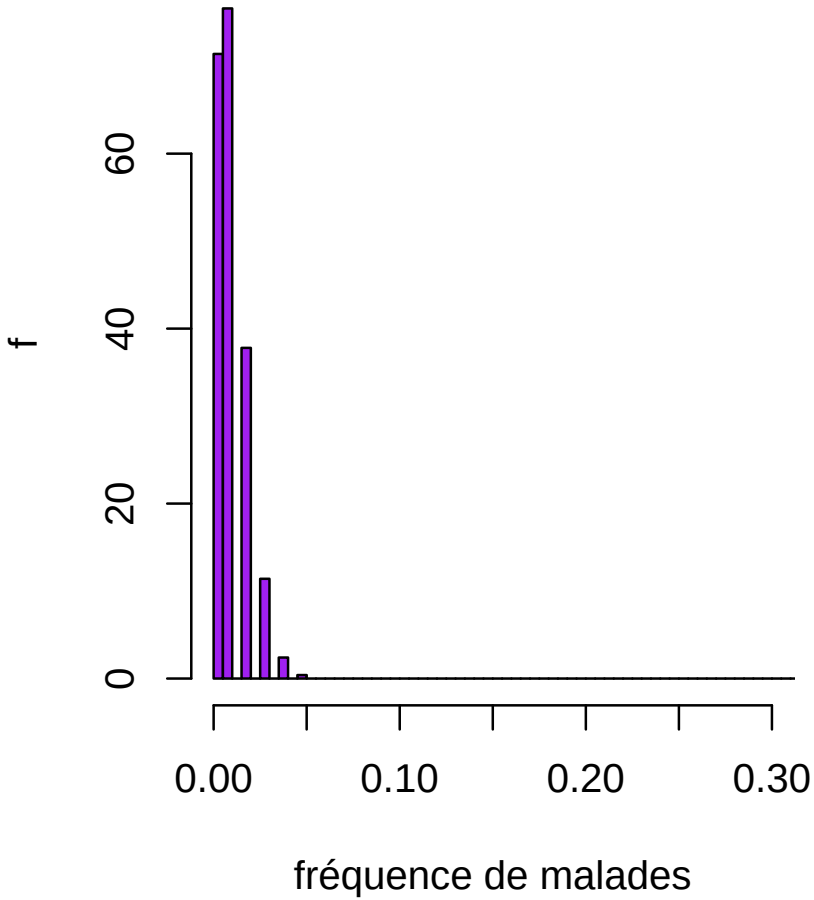
On visualise ensuite la distribution d'échantillonnage de la fréquence de malades (distribution des F_i)

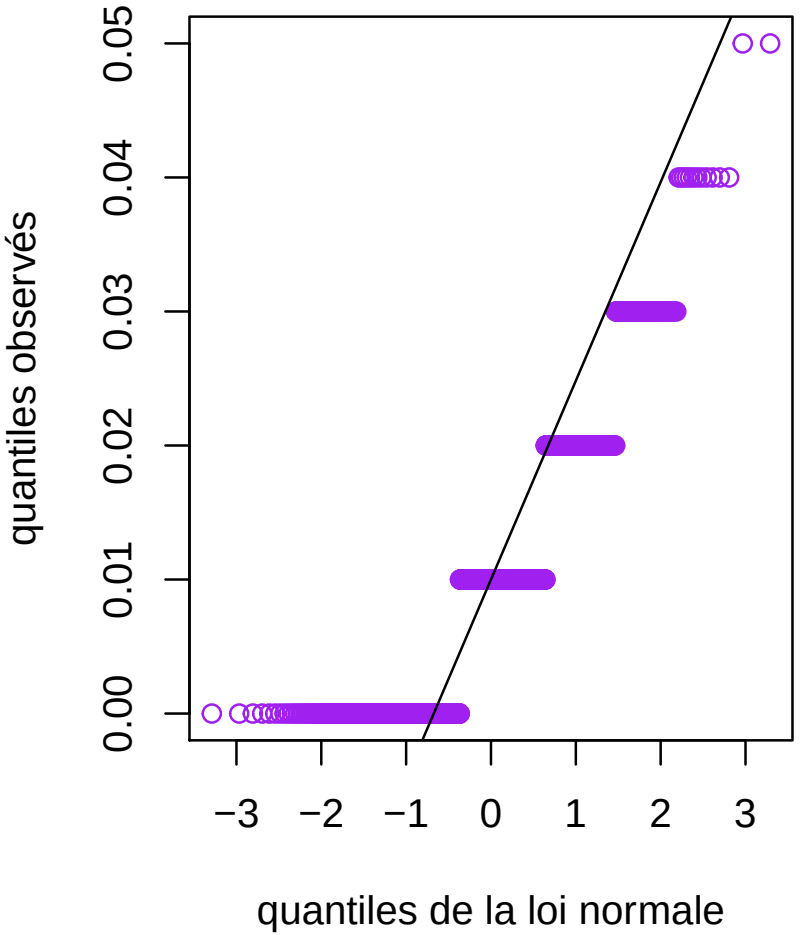
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.01$ et $N = 10$



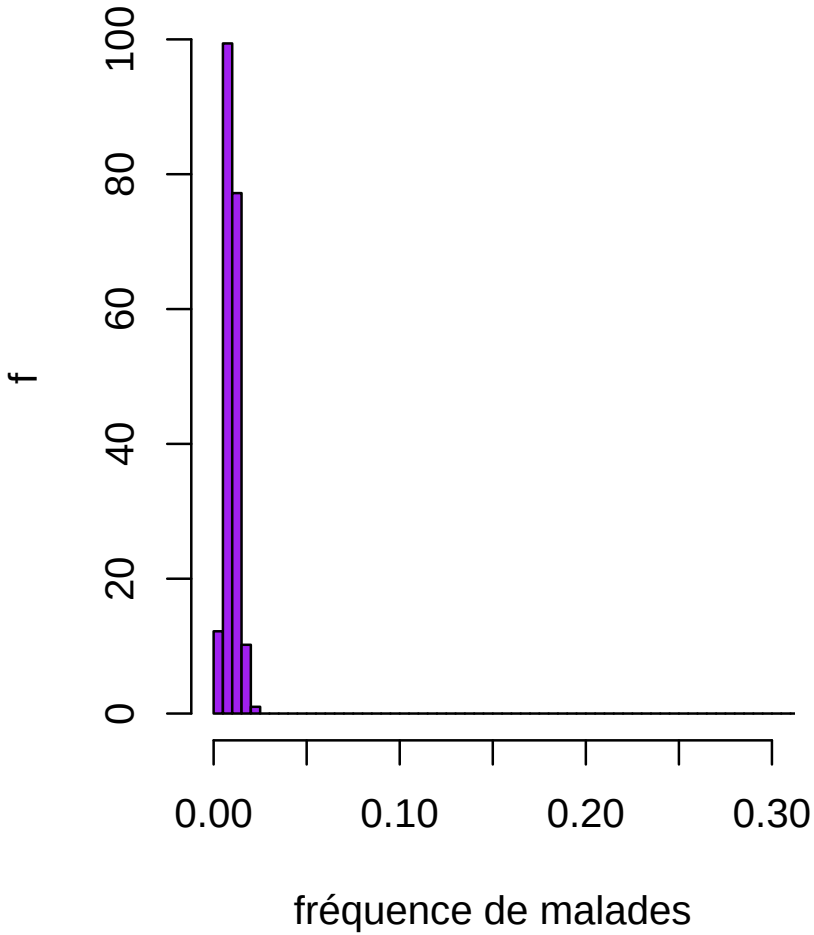


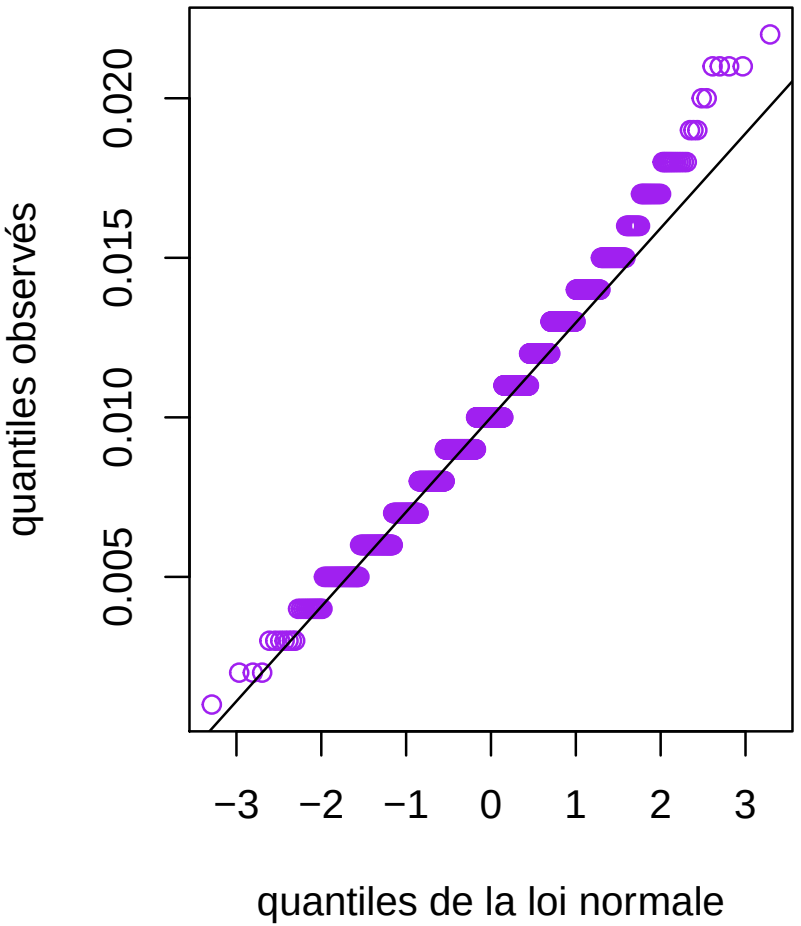
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.01$ et $N = 100$



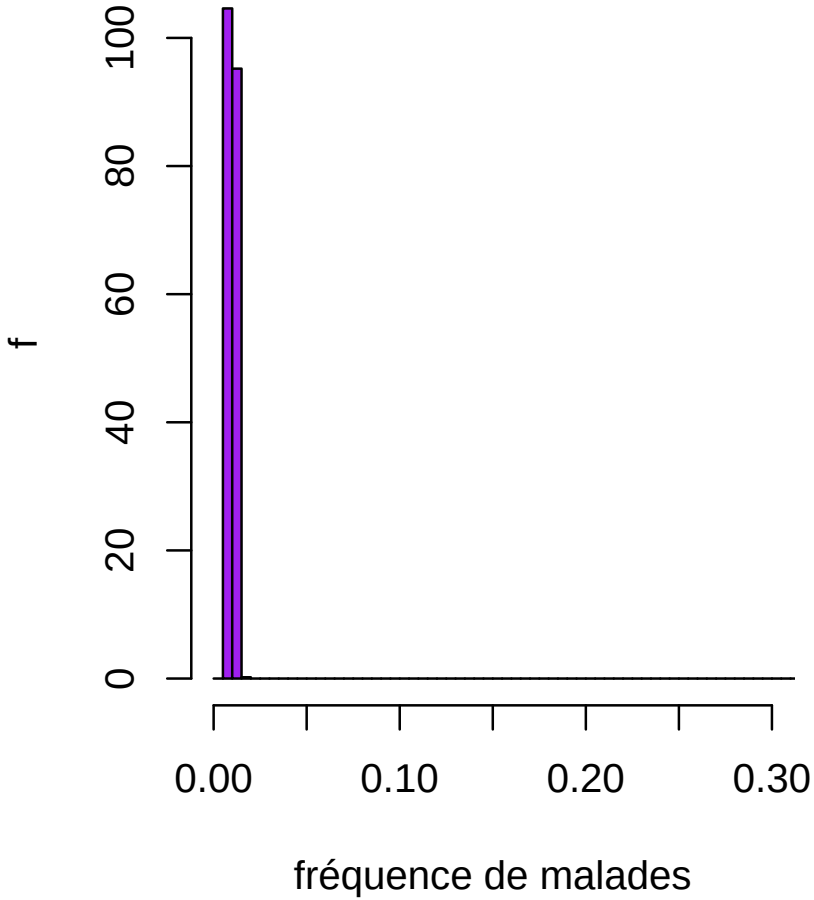


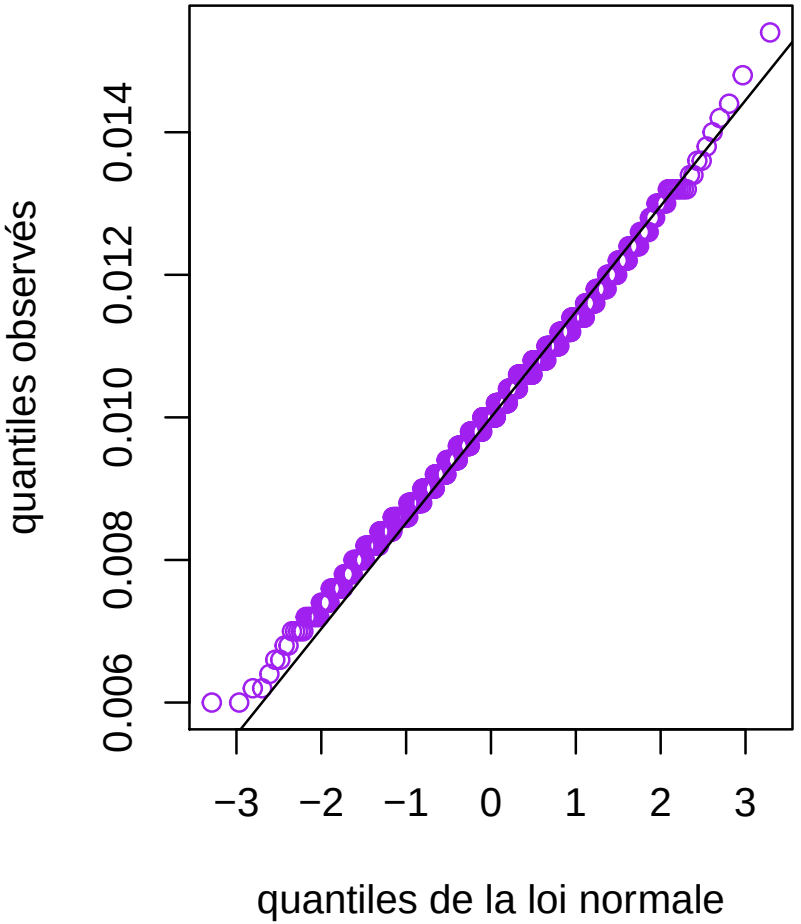
Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.01$ et $N = 1000$





Distribution d'échantillonnage d'une fréquence pour $p_0 = 0.01$ et $N = 5000$





Théorème de l'approximation normale pour l'estimation d'une fréquence

Théorème de l'approximation normale
Cas d'une variable qualitative bimodale

Pour des échantillons aléatoires simples de taille N, la fréquence F d'un caractère étudié varie autour de la proportion p_0 de ce caractère dans la population, avec une erreur standard $\sigma_F = \sqrt{\frac{p_0(1-p_0)}{N}}$.
Lorsque l'effectif N est suffisamment grand, la loi de F s'approche de la loi normale $N(p_0, \sqrt{\frac{p_0(1-p_0)}{N}})$.

Vous avez certainement déjà vu ce résultat dans un cours sur les probabilités (convergence de la loi binomiale vers la loi normale).

Conditions d'application du théorème pour l'estimation d'une fréquence

L'effectif requis pour pouvoir appliquer le théorème de l'approximation normale pour l'estimation de la fréquence F d'un caractère étudié dépend de la proportion p_0 de ce caractère dans la population.
Plus p_0 est proche de 0 (caractère rare) ou de 1 (caratère très répandu), plus N devra être grand.

Plan

Objectif de l'estimation ponctuelle d'un paramètre statistique

Le paramètre θ caractérisant la population étudiée est supposé fixe mais inconnu du fait qu'on n'a pas accès à la population entière.
A partir d'un échantillon de la population, on souhaite estimer au mieux sa vraie valeur sur la population.
On exige souvent d'un estimateur T de θ qu'il soit sans biais, c'est-à-dire qu'en moyenne il ne se trompe pas,
autrement dit que la moyenne de la distribution d'échantillonnage de T soit égale à θ : $E(T) = \theta$.

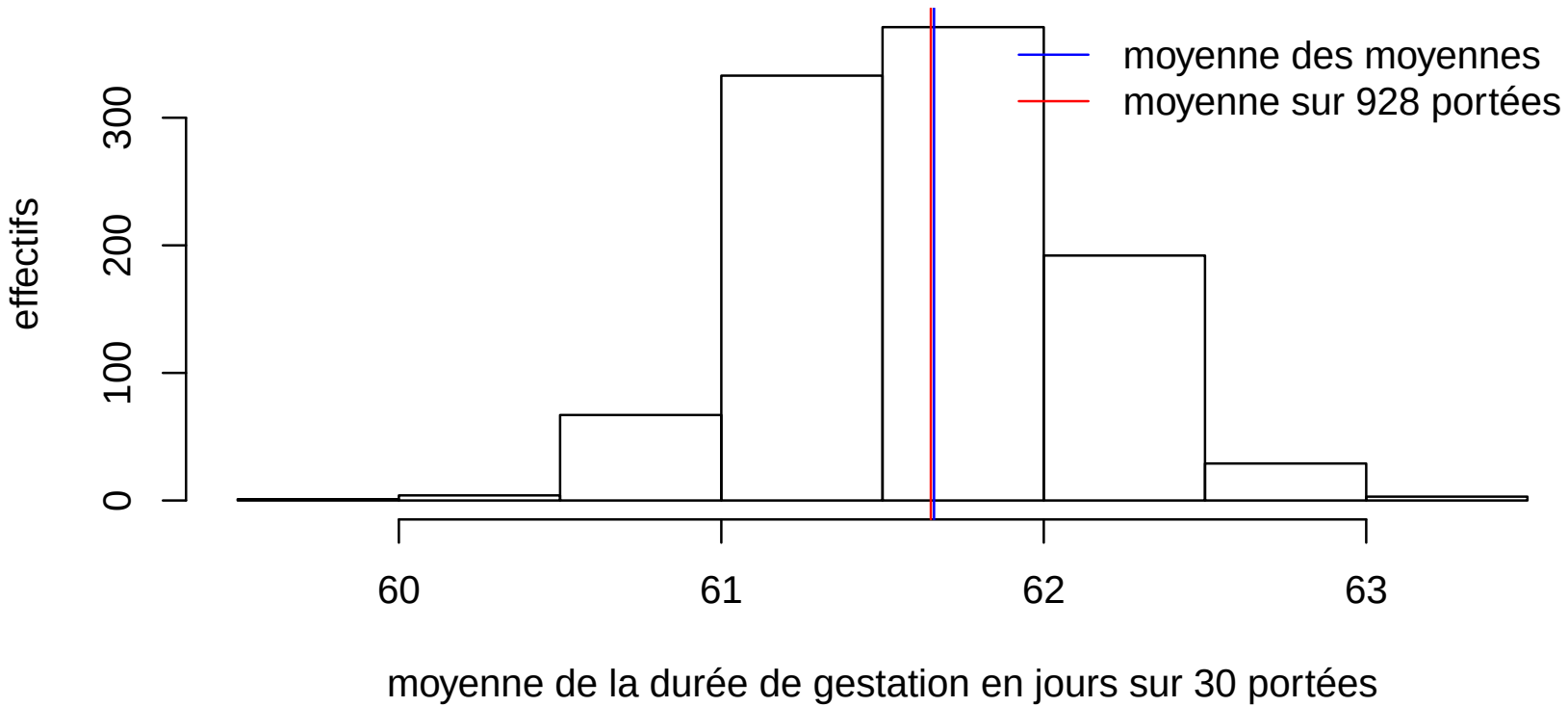
Estimation sans biais d'une moyenne μ

D'après le théorème de l'approximation normale, la moyenne d'une variable quantitative calculée à partir d'un échantillon aléatoire simple de la population est un estimateur sans biais de la moyenne de la population : $E(\overline{X}) = \mu$.

$$\hat{\mu} = \overline{X}$$

Exemple de l'estimation de la durée de gestation de chiennes d'élevage à partir d'un échantillon de 30 portées

Distribution d'échantillonnage de la moyenne : histogramme sur 1000 échantillons de 30 portées tirées au hasard parmi les 928 observées dans la thèse vétérinaire de Mathilde Poinssot.



Estimation sans biais d'une fréquence p_0

D'après le théorème de l'approximation normale, la fréquence F d'un caratère étudiée, calculée à partir d'un échantillon aléatoire simple de la population, est un estimateur sans biais de la fréquence de ce caratère dans la population : $E(F) = p_0$.

$$\hat{p_0} = F$$

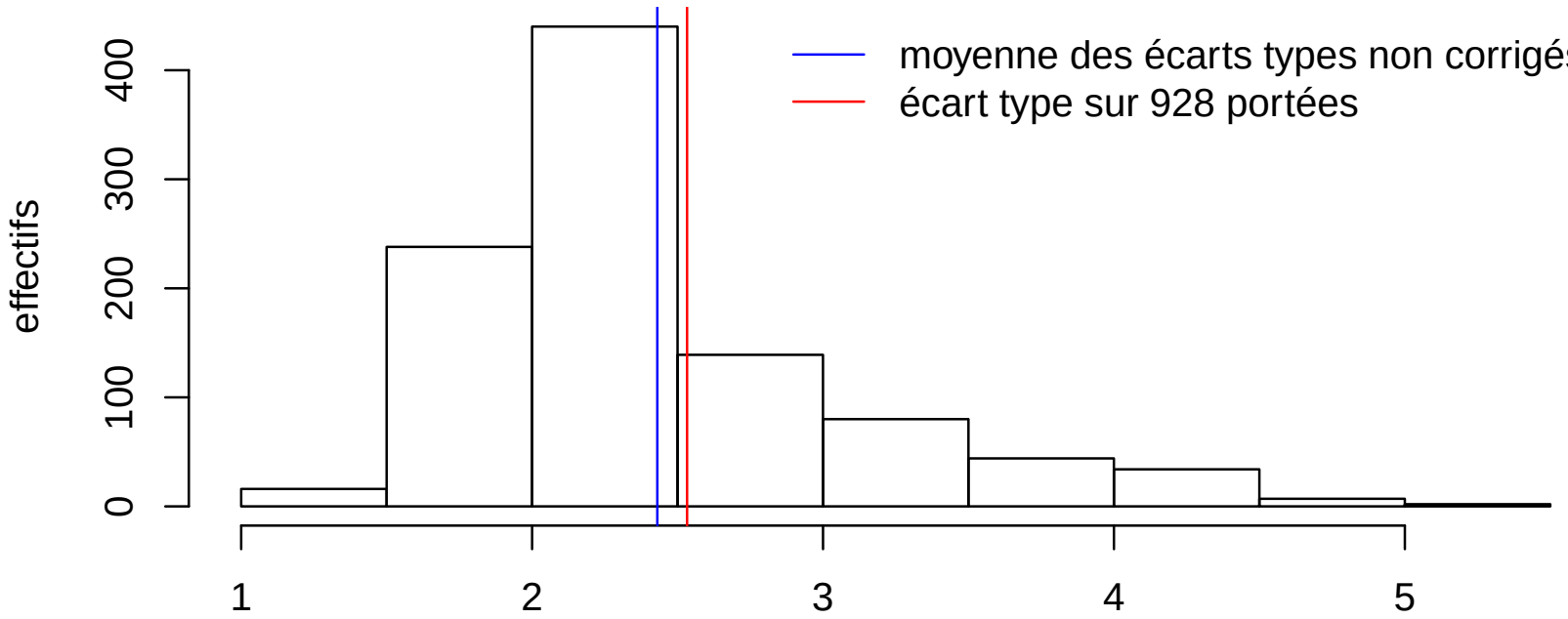
Estimation sans biais d'une variance σ^2

On peut montrer que la variance $V(X) = \frac{1}{N}\sum_{k=1}^{N}(x_i - \overline{x})^2$ d'une variable quantitative calculée à partir d'un échantillon aléatoire simple de taille N de la population est un estimateur biaisé de la variance : $E(V(X)) = \frac{N-1}{N}\sigma^2$.
On obtient un estimateur sans biais de la variance en corrigeant le biais :

$$\hat{\sigma^2} = \frac{N}{N-1}V(X) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i - \overline{x})^2$$

Exemple de l'estimation de la durée de gestation de chiennes d'élevage à partir d'un échantillon de 30 portées

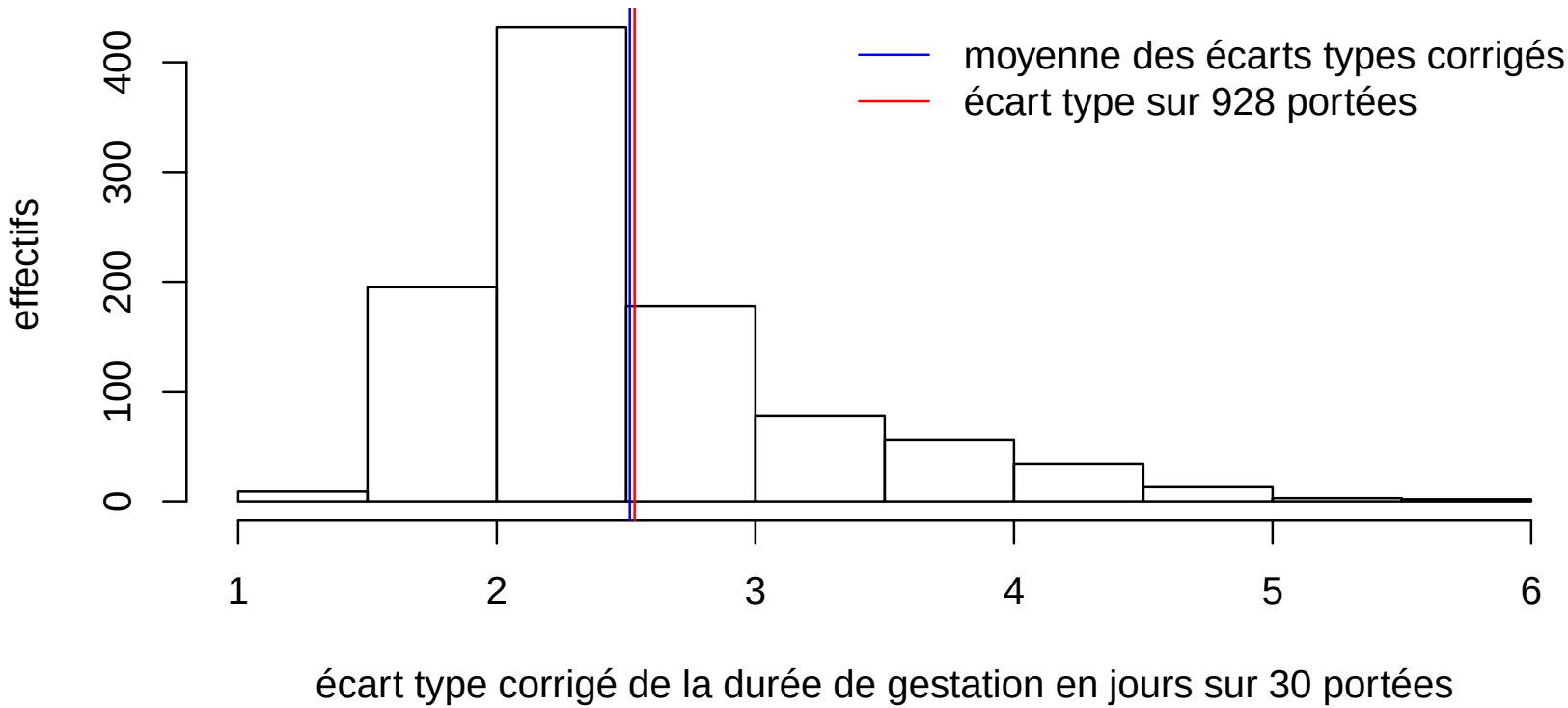
Distribution d'échantillonnage de l'écart type non corrigé : histogramme sur 1000 échantillons de 30 portées tirées au hasard parmi les 928 observées dans la thèse vétérinaire de Mathilde Poinssot.



écart type non corrigé de la durée de gestation en jours sur 30 portées

Exemple de l'estimation de la durée de gestation de chiennes d'élevage à partir d'un échantillon de 30 portées

Distribution d'échantillonnage de l'écart type corrigé : histogramme sur 1000 échantillons de 30 portées tirées au hasard parmi les 928 observées dans la thèse vétérinaire de Mathilde Poinssot.



Notion d'intervalle de confiance

Quelle est la précision de l'estimation ponctuelle ?
Quelle confiance peut-on accorder à une estimation sur un échantillon unique (pratique courante) ?
Comment répondre à cette question sans répéter l'échantillonnage ?
En construisant un **intervalle de confiance** autour de l'estimation.

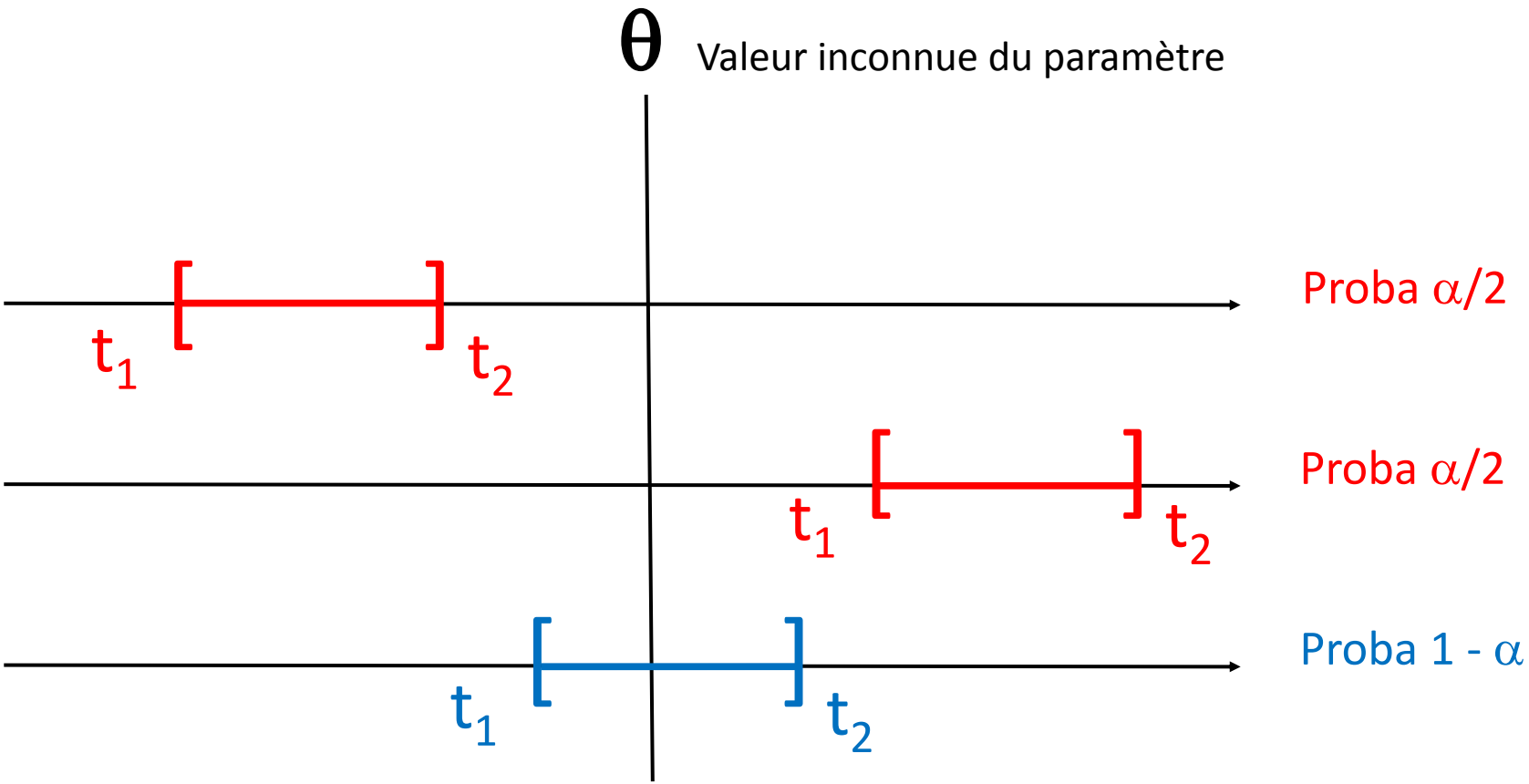
Définition d'un intervalle de confiance bilatéral

Intervalle $[t_1; t_2]$
construit de façon à ce qu'**en terme de distribution d'échantillonnage** (sous-entendu si on répétait l'échantillonnage)

$$Pr(t_1 \geq \theta) = Pr(t_2 \leq \theta) = \frac{\alpha}{2}$$
$$\text{donc } Pr(t_1 \leq \theta \leq t_2) = 1 - \alpha$$

t_1 et t_2 sont appelées les limites de confiance.
$1 - \alpha$ est appelé le seuil de confiance.
Généralement α est fixé à 5% et l'on parle d'**intervalles de confiance à** 95%.

Illustration de la définition d'un intervalle de confiance bilatéral



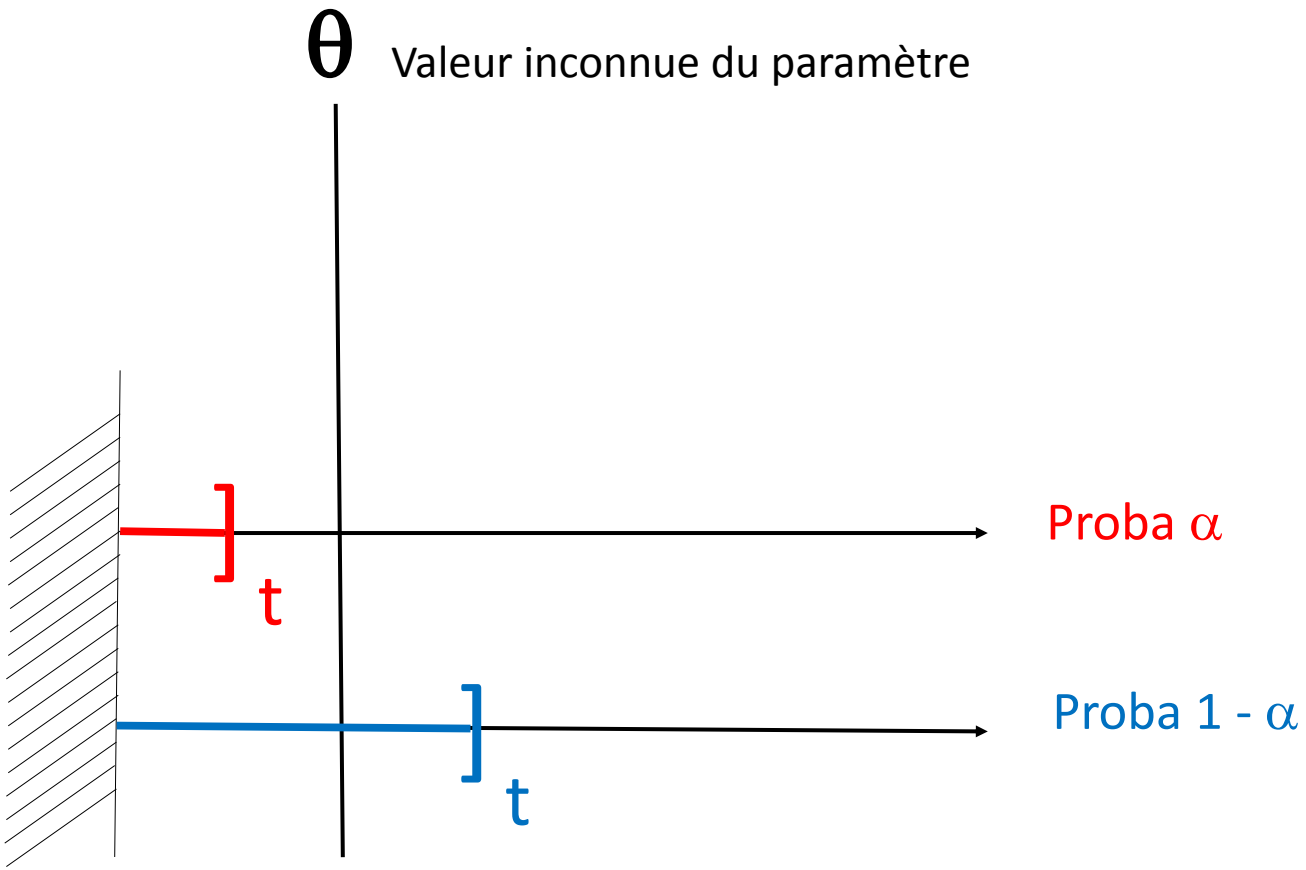
Définition d'intervalles de confiance unilatéraux

Dans certains cas particuliers on définira des intervalles de confiance unilatéraux (une seule limite de confiance) ayant toujours une probabilité $1 - \alpha$ de contenir la vraie valeur du paramètre.

Exemple classiques :

- calcul du seuil au dessous duquel on veut pouvoir dire avec une confiance de 95% que se trouve une proportion d'animaux malades dans un pays :
  Intervalle du type $[0, t]$.
- calcul du seuil au dessus duquel on veut pouvoir dire avec une confiance de 95% que se trouve la sensibilité d'un test diagnostique :
  Intervalle du type $[t, 1]$.

Illustration de la définition d'un intervalle de confiance unilatéral



Calcul d'un intervalle de confiance

A partir du théorème de l'approximation normale et/ou d'autres résultats de la statistique théorique des intervalles de confiance ont été proposés pour les cas classiques, par exemple :

Intervalle de confiance bilatéral autour d'une fréquence

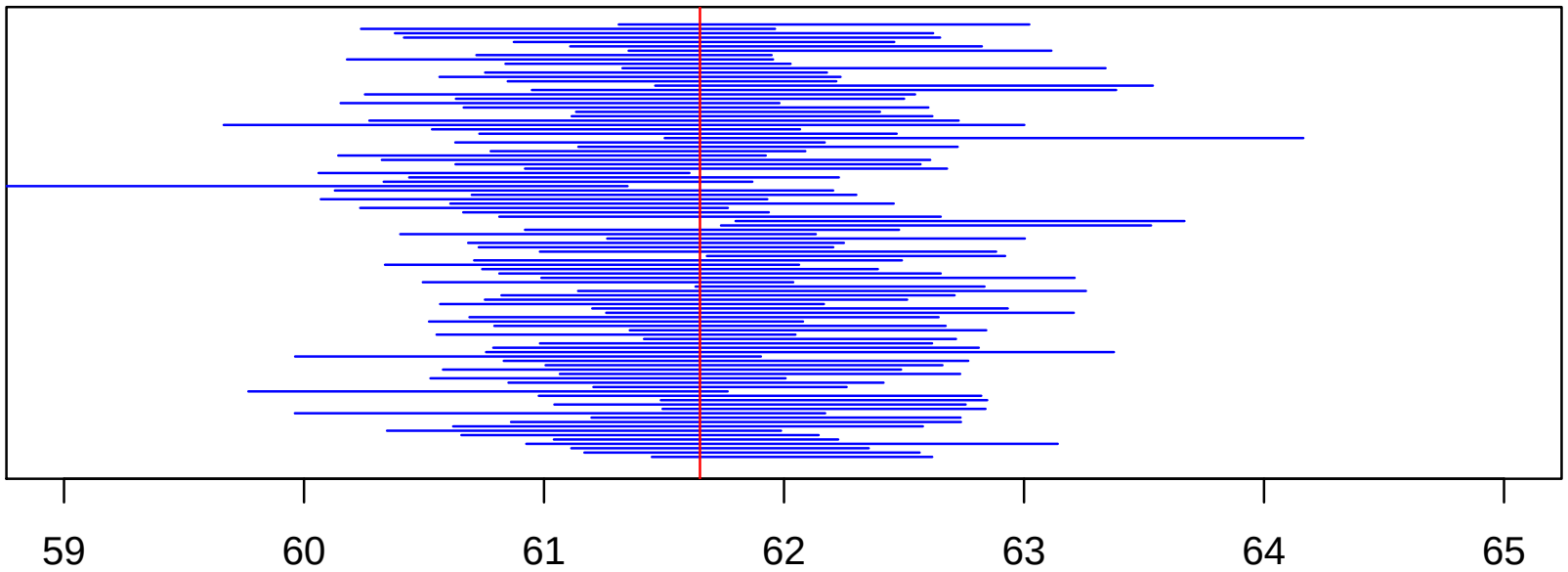
$p_0 = f \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f(1-f)}{N}}$
avec $u_{1-\frac{\alpha}{2}}$ le quantile à $1-\frac{\alpha}{2}$ de la distribution normale $N(0,1)$.
utilisable si $nf \geq 20$ et $n(1-f) \geq 20$

Intervalle de confiance bilatéral autour d'une moyenne

$\mu = \overline{x} \pm t_{N-1;1-\frac{\alpha}{2}} \times \frac{\hat{\sigma}}{\sqrt{N}}$
avec $t_{N-1;1-\frac{\alpha}{2}}$ le quantile à $1-\frac{\alpha}{2}$ de la distribution de Student de degré de liberté $N-1$ (T_{N-1}).
utilisable si théorème de l'approximation normale applicable

Calcul d'intervalles de confiance à 95% sur la durée moyenne de gestation à partir d'un échantillon de 30 portées

Visualisation des intervalles de confiance calculés sur 100 échantillons et de la vraie valeur du paramètre (rouge)



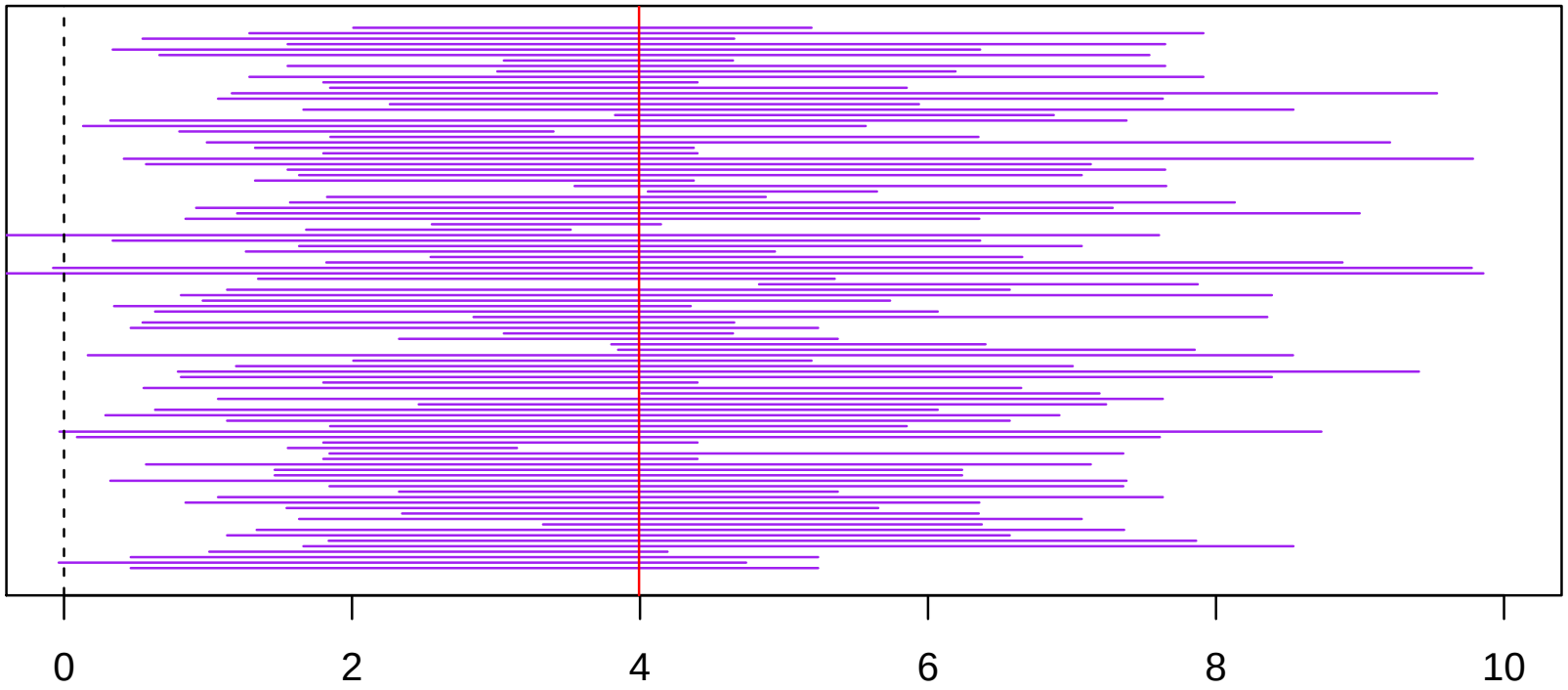
estimation par intervalle de la durée moyenne de gestation en jours

Interprétation d'un intervalle de confiance

En pratique on calcule un intervalle de confiance sur un seul échantillon.
On n'a donc aucun moyen de savoir si cet intervalle de confiance contient bien la vraie valeur du paramètre.
On peut juste se dire qu'en moyenne, lorsqu'on calcule des intervalles de confiance à 95%, on se trompe une fois sur 20 (5% des échantillons).

Respect impératif des conditions d'utilisation ! (1)

Mauvais exemple de calcul d'intervalles de confiance sur la moyenne des âge la mise bas sur 100 échantillons de 4 chiennes (avec quelques intervalles aberrants du fait de la non applicabilité du théorème de l'approximation normale)



estimation par intervalle des âges à la mise bas

Respect impératif des conditions d'utilisation ! (2)

Autre mauvais exemple de calcul de l'intervalle de confiance à 95% sur la prévalence d'une maladie à partir d'un échantillon de 100 animaux sur lesquels 2 sont malades.
Si on utilise la formule donnée précédemment
$(p_0 = f \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f(1-f)}{N}})$ sans vérifier ses conditions d'utilisation (non applicables ici : $nf = 2 < 20$) on obtient
$0.02 \pm 1.96 \times \sqrt{\frac{0.02 \times 0.98}{100}} = 0.02 \pm 0.0274 = [-0.0074; 0.0474]$
soit un intervalle erroné comprenant même des valeurs négatives.
A l'aide du logiciel **R** on pourra plus tard calculer correctement cet intervalle de confiance à l'aide la loi binomiale et on obtiendra un intervalle dissymétrique : [0.0024 ; 0.0704]

Conclusion

Il est très important de savoir juger, à partir d'un échantillon, du respect des conditions d'application du théorème de l'approximation normale. De très nombreux outils statistiques (estimateurs ponctuels et par intervalle, test statistiques) sont basés sur le théorème de l'approximation normale et nécessitent donc la vérification au préalable de ses conditions d'utilisation.
Il est IMPORTANT de se souvenir que la vérification des ces conditions d'utilisation ne peut pas se faire en regardant uniquement la taille de l'échantillon pour une variable quantitative. L'examen de la distribution est indispensable.