

# Description et inférence: formulaire, définitions, démonstrations, lois CM1 et 2

Julien CLAUDE

5 septembre 2022

## 1 Définitions générales

- **Population** : un ensemble d'êtres vivants ou d'objets à partir desquels on pense pouvoir appliquer des lois et définir des propriétés.
- **Echantillon** : un groupe restreint d'individus dans la population
- **Individu** : en statistique, c'est un objet dans l'échantillon. La ou les propriétés de cet objet forme(nt) une observation.
- **Variable statistique** : quantité, propriété ou caractère défini sur une population et qui diffère selon les individus (ou encore les observations)
- **Variable qualitative** : variable dont peut nommer les différents états (ou catégories). On parle parfois de facteur.
- **Variable ordonnée** : les états peuvent être rangés dans un ordre ascendant ou descendant.
- **Variable non ordonnée** : pas d'ordre possible entre les états.
- **Variable quantitative** : variable définie par des nombres.
- **Variable quantitative discrète** : variable dont les états prennent un nombre limité de valeurs numériques.
- **Variable quantitative continue** : variable dont les états possibles sont en théorie infini.

## 2 Distribution, fréquences, fréquences cumulées, quantiles et polygone des fréquences cumulées

- **Distribution** : Ensemble des états d'une variable associés à leurs effectifs. La distribution associe pour chaque modalité ou valeur  $x_i$  de la variable  $X$ , un effectif  $n_i$ ,  $i$  variant de 1 à  $p$ ;  $p$  étant le nombre possible de catégorie.
- **Taille de l'échantillon** =  $\sum_i^p n_i = N$ .
- **Fréquences** : proportion associées à chaque catégorie  $f_i$  :  $f_i = n_i/N$ ;  $f_i$  en pourcents  $\in [0, 100]$ .
- **Tableau d'effectifs** : Tableau qui associe pour chaque effectif  $n_i$ , une modalité  $x_i$ .
- **Tableau de fréquences** : Tableau qui associe pour chaque fréquence  $f_i$ , une modalité  $x_i$ .
- **Tableau de fréquences cumulées** : Tableau qui associe pour chaque fréquence cumulée  $F_i$ , une modalité  $x_i$
- **Fréquences cumulées** : Soit une distribution  $(x_i, n_i)$ , avec  $i \in [1, p]$ , la fréquence cumulée d'ordre  $k$   $F_k$  est la somme des  $k$  premières fréquences.

$$F_k = f_1 + f_2 + \dots + f_k = \sum_i^k f_i$$

- **Quantile** : valeur de  $x_k$  de  $X$  atteignant la fréquence cumulée  $F_k$ . On peut écrire le quantile  $x_k$  de  $X$  est défini de sorte que que  $P(X \leq x_k) = F_k$ . C'est alors le quantile à  $F_k \times 100\%$ .
- **Tableau de contingence à deux facteurs** : Tableau qui recense l'effectif pour la combinaison de deux facteurs (par exemple  $X$  et  $Y$  avec  $X$  prenant des valeurs de  $x_1$  à  $x_p$  et  $Y$  prenant des valeurs de  $y_1$  à  $y_q$ ).
- **Longueur ou amplitude d'une classe pour une variable quantitative** : différence entre la valeur élevée et la valeur minimale des bornes de la classe  $[x_i, x_i]$ . Si toutes les classes ont la même longueur, alors la longueur de la classe est alors appelé le pas et celui ci vaut :  $\frac{x_p - x_0}{p}$ .
- **Polygone des fréquences cumulées pour une variable quantitative**  
Soit une variable  $X$ , découpée en classes  $[x_0, x_1], [x_1, x_2], \dots, [x_{p-1}, x_p]$ , et associées au fréquences cumulées  $F_1, F_2, \dots, F_p$ , le polygone des fréquences cumulées est la courbe polygonale qui associe les couples  $(x_k, F_k)$ .

- **Interpolation linéaire à partir du Polygone des fréquences cumulées.** Connaissant la fréquence cumulée  $F_a$  et que  $F_a$  est encadrée par  $F_{k-1}$  et  $F_k$ , associées à  $x_k$  et  $x_{k-1}$ , on peut associer la valeur  $x_a$  correspondant par interpolation linéaire.

$$x_a = x_{k-1} + \frac{F_a - F_{k-1}}{F_k - F_{k-1}} \times (x_k - x_{k-1}) \quad \text{avec } F_{k-1} < F_a < F_k$$

Ainsi on peut définir la **médiane** correspond qui à la valeur  $x$  de la variable aléatoire  $X$  pour laquelle 50 % de la distribution de  $x$  est atteinte. On dit de la médiane que c'est le quantile à 50 %. On peut aussi définir le **premier et dernier quartiles** comme les quantiles à 25% et 75%. La différences entre le premier et dernier quartile porte le nom d'**intervalle interquartile**.

Les quantiles permettent dans le cas de certaines distributions de détecter des **valeurs extrêmes**. Le graphique de type boîtes à moustaches classique représente la médiane, l'intervalle entre les quartiles :  $IQ$  sous forme de boîte, et le minimum et le maximum sous forme de "moustache" pourvu que ceux ci soit compris dans l'intervalle  $[X_{25\%} - 1.5 \times IQ; X_{75\%} + 1.5 \times IQ]$ ; si des valeurs sont en dehors de cet intervalle, elles peuvent être considérées comme des valeurs extrêmes (dites encore outliers), il convient alors de s'assurer que ces valeurs ne soient pas abhèrentes (erreurs de mesure, erreurs de saisie, ...etc).

### 3 Moyenne

La **moyenne** est définie par la quantité

$$\mu_x = \frac{1}{n} \sum_{i=1}^p n_i x_i \quad \text{avec } \sum_i n_i = n$$

Elle peut aussi être écrite d'après la fréquence ou la probabilité de chacune des observations :

$$\mu_x = \sum_i^p f_i x_i$$

La moyenne est associée à la notion d'**espérance**. L'espérance est le résultat moyen d'un événement aléatoire. Si la probabilité de l'évènement  $x_i$  est  $p_i$ , alors l'espérance est :

$$\mathbb{E}(X) = \sum_{i=1}^n p_i x_i$$

L'estimateur de l'espérance est moyenne empirique.

### 4 Théorème centrale limite et loi des grands nombres

- **Loi des grands nombres** : Cette loi stipule que plus on augmente la taille de l'échantillon, plus les caractères statistiques de l'échantillon se rapproche des caractères de la population.
- **Théorème centrale limite** : Si  $X_1, X_2, \dots, X_n$  est une suite de variables aléatoires appartenant à des lois de distribution identiques avec une espérance  $\mu$  et un écart type  $\sigma$ , et que ces variables sont indépendantes, alors la somme  $S_n = X_1 + X_2 + \dots + X_n$  a une espérance  $n\mu$  et l'écart type de cette somme faut  $\sigma\sqrt{n}$ . 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 aussi écrire d'après ce théorème que :

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

Par ailleurs, grace à ce théorème et au delà d'un certains effectifs, la plupart des lois peuvent être approchées par une loi normale sous condition d'indépendance (voir application pour la construction de l'intervalle de confiance pour une proportion).

- **L'estimateur de la moyenne pour un échantillon** est donné par la moyenne arithmétique donc par

$$\hat{\mu}_x = \sum_i^p f_i x_i$$

Par la suite on utilisera l'accent circonflexe pour montrer qu'on fait une estimation à partir d'un échantillon.

- **Extensions** : Le théorème centrale limite (élaboré par Laplace au 19eme siècle) a été généralisé en particulier pour les conditions dites de Lyapounov. Ainsi, pour tendre vers une loi normale, les lois de probabilité que suivent chaque variables aléatoires, si elles sont indépendantes, n'ont même pas besoin d'être identiques, à condition qu'aucune de ces variables aléatoires ne soit prépondérante. Dès lors, les traits déterminés par de très nombreux facteurs en biologie présentent très souvent une distribution gaussienne.

## 5 Variance

**Variance** : Pour une population, il s'agit de la moyenne des écarts aux carrés à la moyenne.

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

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

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

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

### Développement de Koenig : Démonstration

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

$$\begin{aligned} \sum_i^p n_i (x_i - \mu_x)^2 &= \sum_i^p n_i (x_i^2 - 2x_i \mu_x + \mu_x^2) \\ &= \sum_i^p n_i x_i^2 - 2\mu_x \sum_i^p n_i x_i + \sum_i^p n_i \mu_x^2 \end{aligned}$$

or on sait que  $\sum_i^p n_i = n$  et que  $\sum_i^p n_i x_i = n\mu_x$

$$\text{donc } \sum_i^p n_i (x_i - \mu_x)^2 = \sum_i^p n_i x_i^2 - n\mu_x^2$$

$$\text{donc } var(x) = \frac{1}{n} (\sum_i^p n_i x_i^2 - n\mu_x^2) = \left( \frac{1}{n} (\sum_i^p n_i x_i^2) \right) - \mu_x^2$$

### Estimateur de la variance pour un échantillon

Du fait de l'incertitude sur la moyenne de la population établie par rapport à un échantillon, la formule précédente tend à sous estimer la variance. Quand on estime la variance à partir d'un échantillon, il convient de lever ce biais, en utilisant la formule :

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

Le développement de Koenig donne :

$$\hat{\sigma}_x^2 = \left( \frac{1}{n-1} \sum_i^p x_i^2 \right) - \frac{n}{n-1} \hat{\mu}_x^2$$

### Démonstration

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

Soit la quantité analogue mais calculer sur un échantillon :  $S_x^2 = \frac{\sum_i^n (x_i - \hat{\mu}_x)^2}{n}$

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

$$\mathbb{E}[\sigma_x^2 - S_x^2] = \mathbb{E}\left[\frac{1}{n} \sum_i^p (x_i - \mu_x)^2 - \frac{1}{n} \sum_i^p (x_i - \hat{\mu}_x)^2\right]$$

$$\begin{aligned}
&= \mathbb{E}\left[\frac{1}{n} \sum_i^p ((x_i^2 - 2\mu_x x_i + \mu_x^2) - (x_i^2 - 2\hat{\mu}_x x_i + \hat{\mu}_x^2))\right] \\
&= \mathbb{E}\left[\frac{1}{n} \sum_i^p (\mu_x^2 - \hat{\mu}_x^2 + 2x_i(\hat{\mu}_x - \mu_x))\right] \\
&= \mathbb{E}\left[\mu_x^2 - \hat{\mu}_x^2 + \frac{1}{n} \sum_i^p (2x_i(\hat{\mu}_x - \mu_x))\right] \\
&= \mathbb{E}\left[\mu_x^2 - \hat{\mu}_x^2 + 2\hat{\mu}_x(\hat{\mu}_x - \mu_x)\right] \\
&= \mathbb{E}\left[\mu_x^2 + \hat{\mu}_x^2 - 2\hat{\mu}_x\mu_x\right] \\
&= \mathbb{E}\left[(\hat{\mu}_x - \mu_x)^2\right] \\
&= \text{Var}(\hat{\mu}_x)
\end{aligned}$$

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

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

Ainsi si on veut utiliser un estimateur non biaisé de la variance à partir d'un échantillon, on ne prend pas  $S_x^2$  mais  $\frac{n}{n-1}S_x^2$ . Cette correction est appelée **correction de Bessel**.

**Ecart type** : racine carrée de la variance (ou de son estimateur pour des échantillons).

## 6 Loi de probabilité

### 6.1 Définitions :

- **Loi de probabilité** : Loi (ou formulation mathématique) qui permet de décrire le comportement aléatoire d'un phénomène au hasard. Elle associe une probabilité avec les modalités d'une variable aléatoire ou bien elle associe une probabilité avec des classes connues de  $X$ .
- **Fonction de masse** : Pour des variables discrètes, la loi associe les valeurs de  $X$  avec leur probabilité attendue, on parle de fonction de masse  $f(x) = P(X = x)$ . Dans le cas des variables continues comme  $X$  peut prendre une infinité de valeurs, on utilise la densité.
- **Densité de Probabilité** : Une densité de probabilité est une fonction qui permet de représenter une loi de probabilité sous forme d'intégrales. Ceci est très utile dans le cas continu car s'il n'est pas possible d'associer une valeur unique de  $X$  avec sa probabilité (par définition infiniment petit, il est possible d'associer un intervalle de valeur de  $x$  avec la probabilité prise par cet intervalle).

En pratique, la densité de probabilité  $f$  permet de définir la probabilité de  $x$  sur un intervalle compris entre les valeurs  $X = a$  et  $X = b$  et on peut alors écrire :

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

Dans le cas discret ou continu, les lois de probabilité peuvent être caractérisées par une **fonction de répartition**.

La fonction de répartition  $F(x)$  est l'association de  $X$  avec sa fréquence cumulée.

$$F(x) = P(X \leq x)$$

C'est en quelque sorte une formulation mathématique du polygone des fréquences cumulées.

Dans le cas continu, la fonction de répartition  $F(x)$  est l'intégrale de la densité  $f(x)$  de  $-\infty$  à la valeur de  $X = x$ .

$$F(x) = \int_{-\infty}^x f(x) dx$$

## 6.2 Loi de probabilité uniforme

Cas discret :

Dans le cas d'une variable discrète, tous les événements ont la même probabilité. Par exemple pour un jeu de dé à un tirage  $P(X = 1) = P(X = 2), P(X = 3), P(X = 4), P(X = 5), P(X = 6) = 1/6$  tandis que la probabilité que  $X$  soit différente de ces 6 valeurs = 0.

Loi uniforme continue sur un intervalle borné  $[a; b]$  :

Dans le cas des variables continues, il y a une infinité d'issues possibles, pour établir la loi on utilise la fonction de répartition (= polygone des fréquences cumulées) ou bien alors les densités de probabilité. Dans ce cas :

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a; b] \\ 0 & \text{sinon} \end{cases}$$

L'espérance et la variance variable  $X$  suivant cette loi sur l'intervalle  $[a; b]$  sont données par :

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

## 6.3 Loi de Bernouilli

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

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

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

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

$$Var(X) = pq$$

Un tirage de  $X$  porte le nom d'épreuve ou d'évènement de Bernouilli.

## 6.4 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 :

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

$\binom{n}{k}$  correspond au nombre de combinaisons de  $k$  parmi  $n$  et porte également le nom de coefficient binomial (cf. cours de première ou terminale), il est parfois écrit  $C_n^k$ . A titre de rappel :

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

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

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

$$Var(X) = npq$$

Cette loi peut être utilisée pour modéliser le comportement de proportions, sachant qu'une proportion correspond à un nombre de succès sur  $n$  tirages.

## 6.5 Loi de Poisson

Loi discrète qui caractérise la probabilité d'un nombre  $k$  d'événements qui se produisent dans un temps ou un espace fixé et de paramètres  $\lambda$ ; ce paramètre étant le nombre moyen d'événements 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  $\lambda$ .

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

$$Var(X) = \lambda$$

Cette loi est très souvent utilisée pour modéliser des paramètres biologiques liés à un échantillonnage restreint dans le temps ou l'espace.

## 6.6 Loi Géométrique

Cette loi discrète établit la probabilité associée au nombre  $k$  tirages pour obtenir le premier succès de probabilité  $p$ .

La loi de masse est donnée par

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

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

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

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

Elle peut être utilisée pour estimer un effort d'échantillonnage suffisant qu'un événement se réalise.

## 6.7 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  $p$ .

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

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

Cette loi est souvent utilisée en épidémiologie.

## 6.8 Loi normale et loi normale centrée réduite.

**Loi Normale centrée réduite** : Loi continue symétrique caractérisée par la moyenne et la variance et permettant de comprendre le comportement d'une suite d'expériences aléatoires quand le nombre d'essais est grand (liée au théorème centrale limite). Quand une variable suit cette loi, sa médiane se rapproche de sa moyenne.

$$Var = \sigma^2$$

$$\mathbb{E} = \mu$$

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

Pour passer de  $\mathcal{N}(0, 1)$  à  $\mathcal{N}(\mu, \sigma)$ , on multiplie les quantiles par l'écart type et on ajoute à cette quantité la moyenne  $\mu$

Pour certaines lois, on utilise le **concept de degrés de liberté** car l'estimation d'un paramètre peut dépendre d'autres estimations ou de biais. Ce concept permet de corriger une estimation du fait de dépendance entre les estimateurs calculés (e.g. par exemple l'estimation de la variance dépend de l'estimation de la moyenne). De manière basique, le nombre de degré de liberté est égal au nombre d'observations moins le nombre de relations entre ces observations.

## 6.9 Loi de Student

Loi continue centrée et réduite qui permet d'apprécier l'erreur sur la moyenne à partir de petit échantillon. Elle est dérivé du théorème centrale limite ou on sait que la quantité  $\frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}}$  tend vers la loi normale centrée réduite quand  $n$  est grand. Quand  $n$  est petit, cette quantité  $T_n = \frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}}$  tend vers la variable aléatoire de Student  $t$  à  $n - 1$  degrés de liberté.

## 6.10 Autres lois

Il existe de très nombreuses lois de probabilité pour décrire des phénomènes aléatoires. Concernant les lois discrètes utiles en biologie on peut citer la loi hypergéométrique. Pour les lois à densité continues on retiendra la loi du Chi 2 ( $\chi^2$ ) et la loi de Fisher Snedecor qui permettent de modéliser la distribution de variances ou la distribution du ratio de variances. Vous pourrez consulter les pages wikipedia qui leur sont dédiées sachant que les tables de ces lois sont données dans le formulaire ; la page wikipedia sur la loi hypergéométrique est directement associée à une application en biologie, tandis que vous utiliserez les deux autres plus tard dans votre cursus ou pour calculer l'intervalle de confiance sur la variance. Ces pages sont donc à consulter.

## 7 Intervalles de confiance

**Intervalle de confiance pour un paramètre** : c'est la gamme de valeurs pour lesquelles qui permette d'encadrer le paramètre que l'on estime à  $1 - \alpha$  pourcents de chance de se tromper.  $\alpha$  correspond au risque que le paramètre que l'on estime soit en dehors de l'intervalle de confiance.

### 7.1 Intervalles de confiance sur la moyenne

D'après le théorème central limite, si  $X$  est une variable aléatoire d'espérance  $\mu$  et d'écart type  $\sigma$ , la moyenne estimée de cette variable à une espérance  $\mu$  et un écart type de  $\frac{\sigma}{\sqrt{n}}$ . La quantité  $\frac{\sigma}{\sqrt{n}}$  porte le nom d' **erreur type** sur la moyenne. En d'autres termes  $\mu$  est estimé par  $\hat{\mu}$  et la variance de  $\hat{\mu}$  attendue est de  $\frac{\sigma^2}{n}$ . On peut estimer  $\sigma^2$  par  $\hat{\sigma}^2 = \frac{n}{n-1} \times \sigma^2$ . La distribution de  $\hat{\mu}$  est symétrique (gaussienne). On peut estimer la demi longueur de l'intervalle de confiance à 95% par :

$a_\alpha = t_\alpha \times \frac{\sigma}{\sqrt{n}}$  avec  $t_\alpha$  le quantile de la loi de Student à  $n - 1$  degré de liberté à 97.5%. Pour les grands échantillons et comme le prédit le théorème centrale limite,  $t_\alpha$  peut être estimé par  $z_\alpha$  le quantile de la loi Normale centrée réduite à 97.5%.

Pour un intervalle à 99%, les quantiles seront alors pris à 99.5% (la loi normale est symétrique donc les valeurs attendues seront alors comprises entre l'erreur type et les quantiles extrêmes à  $t_1$  0.5% et  $t_2$  à 99.5% couvrant 99% de la variation attendue de  $\mu$ ).

L'intervalle de confiance est donc donné par :

$$\left[ \hat{\mu} + t_{1\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}}; \hat{\mu} + t_{2\alpha} \times \frac{\hat{\sigma}}{\sqrt{n}} \right] \quad \text{avec } t_{1\alpha} \quad \text{le quantile à } \alpha/2\% \quad \text{et } t_{2\alpha} \quad \text{le quantile à } 1 - \alpha/2\%$$

ou bien par

$$\left[ \hat{\mu} - t_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}}; \hat{\mu} + t_\alpha \times \frac{\hat{\sigma}}{\sqrt{n}} \right] \quad \text{avec } t_\alpha \quad \text{le quantile à } 1 - \alpha/2\%$$

## 7.2 Intervalle de confiance pour la variance

L'intervalle de confiance sur la variance se fonde sur le fait que la quantité  $(n-1)\frac{\hat{\sigma}^2}{\sigma^2}$  suit une loi de distribution du  $\chi^2$  à  $n-1$  degrés de liberté. On va donc chercher à encadrer  $\sigma^2$  à l'aide des quantiles à  $\frac{\alpha}{2} \times 100\%$  et  $1 - \frac{\alpha}{2} \times 100\%$  qu'on définira respectivement par  $\chi_1^2$  et  $\chi_2^2$  à  $n-1$  degrés de liberté. L'intervalle de confiance pour la variance se définit donc par :

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

## 7.3 Intervalles de confiance pour une proportion

Le théorème centrale limite nous dit 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

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

A partir de là on peut construire l'intervalle de confiance pour une proportion. En effet pour la loi de Bernouilli la variance attendu  $\sigma^2 = p \times q$  On voit en fait qu'on substitue dans la formule de l'intervalle de confiance sur la moyenne  $\sigma^2$  par le produit  $pq$ . Ainsi le demi intervalle de confiance était approximé par  $z_\alpha \times \frac{\sigma}{\sqrt{n}}$  quand  $n$  était grand, donc on aura pour demi intervalle  $z_\alpha \times \sqrt{\frac{pq}{n}} = z_\alpha \times \sqrt{\frac{p(1-p)}{n}}$ . L'intervalle se calcule alors par :

$$\left[ \hat{p} - z_\alpha \times \sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_\alpha \times \sqrt{\frac{p(1-p)}{n}} \right]$$

avec  $z_\alpha$  correspondant au quantile à  $1 - \frac{\alpha}{2} \times 100$  pourcents de la distribution normale centrée réduite.

## 8 Taille nécessaire d'un échantillon pour une précision donnée

### 8.1 Cas d'une proportion

On souhaite par exemple que le demi écart  $a_\alpha$  de confiance pour une proportion soit inférieur à une valeur  $\lambda$ .

Sachant que pour une proportion le demi intervalle de confiance vaut :  $z_\alpha \sqrt{\frac{p(1-p)}{n}}$ .

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

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

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

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

### 8.2 Cas de la moyenne

Dans le cas d'un intervalle de confiance sur une moyenne le raisonnement est analogue :

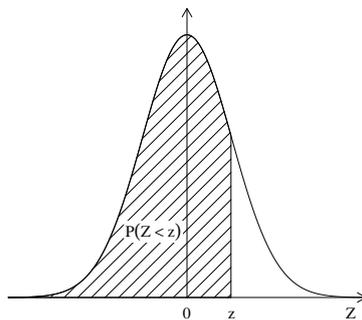
On veut  $a_\alpha < \lambda$  donc

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

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

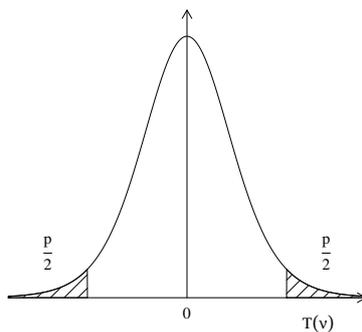
avec  $t_\alpha$ , le quantile de la loi de Student correspondant à la probabilité associée à un intervalle de confiance à  $1 - \alpha \times 100$  pourcents. Comme ce quantile varie légèrement, on peut l'approximer par le quantile de la loi Normale centrée réduite si on trouve que  $n$  est grand ( $\geq 30$ ). Si  $n$  est petit alors on optera pour son approximation par la loi de Student et on procédera par tâtonnement.

## 9 Fonction de répartition de la loi normale centrée réduite



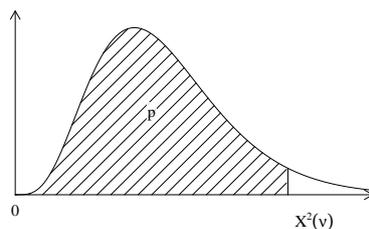
$z$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0.1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0.2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0.3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0.4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0.5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0.6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0.7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0.8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0.9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1.1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1.2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1.3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1.4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1.5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1.6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1.7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1.8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1.9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2.1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2.2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2.3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2.4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2.5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2.6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2.7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2.8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2.9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

# 10 Loi de Student



$\nu$	$P$												
	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0,1584	0,3249	0,5095	0,7265	1	1,3764	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6192
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,92	4,3027	6,9646	9,9248	31,5991
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,924
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,941	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,015	2,5706	3,3649	4,0321	6,8688
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,896	1,1192	1,4149	1,8946	2,3646	2,998	3,4995	5,4079
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,306	2,8965	3,3554	5,0413
9	0,1293	0,261	0,3979	0,5435	0,7027	0,8834	1,0997	1,383	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,201	2,7181	3,1058	4,437
12	0,1283	0,259	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	2,681	3,0545	4,3178
13	0,1281	0,2586	0,394	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,128	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,345	1,7613	2,1448	2,6245	2,9768	4,1405
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1314	2,6025	2,9467	4,0728
16	0,1277	0,2576	0,3923	0,535	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,5835	2,9208	4,015
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,069	1,3334	1,7396	2,1098	2,5669	2,8982	3,9651
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,862	1,0672	1,3304	1,7341	2,1009	2,5524	2,8784	3,9216
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,861	1,0655	1,3277	1,7291	2,093	2,5395	2,8609	3,8834
20	0,1273	0,2567	0,3909	0,5329	0,687	0,86	1,064	1,3253	1,7247	2,086	2,528	2,8453	3,8495
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,5176	2,8314	3,8193
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,5083	2,8188	3,7921
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,4999	2,8073	3,7676
24	0,127	0,2562	0,39	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,4922	2,7969	3,7454
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,4851	2,7874	3,7251
26	0,1269	0,256	0,3896	0,5309	0,684	0,8557	1,0575	1,315	1,7056	2,0555	2,4786	2,7787	3,7066
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,4727	2,7707	3,6896
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,056	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739
29	0,1268	0,2557	0,3892	0,5302	0,683	0,8542	1,0553	1,3114	1,6991	2,0452	2,462	2,7564	3,6594
30	0,1267	0,2556	0,389	0,53	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,4573	2,75	3,646
40	0,1265	0,255	0,3881	0,5286	0,6807	0,8507	1,05	1,3031	1,6839	2,0211	2,4233	2,7045	3,551
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,3739	2,6387	3,4163
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6577	1,9799	2,3578	2,6174	3,3735
$\infty$	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,96	2,3264	2,5758	3,2905

# 11 Loi du $\chi^2$



$\nu$	$p$												
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	7,879	6,635	5,024	3,841	2,706	1,323	0,4549	0,1015	0,01579	0,003932	0,0009821	0,0001571	3,927e-05
2	10,6	9,21	7,378	5,991	4,605	2,773	1,386	0,5754	0,2107	0,1026	0,05064	0,0201	0,01003
3	12,84	11,34	9,348	7,815	6,251	4,108	2,366	1,213	0,5844	0,3518	0,2158	0,1148	0,07172
4	14,86	13,28	11,14	9,488	7,779	5,385	3,357	1,923	1,064	0,7107	0,4844	0,2971	0,207
5	16,75	15,09	12,83	11,07	9,236	6,626	4,351	2,675	1,61	1,145	0,8312	0,5543	0,4117
6	18,55	16,81	14,45	12,59	10,64	7,841	5,348	3,455	2,204	1,635	1,237	0,8721	0,6757
7	20,28	18,48	16,01	14,07	12,02	9,037	6,346	4,255	2,833	2,167	1,69	1,239	0,9893
8	21,95	20,09	17,53	15,51	13,36	10,22	7,344	5,071	3,49	2,733	2,18	1,646	1,344
9	23,59	21,67	19,02	16,92	14,68	11,39	8,343	5,899	4,168	3,325	2,7	2,088	1,735
10	25,19	23,21	20,48	18,31	15,99	12,55	9,342	6,737	4,865	3,94	3,247	2,558	2,156
11	26,76	24,72	21,92	19,68	17,28	13,7	10,34	7,584	5,578	4,575	3,816	3,053	2,603
12	28,3	26,22	23,34	21,03	18,55	14,85	11,34	8,438	6,304	5,226	4,404	3,571	3,074
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34	9,299	7,042	5,892	5,009	4,107	3,565
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34	10,17	7,79	6,571	5,629	4,66	4,075
15	32,8	30,58	27,49	25	22,31	18,25	14,34	11,04	8,547	7,261	6,262	5,229	4,601
16	34,27	32	28,85	26,3	23,54	19,37	15,34	11,91	9,312	7,962	6,908	5,812	5,142
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34	12,79	10,09	8,672	7,564	6,408	5,697
18	37,16	34,81	31,53	28,87	25,99	21,6	17,34	13,68	10,86	9,39	8,231	7,015	6,265
19	38,58	36,19	32,85	30,14	27,2	22,72	18,34	14,56	11,65	10,12	8,907	7,633	6,844
20	40	37,57	34,17	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,591	8,26	7,434
21	41,4	38,93	35,48	32,67	29,62	24,93	20,34	16,34	13,24	11,59	10,28	8,897	8,034
22	42,8	40,29	36,78	33,92	30,81	26,04	21,34	17,24	14,04	12,34	10,98	9,542	8,643
23	44,18	41,64	38,08	35,17	32,01	27,14	22,34	18,14	14,85	13,09	11,69	10,2	9,26
24	45,56	42,98	39,36	36,42	33,2	28,24	23,34	19,04	15,66	13,85	12,4	10,86	9,886
25	46,93	44,31	40,65	37,65	34,38	29,34	24,34	19,94	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	35,56	30,43	25,34	20,84	17,29	15,38	13,84	12,2	11,16
27	49,64	46,96	43,19	40,11	36,74	31,53	26,34	21,75	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	37,92	32,62	27,34	22,66	18,94	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	39,09	33,71	28,34	23,57	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43,77	40,26	34,8	29,34	24,48	20,6	18,49	16,79	14,95	13,79
40	66,77	63,69	59,34	55,76	51,81	45,62	39,34	33,66	29,05	26,51	24,43	22,16	20,71
50	79,49	76,15	71,42	67,5	63,17	56,33	49,33	42,94	37,69	34,76	32,36	29,71	27,99
60	91,95	88,38	83,3	79,08	74,4	66,98	59,33	52,29	46,46	43,19	40,48	37,48	35,53
70	104,2	100,4	95,02	90,53	85,53	77,58	69,33	61,7	55,33	51,74	48,76	45,44	43,28
80	116,3	112,3	106,6	101,9	96,58	88,13	79,33	71,14	64,28	60,39	57,15	53,54	51,17
90	128,3	124,1	118,1	113,1	107,6	98,65	89,33	80,62	73,29	69,13	65,65	61,75	59,2
100	140,2	135,8	129,6	124,3	118,5	109,1	99,33	90,13	82,36	77,93	74,22	70,06	67,33