Polytech Nancy **2020 / 2021**
3e année

Statistiques

Chapitre 2 – Estimation

Marie-Christine SUHNER

marie-christine.suhner@univ-lorraine.fr

2.1 Echantillonnage

- **Echantillonnage d'une population**
  - Objectif : découvrir des renseignements au sujet d'une population particulière
  - Moyen : prélever un échantillon représentatif de la population
- **Echantillonnage aléatoire**
  - Soit une population de N individus sur laquelle on prélève un échantillon de taille n
  - On prélève, par tirage au sort, n numéros constituant l'échantillon
  - Echantillon aléatoire simple : chaque sous-ensemble de n individus parmi N a la même probabilité d'être choisi
- **Principe de construction d'un échantillon**
  - **Tirage sans remise (ou exhaustif)**
    Les unités prélevées ne sont pas remises dans la population
  - **Tirage avec remise (indépendant ou non exhaustif)**
    Chaque unité prélevée au hasard est observée puis remise dans la population
  - **Remarque**
    Lorsque n est petit devant N, les résultats obtenus par les deux types de tirages se confondent
- **Méthodes de construction d'un échantillon**
  - **Utilisation d'une table de nombres aléatoires**
  - **Utilisation d'un générateur de nombres pseudo-aléatoires**
    Obtention par un procédé informatique de nombres tirés au hasard, de façon équiprobable et indépendante (en fait, ces nombres sont générés par une fonction, mais tellement complexe que tout se passe comme s'ils étaient indépendants les uns des autres)

2.1 Théorèmes fondamentaux

Les résultats asymptotiques en probabilité concernent les propriétés des suites $X_1, X_2, X_3, \ldots, X_n$ de variables aléatoires si n tend vers l'infini

❖ Loi faible des grands nombres

Soient $X_1, X_2, X_3, \ldots, X_n$ des variables aléatoires indépendantes et supposons que $E(X_i) = \mu$ et $V(X_i) = \sigma^2$ existent, alors

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[p]{} \mu \qquad n \to \infty$$

La moyenne $\frac{X_1 + \cdots + X_n}{n}$ converge en probabilité vers l'espérance μ

❖ Loi forte des grands nombres

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes de même loi admettant une espérance notée μ, alors

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[p.s.]{} \mu \qquad n \to \infty$$

❖ Théorème central limite

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes ayant respectivement comme moyenne $E(X_i) = \mu_i$ et comme variance $V(X_i) = \sigma_i^2$

$$Y = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

$$E(Y) = \sum_{i=1}^{n} \mu_i \quad \text{et} \quad V(Y) = \sum_{i=1}^{n} \sigma_i^2$$

Alors $$U = \frac{Y - E(Y)}{\sigma_Y} \rightarrow \text{LNCR} \qquad n \to \infty$$

Cas particulier :

Si les variables possèdent la même loi, alors,

$$U = \frac{Y - n\mu}{\sigma \sqrt{n}} \rightarrow \text{LNCR} \qquad n \to \infty$$

2.1 Application des théorèmes fondamentaux

❖ Loi suivie par la moyenne d'un échantillon $\overline{X}$

On prélève au hasard un échantillon de taille n (tirage avec remise) dont les éléments possèdent un caractère mesurable X de moyenne μ et d'écart type σ

On crée une suite de n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ dont chacune a la même distribution X

$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$ tend vers une loi normale de paramètres

$$E(\overline{X}) = \frac{1}{n} E(X_1 + \ldots + X_n) = \mu$$

$$V(\overline{X}) = \frac{1}{n^2} V(X_1 + \ldots + X_n) = \frac{\sigma^2}{n}$$

et ce, d'autant plus que la taille de l'échantillon est grande

❖ Approximation d'une loi binomiale par une loi normale

Si X est est le nombre de succès en n épreuves de Bernoulli, alors, si n est suffisamment grand

$$U = \frac{X - n\,p}{\sqrt{n\,p\,(1-p)}} \rightarrow \text{LNCR}$$

Ne pas oublier de tenir compte de la correction de continuité

❖ Règles pratiques pour définir n

Distribution symétrique et unimodale

n de l'ordre de 4 ou 5

Distribution symétrique sans mode dominant

$n \geq 12$ (on apparente cette distribution à une loi uniforme)

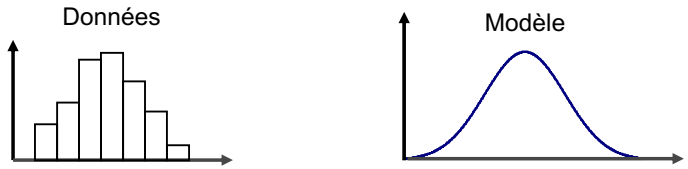
Distribution quelconque

$n \geq 30$

2.1 Principe de l'estimation

❖ Modèle statistique

Observation x_i = réalisation d'une variable aléatoire X avec une distribution partiellement inconnue



Exemple 1

On lance une pièce de monnaie 6 fois et on obtient les résultats (Pile, Pile, Face, Pile, Face, Pile) = (1, 1, 0, 1, 0, 1)

Chaque observation x_i est une réalisation d'une variable aléatoire X_i (variable de Bernoulli)

Modèle $X = (X_1, \ldots, X_6)$

Les jets X_i sont indépendants et suivent tous une loi de Bernoulli Ber (1 ; p) avec p inconnu entre 0 et 1

Exemple 2

On effectue plusieurs mesures d'une quantité physique μ

De telles mesures ont une composante aléatoire (due aux erreurs de mesure, par exemple) :

mesure = valeur vraie théorique + erreur de mesure

$$X = \mu + \varepsilon$$

$$E(\varepsilon) = 0 \text{ et } V(\varepsilon) = \sigma^2$$

Modèle $X \rightarrow N(\mu ; \sigma)$ avec μ et σ inconnus

2.1 Principe de l'estimation

❖ Estimation de paramètres inconnus

Une fois le modèle choisi, il s'agit d'estimer le (ou les) paramètre(s) inconnu(s), à l'aide des réalisations (de la variable aléatoire) observées

Exemple 1

Séquence observée (1, 1, 0, 1, 0, 1)

Un estimateur naturel de p (probabilité d'un Pile) est :

$$\hat{P}(\text{données}) = \hat{P} = \frac{\text{nombre de Pile}}{\text{nombre d'essais}}$$

$$\hat{p} = 4/6 = 0{,}667$$

Un estimateur est une fonction des données

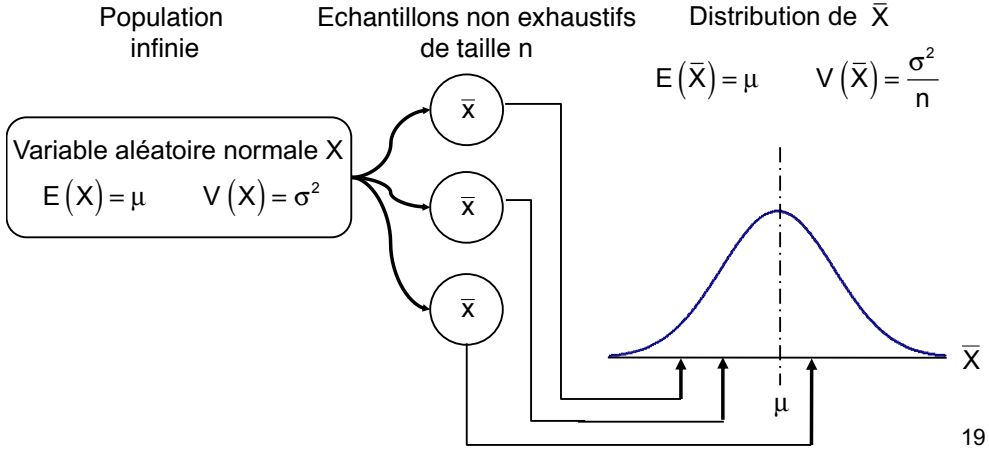
L'estimateur est construit de telle façon que sa valeur soit proche de la vraie valeur du paramètre de la distribution sous-jacente

Il faut distinguer la valeur vraie p et la valeur estimée $\hat{p}$

❖ Distribution d'échantillonnage

Un estimateur est une variable aléatoire dont on peut déterminer la distribution (ou loi) de probabilité, l'espérance et la variance

Distribution d'échantillonnage de la moyenne de mesures



2.2 Estimateur ponctuel

❖ Définitions

- Estimation ponctuelle : nombre qui estime une caractéristique d'une population, à partir des résultats d'un échantillon
- L'**estimation ponctuelle** est calculée à partir d'un **estimateur** qui est une variable aléatoire dépendant des observations d'un échantillon

❖ Propriétés des estimateurs ponctuels

– Biais d'un estimateur

Paramètre inconnu : θ

Estimateur : $\hat{\theta}(\text{données}) = \hat{\theta}$

La qualité de cet estimateur dépend de la différence

$$\hat{\theta} - \theta$$

L'espérance de cette quantité est appelée le biais

$$\text{Biais}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

– Carré moyen de l'erreur (CME) ou erreur quadratique

Le carré de l'erreur d'un estimateur est défini comme

$$(\hat{\theta} - \theta)^2$$

L'espérance de ce carré est le CME

$$\text{CME}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

$$\text{CME}(\hat{\theta}) = V(\hat{\theta}) + \left[\text{Biais}(\hat{\theta})\right]^2$$

Certains paramètres peuvent avoir plusieurs estimateurs sans biais

Le choix s'effectue alors en comparant les variances des estimateurs

2.2 Estimateur ponctuel

❖ Propriétés des estimateurs ponctuels (suite)

– Inégalité de Cramer Rao

On considère une variable aléatoire X qui suit une distribution de probabilité définie par $f(x;\theta)$ où θ est un paramètre inconnu

On veut estimer θ au moyen d'un échantillon de taille n

Les observations $x_1, x_2, \ldots, x_n$ sont considérées comme les valeurs prises par n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ issues de la même distribution de probabilité que X :

$$f(x_1;\theta),\ f(x_2;\theta), \ldots\ f(x_n;\theta)$$

On définit la fonction de vraisemblance L

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_i f(x_i;\theta)$$

La variance d'un estimateur sans biais possède une borne inférieure telle que :

$$V(\hat{\theta}) \geq \frac{1}{I_n(\theta)} \quad \text{où} \quad I_n(\theta) = E\left\{\left[\frac{d}{d\theta} \ln L(x_1, x_2, \ldots, x_n; \theta)\right]^2\right\}$$

$$= n\, I_1(\theta) = n\,.\,E\left\{\left[\frac{d}{d\theta} \ln f(x;\theta)\right]^2\right\}$$

si le domaine de définition de X ne dépend pas de θ

– Estimateur efficace

Un estimateur sans biais est efficace si sa variance est la plus faible parmi les autres variances des estimateurs sans biais

– Estimateur convergent

Un estimateur est convergent si sa distribution tend à se concentrer autour de la valeur inconnue à estimer, à mesure que la taille de l'échantillon augmente

$$V(\hat{\theta}) \to 0 \quad \text{quand} \quad n \to \infty$$

– Estimateur absolument correct

Un estimateur sans biais et convergent est dit absolument correct

2.2 Estimateur ponctuel

❖ Construction d'un estimateur ponctuel

Objectif : obtenir une estimation de paramètres inconnus à l'aide de données

Plusieurs méthodes possibles :

méthode des moments (principe : identifier les moments théoriques avec les moments empiriques),

méthode des moindres carrés, du maximum de vraisemblance…

❖ Méthode du maximum de vraisemblance

– Principe

On considère une variable aléatoire X qui suit une distribution de probabilité définie par $f(x;\theta)$ où θ est un paramètre inconnu

On veut estimer θ au moyen d'un échantillon de taille n

Les observations $x_1, x_2, \ldots, x_n$ sont considérées comme les valeurs prises par n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ issues de la même distribution de probabilité que X :

$$f(x_1;\theta),\ f(x_2;\theta),\ldots\ f(x_n;\theta)$$

Fonction de vraisemblance L

$$L(x_1, x_2, \ldots, x_n;\theta) = \prod_i f(x_i;\theta)$$

On recherche une valeur de θ qui maximise L

$$\frac{dL}{d\theta} = 0 \quad \text{et} \quad \frac{d^2L}{d\theta^2} < 0$$

En pratique, on travaille plutôt sur ln L

$$\frac{d\ln L}{d\theta} = 0 \quad \text{et} \quad \frac{d^2\ln L}{d\theta^2} < 0$$

– Construction d'un estimateur d'une proportion

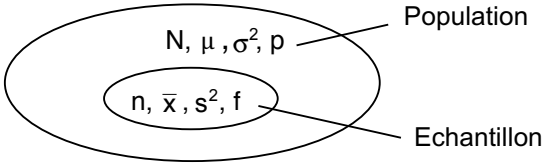
Soit $x_1, x_2, \ldots, x_n$ un échantillon issu d'une loi de Bernoulli Ber (1 ; p)

$$\Pr(X = x_i) = p^{x_i}(1-p)^{1-x_i} \quad \text{avec} \quad x_i = 0 \text{ ou } 1$$

$$L(x_1, x_2, \ldots, x_n; p) = p^{x_1+\ldots+x_n}(1-p)^{n-x_1-\ldots-x_n}$$

$\hat{P} = \dfrac{X_1+\ldots+X_n}{n}$ est l'estimateur du maximum de vraisemblance du paramètre p

2.2 Estimation ponctuelle moyenne, variance, proportion



❖ Estimation ponctuelle de la moyenne

- Estimateur **sans biais** de la moyenne

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Estimation ponctuelle

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance de l'estimateur sans biais de la moyenne *(variance population connue)*

Echantillon non exhaustif

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

Echantillon exhaustif

Dans le cas d'échantillonnage sans remise, à partir d'une population finie de taille N

$$V(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \cong \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$$

n / N représente le taux de sondage

En pratique, on peut négliger le taux de sondage s'il est inférieur à 5 %

2.2 Estimation ponctuelle moyenne, variance, proportion

❖ Estimation ponctuelle de la variance

- Estimateur **sans biais** de la variance *(moyenne population connue)*

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

- Estimation ponctuelle de la variance *(moyenne population connue)*

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

- Estimateur **sans biais** de la variance *(moyenne population inconnue)*

$$\hat{\sigma}^2 = S^{*2} = \frac{n}{n-1}S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Justification

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2$$

$$E\left(S^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}E\left(X_i^2\right) - E\left(\bar{X}^2\right) = E\left(X^2\right) - E\left(\bar{X}^2\right)$$

$$E\left(S^2\right) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2\left(\frac{n-1}{n}\right)$$

- Estimation ponctuelle de la variance *(moyenne population inconnue)*

$$\hat{\sigma}^2 = s^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

❖ Estimation ponctuelle d'une proportion

- Estimateur **sans biais** de la proportion

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

- Estimation ponctuelle

$$\hat{p} = \frac{x_1 + \ldots + x_n}{n} = f$$

2.3 Estimation par intervalle

Les estimations ponctuelles ne fournissent aucune information sur la précision des estimations

Elles ne tiennent pas compte de l'erreur attribuable aux fluctuations d'échantillonnage

❖ Définition

Un intervalle de confiance [I ; S] pour un paramètre θ ayant un niveau de confiance $(1-\alpha)$ vérifie :

$$\Pr\left\{ I \le \theta \le S \right\} = 1 - \alpha$$

Les statistiques I et S sont dites bornes de confiance inférieure et supérieure

Ces bornes dépendent uniquement des données

❖ Interprétation

Cet intervalle signifie que, si l'on répète l'expérience un grand nombre de fois (prélever plusieurs fois un échantillon de taille n de la même population), dans 100 $(1-\alpha)$ cas sur 100, l'intervalle recouvre la vraie valeur du paramètre

- L'intervalle ainsi défini est un intervalle aléatoire puisqu'avant expérience, les limites de l'intervalle sont des variables aléatoires, fonction des observations de l'échantillon
- L'élément aléatoire réside dans les bornes de confiance et non dans le paramètre. Ce sont I et S qui changent lorsque l'on répète l'expérience et non pas θ
- On ne peut pas interpréter le niveau de confiance comme une probabilité. Etre confiant à 95 % que $\theta \in [I ; S]$ ne veut pas dire que le paramètre θ est devenu une quantité aléatoire qui est avec une probabilité 0,95 dans l'intervalle [I ; S]
- La quantité α correspond au risque qu'a l'intervalle de ne pas contenir la vraie valeur du paramètre
  - θ est ou n'est pas dans l'intervalle [I ; S]
  - Si on affirme que $\theta \in [I ; S]$, on ne se trompera en moyenne que 100 α sur 100

2.3 Estimation par intervalle
Cas de la moyenne

❖ Construction de l'intervalle

- Il s'agit de calculer à partir de la moyenne de l'échantillon, un intervalle dans lequel il est vraisemblable que la valeur vraie μ se trouve
- On obtient cet intervalle en calculant deux limites auxquelles est associée une certaine assurance de contenir la vraie valeur de μ

$$\Pr\left\{ \bar{X} - k \le \mu \le \bar{X} + k \right\} = 1 - \alpha$$

- Les limites prendront, après avoir prélevé l'échantillon et calculé l'estimation de la moyenne la forme suivante

$$\bar{x} - k \le \mu \le \bar{x} + k$$

- k est déterminé à l'aide de l'écart type de la distribution d'échantillonnage de $\bar{X}$ et du niveau de confiance $(1-\alpha)$ choisi

D'après le théorème central limite, si l'on prélève un échantillon aléatoire non exhaustif de taille n d'une population normale de variance connue,

$\bar{X}$ $\rightarrow$ Loi normale avec $E(\bar{X}) = \mu$ $V(\bar{X}) = \sigma^2 / n$

donc $$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \rightarrow \text{LNCR}$$

et $$\Pr\left\{ -u_{1-\alpha/2} \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le u_{1-\alpha/2} \right\} = 1 - \alpha$$

$$\Pr\left\{ \bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

d'où $$k = u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

2.3 Estimation par intervalle
Cas de la moyenne

❖ Interprétation de l'intervalle de confiance

- Avant toute expérience, la probabilité que l'intervalle aléatoire

$$\left[\bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} ; \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

contienne la valeur vraie μ est $(1-\alpha)$

- Ces deux limites sont des variables aléatoires qui prendront des valeurs numériques une fois que l'échantillon a été choisi et que l'on a obtenu la valeur $\bar{x}$
- On en déduit un intervalle d'extrémités fixes

$$\bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

et on lui attribue un niveau de confiance de 100 $(1-\alpha)$ % de contenir la valeur vraie de μ

❖ Remarques

- L'intervalle de confiance sera numériquement différent à chaque prélèvement d'échantillon de taille n puisqu'il est centré sur la moyenne de l'échantillon
- Le niveau de confiance est associé à l'intervalle et non au paramètre μ
- Plus le niveau de confiance est élevé, plus l'amplitude de l'intervalle est grande
- Pour la même taille d'échantillon, on perd de la précision en gagnant une plus grande confiance

2.4 Estimation ponctuelle et par intervalle d'une moyenne

- **Estimation ponctuelle**

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Estimation par intervalle**
  - **Population normale de variance population connue**

$$\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  - **Population quelconque de variance population connue avec un grand échantillon (n ≥ 30)**

$$\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  - **Population normale de variance population inconnue mais estimée avec un petit échantillon (n < 30)**

$$\bar{x} - t_{1-\alpha/2;\,\nu=n-1}\frac{s^*}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2;\,\nu=n-1}\frac{s^*}{\sqrt{n}} \quad \text{avec} \quad s^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

    Dans ce cas, le théorème central limite ne s'applique pas et

$$\frac{\bar{X} - \mu}{s^*/\sqrt{n}}$$

    n'est plus distribué selon la LNCR mais selon la distribution de Student à $\nu = n-1$ degrés de liberté (DL)

  - **Population quelconque de variance population inconnue mais estimée avec un grand échantillon (n ≥ 30)**

$$\bar{x} - t_{1-\alpha/2;\,\nu=n-1}\frac{s^*}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2;\,\nu=n-1}\frac{s^*}{\sqrt{n}} \quad \text{avec} \quad s^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  - **Remarque**

    Lorsque $n \ge 30$, $t_{1-\alpha/2;\,\nu=n-1}$ peut être approximé par $u_{1-\alpha/2}$

2.4 La distribution de Student et la notion de degrés de liberté

- **La distribution de Student (ou Student Fisher)**

$$T = \frac{\bar{X} - \mu}{s^*/\sqrt{n}} \quad \rightarrow \quad \text{Loi de Student à } \nu = n-1 \text{ DL}$$

$$E(T) = 0 \quad \text{si} \quad \nu > 1 \qquad \text{et} \qquad V(T) = \frac{\nu}{\nu - 2} \quad \text{si} \quad \nu > 2$$

- **Propriétés**
  - La variable T varie de - ∞ à + ∞
  - La distribution est symétrique par rapport à l'origine et un peu plus aplatie que la LNCR
  - La distribution ne dépend que d'une seule quantité ν appelée nombre de degrés de liberté
  - La variance de T tend vers 1 lorsque ν augmente
  - A mesure que ν augmente, la distribution de Student s'approche de plus en plus de la LNCR

- **Notion de degrés de liberté**
  - **Définition**

    Le nombre de degrés de liberté est une quantité associée à une somme de carrés et représente le nombre d'écarts indépendants dans le calcul de cette somme de carrés c'est-à-dire le nombre d'écarts nécessaire au calcul de la somme de carrés moins le nombre de restrictions de ces écarts

    Pour le calcul de $S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ il faut calculer n écarts mais

    il faut respecter la propriété $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ donc on perd 1 DL
  - **Autre définition**

    Nombre d'écarts nécessaire au calcul de la somme des carrés moins le nombre de paramètres que l'on doit estimer pour effectuer le calcul des écarts
  - **Remarque**

    Une somme de carrés divisée par les degrés de liberté constitue une variance

2.5 Estimation ponctuelle et par intervalle d'une variance

Moyenne population connue

- **Estimation ponctuelle**

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

- **Estimation par intervalle**

X suit une loi normale de moyenne connue

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \quad \rightarrow \quad \text{Loi du Khi-Deux à} \quad \nu = n \quad \text{DL}$$

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\alpha/2\,;\,\nu=n}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\alpha/2\,;\,\nu=n}}$$

Moyenne population inconnue

- **Estimation ponctuelle**

$$\hat{\sigma}^2 = s^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Estimation par intervalle**

X suit une loi normale de moyenne inconnue

$$\frac{(n-1)S^{*2}}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \quad \rightarrow \quad \text{Loi du Khi-Deux à} \quad \nu = n-1 \quad \text{DL}$$

$$\frac{(n-1)s^{*2}}{\chi^2_{1-\alpha/2\,;\,\nu=n-1}} \le \sigma^2 \le \frac{(n-1)s^{*2}}{\chi^2_{\alpha/2\,;\,\nu=n-1}}$$

2.5 La loi du Khi-Deux

- **La loi du Khi-Deux (ou de Pearson)**

Si U_1, U_2, …, U_ν sont des variables aléatoires normales centrées réduites et indépendantes, alors

$$U_1^2 + U_2^2 + \ldots + U_\nu^2 = \sum_{i=1}^{\nu} U_i^2 \quad \rightarrow \quad \text{Loi du Khi-Deux à } \nu \text{ DL}$$

$$E(\chi^2) = \nu \qquad \text{et} \qquad V(\chi^2) = 2\nu$$

- **Propriétés**
  - La quantité χ^2 est une variable aléatoire continue dont la loi de probabilité présente une asymétrie positive
  - Elle ne dépend que du nombre de degrés de liberté
  - A mesure que ν augmente, elle tend vers la loi normale $N\left(\nu\,;\sqrt{2\nu}\right)$
  - Elle possède la propriété d'additivité

- **Remarque**

Si X_1, X_2, …, X_n sont des variables aléatoires normales ayant respectivement pour moyennes $\mu_1, \mu_2, \cdots, \mu_n$ et pour variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$, alors

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} U_i^2 \quad \rightarrow \quad \text{Loi du Khi-Deux à} \quad n \quad \text{DL}$$

2.6 Estimation ponctuelle et par intervalle d'une proportion

❖ Estimation ponctuelle

$$\hat{p} = f = \frac{\sum_{i=1}^{n} x_i}{n}$$

❖ Estimation par intervalle – Approximation de la loi binomiale par la loi normale

- On prélève au hasard un échantillon de grande taille d'une population dont les éléments possèdent un caractère qualitatif dans une proportion p (inconnue)
- On observe sur l'échantillon une proportion f d'éléments qui possèdent le caractère
- Si $n.f \geq 5$ et $n.(1 - f) \geq 5$, alors :

$\hat{P} \rightarrow$ Loi normale avec $E(\hat{P}) = p$ $\quad V(\hat{P}) = \frac{p(1-p)}{n}$

donc $$U = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow \text{LNCR}$$

- En pratique, on utilise l'estimation f pour calculer l'écart type
- L'intervalle de confiance est

$$f - u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}$$

2.7 Marge d'erreur et taille d'échantillon requise

❖ Marge d'erreur (ou précision)

– Cas de la moyenne (variance population connue)

- Marge d'erreur $k = |\bar{x} - \mu|$

  Pour un niveau de confiance $(1-\alpha)$ $\quad |\bar{x} - \mu| \leq u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

- La marge d'erreur quantifie l'erreur attribuable aux fluctuations d'échantillonnage
- Pour σ connu et n fixé, plus le niveau de confiance est élevé, plus la marge d'erreur sera grande

– Cas de la proportion

Pour un niveau de confiance $(1-\alpha)$ $\quad |f - p| \leq u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}$

❖ Taille d'échantillon requise

– Cas de la moyenne (variance population connue)

- Taille minimale de l'échantillon requise pour une marge d'erreur fixée

$$n = \left[\frac{u_{1-\alpha/2}\,\sigma}{k}\right]^2$$

- Pour le même niveau de confiance et le même écart type, plus la marge d'erreur est faible, plus la taille d'échantillon sera élevée

– Cas de la proportion

- Si on connaît une valeur approximative de p obtenue, par exemple, par un sondage préalable sur un autre échantillon,

$$n = \left[\frac{u_{1-\alpha/2}}{k}\right]^2 p(1-p)$$

- Sinon, on fixe la valeur de p à 0,5. Cette valeur représente le cas le plus défavorable (plus grand écart type)