Statistique, L3 maths-éco

L3 MIASHS

Pierre Ailliot

Année 2024-2025

1 Introduction

Selon Wikipedia, "La statistique est la discipline qui étudie des phénomènes à travers la collecte de **données**, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous."

On peut donc distinguer plusieurs étapes dans une analyse statistique:

1. **Collecte des données.** Cette étape est généralement réalisée par des praticiens même si des méthodes statistiques peuvent être utilisées pour 'optimiser' le recueil des données. Cette étape sera peu abordée dans le cadre de ce cours, on supposera que les données ont déjà été collectées.
2. **Phase exploratoire d'analyse des données.** L'objectif de cette étape est de synthétiser l'information contenue dans les données afin de mettre en évidence certaines propriétés et de suggérer des hypothèses. Cela peut être réalisé en calculant des valeurs numériques (moyennes, écart types, proportions, ...), ou sous la forme de graphiques (histogrammes, camenberts...) et de tableaux. On peut également utiliser des méthodes plus sophistiquées comme l'analyse en composantes principales vue dans le cours d'analyse de données. Cette étape sera également peu abordée dans le cadre de ce cours.
3. **Modélisation.** Cette étape sera l'objet principal de cours. Elle consiste à développer un modèle mathématique adapté aux données avec généralement pour objectif de valider ou d'infirmer les hypothèses faites dans la phase exploratoire, de résumer l'information contenue dans les données ou de faire des prédictions.
4. **Interprétation et présentation des résultats (ou du modèle).** Les résultats de la modélisation mathématique peuvent être difficiles à comprendre pour les non-spécialistes, par exemple pour des décideurs qui veulent prendre en compte les résultats obtenus lors d'une analyse statistique. Il faut alors réfléchir à la manière d'expliquer les résultats obtenus, généralement sur la forme de graphiques ou de résumés numériques. Ce point sera partiellement abordé dans le cadre de cours (intervalles de confiance, tests d'hypothèses).

Les trois exemples ci-dessous serviront à illustrer ce cours :

- **Sondage**. Afin d'estimer les intentions de vote lors du deuxième tour d'une élection présidentielle, un institut réalise un sondage. Sur 1000 personnes interrogées au hasard parmi la population française, 520 pensent voter pour le candidat A et 480 pour le candidat B. Que peut-on en déduire sur les intentions de vote dans la population française? Avec quelle précision le sondage effectué permet t'il d'estimer le pourcentage d'intention de vote en faveur du candidat A? Peut-on déduire de ce sondage, avec une certaine confiance, que à la date du sondage le candidat A est en tête?
- **Risque innondation : nombre et montant des innondations couvertes par le régime Cat Nat.** En France, les risques liés aux inondations sont couverts par le régime d'indemnisation des catastrophes naturelles ('régime Cat Nat') créé par la loi du 13 juillet 1982. Dans le cadre de cours, on considérera les deux jeux de données suivants (source https://geoportail.ccr.fr/portal/apps/sites/#/bilancatnat).
  - Nombre de communes reconnues Cat Nat par an au titre des inondations.

– Coût des sinistres Cat Nat par an au titre des inondations (en millions d'euros, montants actualisés en euros 2022).

Les deux jeux de données sont disponibles sur la période 1984-2020 et sont représentés sur les figures ci-dessous. Un actuaire pourrait être amené à se poser les questions suivantes. Que peut-on en déduire sur le nombre moyen de communes touchées par une inondation reconnue Cat Nat? Que peut-on en déduire sur le coût annuel moyen des inondations? Quelle est la probabilité d'avoir plus de 1 milliards d'euros de sinistres Cat Nat inondation en une année?

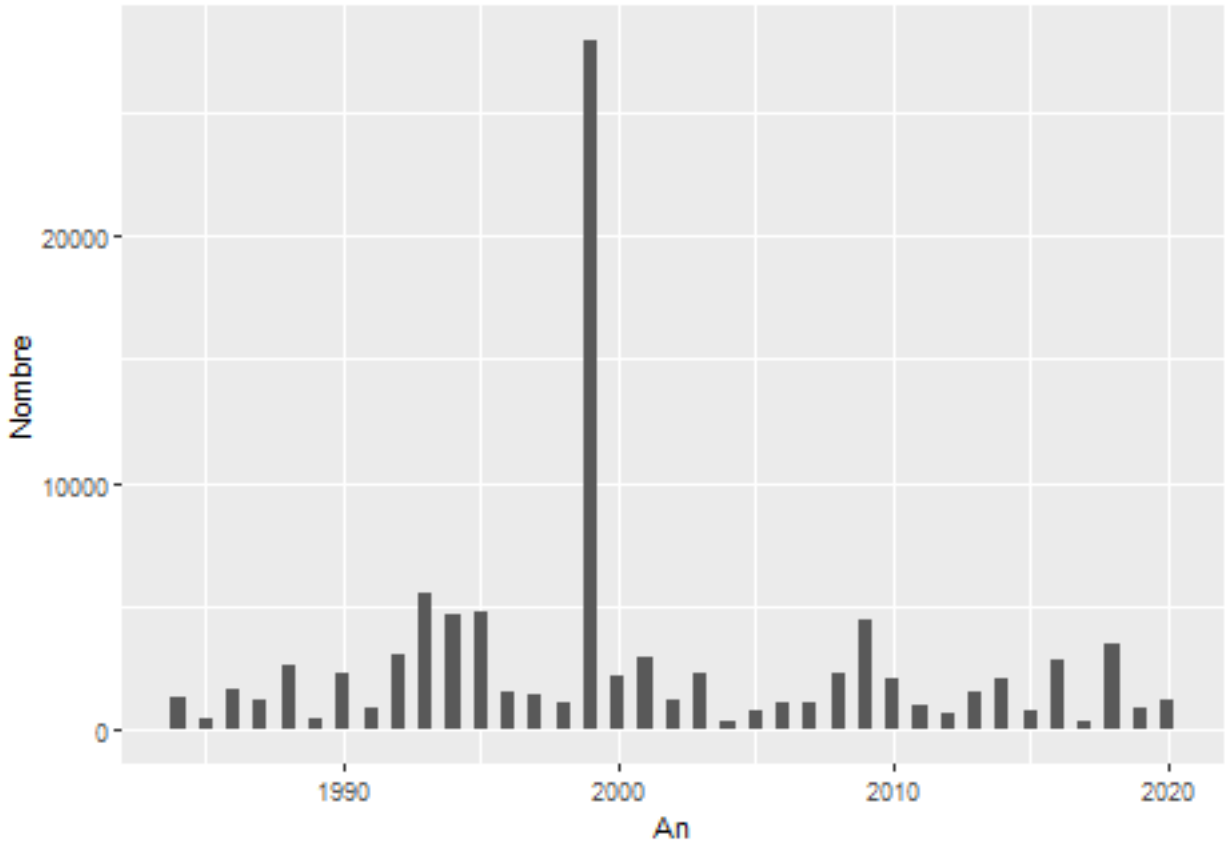
```
head(datasin)
```

```
##     An Montant Nombre
## 1 1984      52   1289
## 2 1985      29    529
## 3 1986     110   1621
## 4 1987     404   1203
## 5 1988     605   2675
## 6 1989      66    531
```

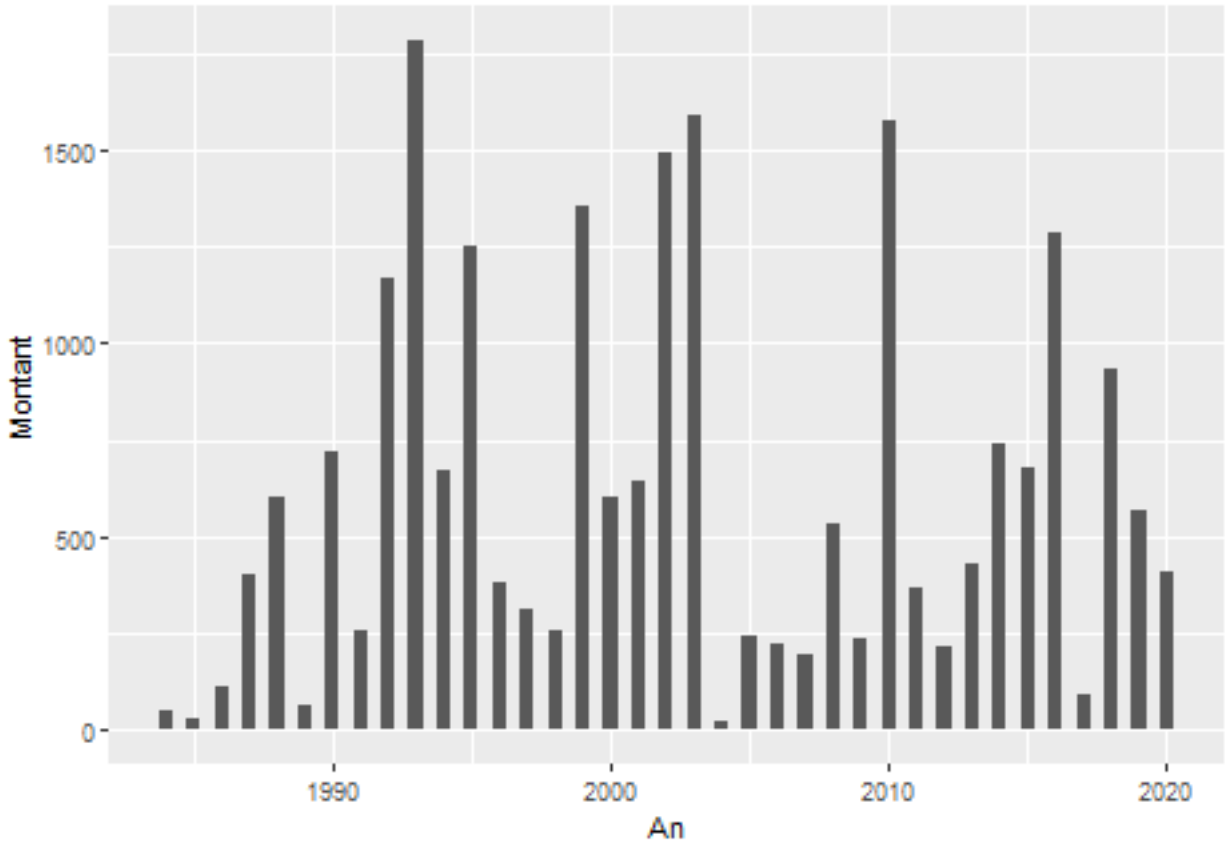
```
library(ggplot2)
```

```
## Warning: le package 'ggplot2' a été compilé avec la version R 4.2.2
```

```
ggplot(datasin, aes(x=An, y=Nombre)) +
  geom_bar(stat = "identity", width=0.5)
```



```r
ggplot(datasin, aes(x=An, y=Montant)) +
  geom_bar(stat = "identity", width=0.5)
```



La modélisation statistique repose principalement sur la **théorie des probabilités** qui sera considérée comme un prérequis de ce cours. D'autres outils sont importants pour un statisticien, notamment **l'informatique** qui permet d'implémenter les méthodes statistiques sur les jeux de données à disposition (souvent volumineux) ainsi que la connaissance du domaine concerné par les données (par exemple actuariat, santé, finance, environnement, etc) qui est généralement apportée par un expert du domaine pour guider la modélisation.

2 Théorie de l'estimation

2.1 Echantillon

On note $(x_1, ..., x_n) \in \mathcal{X}^n$ les données disponibles, avec n le nombre d'individus, x_i la valeur observée pour le ième individu. $\mathcal{X}$ représente l'espace des valeurs possibles pour les observations et dépend donc des données considérées. Revenons sur les exemples du cours.

- **Sondage.** On a $n = 1000$. Les données sont qualitatives, les personnes sondées ayant deux réponses possibles, 'Candidat A' ou 'Candidat B'. Il est toujours possible de recoder les variables qualitatives par des variables discrètes. Par exemple, on pourra supposer que $x_i = 0$ si la ième personne répond 'Candidat A' et $x_i = 1$ si ième personne répond 'Candidat B'. On a alors $\mathcal{X} = \{0, 1\}$.
- **Risque innondation.** Ici les individus sont les années et $n = 37$. Les données qui décrivent le nombre de sinistres sont à valeurs entières et $\mathcal{X} = \mathbb{N}$. Les données qui décrivent le montant des sinistres sont à valeurs réelles et $\mathcal{X} = \mathbb{R}$.

Dans le cadre de ce cours, on supposera que $\mathcal{X} \subset \mathbb{R}$. Cela exclut notamment les jeux de données multivariés comme ceux qui ont été vus en analyse de données au premier semestre (chaque individu était décrit par plusieurs variables, $\mathcal{X} \subset \mathbb{R}^p$).

Il existe généralement des sources d'incertitude dans la collecte des données. D'un point de vue mathématique, ces différentes sources d'incertitudes seront modélisées en supposant que **les observations sont une réalisation d'une expérience aléatoire**, c'est à dire qu'il existe un vecteur aléatoire $(X_1, ..., X_n)$ défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ tel que $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ avec $\omega \in \Omega$. Par exemple, dans les exemples introduits ci-dessus:

- **Sondage.** Les individus considérés sont choisis au hasard parmi un grand nombre d'individus. Si on recommence l'expérience, il y a de fortes chances qu'on choisisse d'autres individus et qu'on obtienne des résultats différents : le résultat de l'expérience est "aléatoire".
- **Risque innondation.** L'incertitude est liée à la complexité des phénomènes étudiés : la survenance d'un évènement naturel telle qu'une inondation ne peut pas être prévue de manière déterministe avec les méthodes scientifiques actuelles.
- **Erreur/imprécisions dans la collecte des données.** Par exemple, le calcul du montant total des sinistres sur une année peut être sujets à différentes erreurs/approximations qui peuvent être modélisées par des variables aléatoires.

Remarque : une variable aléatoire X_i est une application mesurable, la mesurabilité dépend donc de la tribu $\mathcal{A}$ qu'on choisit sur l'espace $\mathcal{X}$. Par défaut, on prendra la tribu des boréliens pour les variables continues et l'ensemble des parties de $\mathcal{X}$ pour les variables aléatoires discrètes ou finies.

L'étape de modélisation consiste à faire des hypothèses sur la loi de probabilité du vecteur aléatoire $(X_1, ..., X_n)$. **Dans le cadre de ce chapitre, on supposera que ce sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d)**. Il s'agit du cadre le plus simple, mais cette hypothèse n'est pas toujours réaliste.

- Lorsqu'on considère des phénomènes indexés par le temps, l'hypothèse d'indépendance n'est généralement pas vérifiée. Par exemple, si $(x_1, x_2, ..., x_n)$ désigne la température à Brest pendant n jours successifs, alors on ne peut généralement pas supposer que les observations successives x_i et x_{i+1} proviennent de variables aléatoires indépendantes (si il fait chaud un jour, il y a plus de chances qu'il fasse aussi chaud le lendemain).
- Dans de nombreuses situations, la loi de la variable dépend d'autres variables (les variables explicatives). L'hypothèse "identiquement distribuée" n'est plus vérifiée. Par exemple sur l'exemple des inondations, la distribution du montant ou du nombre de sinistres est potentiellement différente en 1984 et 2020 à cause du réchauffement climatique, des changements de réglementation, des changements dans les règles d'urbanisme, etc. L'hypothèse 'identiquement distribuée' semble donc discutable ici. Un cadre classique pour modéliser des données non i.i.d. est celui des modèles de régression.

Définition 1 *On appelle* ***n-échantillon*** *d'une loi de probabilité $\mathbb{P}$ une suite $(X_1, ..., X_n)$ de variables aléatoires (v.a.) indépendantes et identiquement distribuées (i.i.d.) qui suivent la loi de probabilité $\mathbb{P}$. On notera $X_1, ..., X_n \sim_{iid} \mathbb{P}$*

A retenir :

- On utilise des lettres minuscules (x_i) pour noter les observations. Ce sont des quantités déterministes (nombres réels) qui sont supposés être une réalisation d'une variable aléatoire.
- On utilise des lettres majuscules (X_i) pour noter les variables aléatoires associées. On suppose dans le cadre de ce cours que ces variables aléatoires sont i.i.d. ("échantillon").
- La quantité d'intérêt pour de nombreuses applications est la loi commune $\mathbb{P}$ de ces variables aléatoires ou certaines caractéristiques de cette loi (par exemple espérance, variance, quantiles).

2.2 Modèle paramétrique

On va supposer dans ce chapitre que la loi de probabilité commune de X_1, X_2, ..., X_n est un loi de probabilité qui dépend d'un **paramètre inconnu** $\theta \in \Theta$ avec $\Theta \subset \mathbb{R}^k$. On parle alors de modèle "paramétrique" et on notera

$$X_1, ..., X_n \sim_{iid} \mathbb{P}_\theta$$

Nous proposons ci-dessous des modèles paramétriques pour les trois exemples du cours.

- **Sondage.** On rappelle que les résultats sont codés de la manière suivante
  - $x_i = 0$ si la ième personne sondée pense voter pour A
  - $x_i = 1$ si la ième personne sondée pense voter pour B

  Il semble naturel de supposer que $(x_1, ..., x_n)$ est une réalisation d'un échantillon $(X_1, ..., X_n)$ d'une loi de Bernoulli de paramètre inconnu θ avec $\theta \in \Theta =]0, 1[$. $\theta = P[X_i = 1]$ représente la probabilité qu'un individu choisi au hasard vote pour le candidat B. Ce modèle paramétrique (appelé "modèle de Bernoulli") sera noté

  $$X_1, ..., X_n \sim_{iid} Ber(\theta)$$

  avec $\theta \in]0, 1[$ un paramètre inconnu.

- **Nombre d'inondations.** Ici les observations sont à valeurs dans $\mathcal{X} = \mathbb{N}$. Une loi usuelle en actuariat pour modéliser le nombre d'évènements ou de sinistres est la loi de Poisson. On rappelle que $X_i \sim Pois(\theta)$ si

  $$P(X_i = x_i) = exp(-\theta)\frac{\theta^{x_i}}{x_i!}$$

  pour $x_i \in \mathbb{N}$. Un modèle paramétrique possible pour décrire les observations consiste alors à supposer (modèle de Poisson) que

  $$X_1, ..., X_n \sim_{iid} Pois(\theta)$$

  avec $\theta \in \mathbb{R}^{+*}$ un paramètre inconnu. D'autres choix de modèles paramétriques seraient possibles ici et il serait intéressant de vérifier que le choix de la loi de Poisson est approprié pour les données. Ceci sera discuté dans la suite du cours (cf paragraphe sur le test du χ^2).

- **Montant des inondations.** Ici les observations sont à valeurs dans $\mathcal{X} = \mathbb{R}$. Le modèle paramétrique le plus classique pour les variables aléatoires réelles est le modèle gaussien. On suppose alors (modèle gaussien) que

  $$X_1, ..., X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$$

  avec $\mathcal{N}(\mu, \sigma^2)$ la loi normale d'espérance μ et de variance σ^2 et $\theta = (\mu, \sigma^2)$ le paramètre inconnu à valeurs dans $\Theta = \mathbb{R} \times \mathbb{R}^{+*}$. D'autres choix de modèles paramétriques seraient possibles, par exemple on utilise souvent la loi Gamma pour décrire le montant des sinistres en actuariat. Le modèle gaussien peut parfois se justifier de manière théorique par le théorème central limite, lorsque les observations sont obtenues en faisant la moyenne sur un grand nombre de valeurs. Nous verrons dans la suite du cours que le choix de la loi normale a de nombreux avantages d'un point de vue analytique et qu'il existe des méthodes pour vérifier si le choix de la loi normale est approprié.

2.3 Estimateur

Une fois le modèle paramétrique choisi, on cherche généralement à estimer le paramètre inconnu θ à partir des observations disponibles.

Définition 2 *Soit $(X_1, ..., X_n)$ un n-échantillon d'une loi P_θ. Un **estimateur** du paramètre inconnu θ est une variable aléatoire $T = g(X_1, ..., X_n)$ qui s'exprime en fonction de $(X_1, ..., X_n)$. Une **estimation** de θ est alors la valeur numérique prise par cette statistique sur la réalisation particulière qui correspond aux données, c'est à dire la quantité $t = g(x_1, ..., x_n)$.*

Nous proposons ci-dessous des estimateurs pour les trois exemples du cours.

- **Sondage.** On considère le modèle paramétrique $X_1, ..., X_n \sim_{iid} Ber(\theta)$. Le paramètre inconnu θ est la probabilité qu'un individu choisi au hasard vote pour le candidat B. Une estimation naturelle de cette probabilité est la **fréquence empirique** de vote pour le candidat B parmi les 1000 personnes sondées, c'est à dire

  $$f = \frac{1}{n} card\{i \in \{1, ..., n\} | x_i = 1\} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Pour le sondage réalisé, l'estimation obtenue est $f = 0.48$. Cette valeur numérique est une réalisation de la variable aléatoire

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

qui est un estimateur du paramètre inconnue θ.

- **Nombre d'inondations.** On considère le modèle paramétrique $X_1, ..., X_n \sim_{iid} Pois(\theta)$ avec $\theta \in \mathbb{R}^{+*}$ le paramètre inconnu. On rappelle que l'espérance d'une loi de Poisson est donnée par $E[X_i] = \theta$. Un estimateur usuel de $E[X_i]$ est la **moyenne empirique** définie par

$$\bar{X} = \frac{X_1 + ... + X_n}{n}.$$

On peut faire l'application numérique sur les données (cf ci-dessous), et on obtient $\bar{x} = 2629$ comme estimation de θ.

```
mean(datasin$Nombre)
```

```
## [1] 2628.865
```

- **Montant des inondations.** On considère le modèle paramétrique

$$X_1, ..., X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$$

avec $\mu = E[X_i]$, $\sigma^2 = var(X_i)$ et $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{+*}$ le paramètre inconnu. Un estimateur naturel de μ est la moyenne empirique introduite ci-dessus. Un estimateur usuel de $\sigma^2 = var(X) = E[X^2] - E[X]^2$ est la **variance empirique** définie par

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

On peut faire l'application numérique sur les données (cf ci-dessous), et on obtient $\bar{x} = 609$ comme estimation de μ et $s^2 = 495$ comme estimation de σ^2. $\bar{x}$ s'interprète comme le montant moyen des sinistres et $s = \sqrt{s^2}$ mesure la dispersion des sinistres autour de la valeur moyenne. Ces deux quantités ont la même unité que les données de départ.

```
mean(datasin$Montant)
```

```
## [1] 608.7838
```

```
mean(datasin$Montant^2)-mean(datasin$Montant)^2
```

```
## [1] 244741
```

```
sqrt(mean(datasin$Montant^2)-mean(datasin$Montant)^2)
```

```
## [1] 494.7131
```

Le tableau ci-dessous récapitule les trois exemples discutés dans le cours.

Dans la suite du cours, les variables aléatoires (par exemple X_i, F, $\bar{X}$ et S) sont notées avec des lettres majuscules, les observations (x_i) et les estimations (f, $\bar{x}$, s) avec des lettres minuscules. Les paramètres inconnus sont notés avec des lettres grecques (par exemple θ, μ, σ).

Sur l'exemple des sinistres inondation, on appelle souvent (abusivement) "moyenne" les quantités μ, $\bar{x}$ et $\bar{X}$. Pourtant ce sont des objets de nature différente!

	$\mathcal{X}$	$\mathcal{A}$	$\mathbb{P}_\theta$	Θ	Estimateur	Estimation
Sondage	$\{0,1\}$	$\mathcal{P}(\{0,1\})$	$Ber(\theta)$	$]0,1[$	$F=\bar{X}$	0.48
Nombre d'inondations	$\mathbb{N}$	$\mathcal{P}(\mathbb{N})$	$Pois(\theta)$	$\mathbb{R}^{+*}$	$\bar{X}$	2629
Montant des inondations	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$	$\mathcal{N}(\mu,\sigma^2)$	$\mathbb{R}\times\mathbb{R}^{+*}$	$(\bar{X},S)$	(609,495)

Table 1: Modèle, estimateur et estimation pour les trois exemples du cours.

2.4 Propriétés des estimateurs

On peut toujours définir une infinité d'estimateurs pour un paramètre inconnu donné et en pratique on cherchera à utiliser le "meilleur" de ces estimateurs. Ceci nécessite de définir ce qu'est un bon estimateur. Dans ce paragraphe, T est un estimateur du paramètre inconnu $\theta \in \mathbb{R}$ (paramètre inconnu scalaire).

2.4.1 Biais d'un estimateur

Définition 3 *On appelle* **biais** *de l'estimateur T la quantité*

$$biais(T) = E(T) - \theta.$$

On dit que l'estimateur T est **sans biais** *lorsque $biais(T) = 0$, c'est à dire lorsque $E[T] = \theta$. Lorsque $biais(T) > 0$ (resp. $biais(T) < 0$) on dit que l'estimateur sur-estime (resp. sous-estime) le paramètre inconnu. Le biais représente "l'erreur moyenne" qui est faite lorsqu'on utilise T pour estimer θ.*

Proposition 1

- *Si $(X_1, ..., X_n)$ est un n-échantillon d'une loi d'espérance $E[X_i] = \mu$, alors $\bar{X}$ est un estimateur sans biais de μ. En particulier, si $(X_1, ..., X_n)$ est un n-échantillon de Bernoulli de paramètre θ alors $F = \bar{X}$ est un estimateur sans biais de θ.*
- *Si on suppose en outre que $var(X_i) = \sigma^2 < \infty$ alors $E[S^2] = \frac{n-1}{n}\sigma^2$. S^2 est donc un estimateur biaisé de σ^2, et on préfère parfois utiliser l'estimateur corrigé*

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

qui est un estimateur sans biais de σ^2.

Preuve. Si $(X_1, ..., X_n)$ est un n-échantillon d'une loi d'espérance μ alors

$$\begin{aligned} E[\bar{X}] &= E[\frac{X_1 + ... + X_n}{n}] \\ &= \frac{E[X_1] + ... + E[X_n]}{n} \\ &= \mu. \end{aligned}$$

On suppose que $\sigma^2 < \infty$. Par définition,

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

et de la décomposition $(X_i - \bar{X}) = (X_i - \mu) - (\bar{X} - \mu)$, on déduit que:

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) + (\bar{X} - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + (\bar{X} - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2. \end{aligned}$$

Donc

$$\begin{aligned} E[S^2] &= E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2]-E[(\bar{X}-\mu)^2] \\ &= \sigma^2-E[(\bar{X}-\mu)^2]. \end{aligned}$$

Il reste à calculer

$$\begin{aligned} E[(\bar{X}-\mu)^2] &= var(\bar{X}) \\ &= var(\frac{1}{n}\sum_{i=1}^{n}X_i) \\ &= \frac{1}{n^2}var(\sum_{i=1}^{n}X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}var(X_i) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Finalement,

$$E[S^2] = \frac{n-1}{n}\sigma^2.$$

■

Remarque 1 *De nombreux logiciels statistiques (Excel, R, SAS mais pas Python) calculent par défaut l'estimateur sans biais de la variance S^2_{corr} défini ci-dessus. Ceci est illustré dans les codes R ci-dessous.*

```
n=10
x=rnorm(n)  #simulation d'un échantillon de taille n d'une loi N(0,1)
x  #valeurs simulées
```

```
##  [1]  0.38321969 -0.85020477 -0.42924106  1.02912046  0.27405655 -1.55197871
##  [7]  0.16244966  0.09936925  1.99601239 -0.71370070
```

```
#la variance théorique de la loi est sigma^2=1.
sum((x-mean(x))^2)/n #estimateur biaisé
```

```
## [1] 0.9110557
```

```
sum((x-mean(x))^2)/(n-1) #estimateur sans biais
```

```
## [1] 1.012284
```

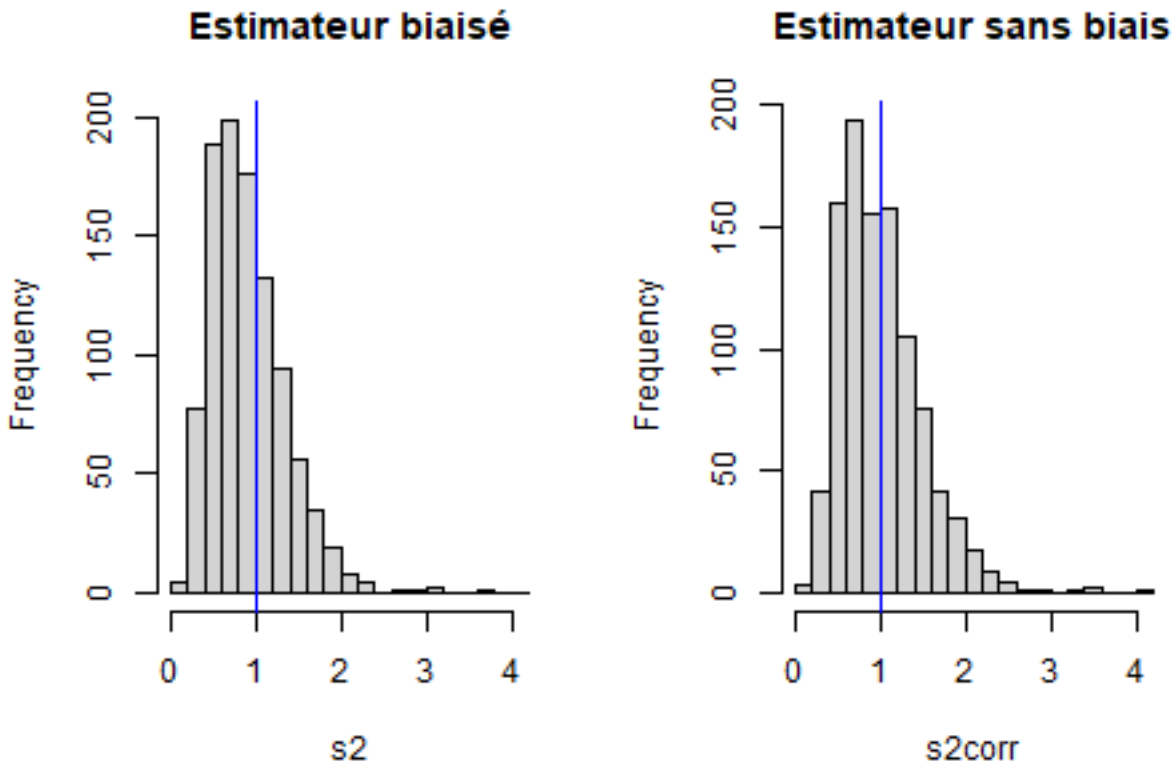
```
var(x) #var calcule l'estimateur sans biais
```

```
## [1] 1.012284
```

Les simulations ci-dessous montrent que l'estimateur biaisé S^2 tend à sous-estimer la variance. Cette sous-estimation systématique disparaît avec l'estimateur sans biais. La différence entre les deux estimateurs peut être importante pour les petits échantillons (n petit) mais devient négligeable pour les grands échantillons.

```r
#répétons l'expérience précédente 1000 fois et stockons les résultats
s2=NULL
s2corr=NULL
for (i in 1:1000){
  x=rnorm(n)  #simulation d'un échantillon de taille n d'une loi N(0,1)
  s2[i]=sum((x-mean(x))^2)/n #estimateur biaisé
  s2corr[i]=sum((x-mean(x))^2)/(n-1) #estimateur sans biais
}

#On peut représenter la distribution empirique des 1000 valeurs simulées avec un histogramme
par(mfrow=c(1,2))
hist(s2, breaks=seq(0,max(s2corr+.2),by=.2),main='Estimateur biaisé')
abline(v=1,col='blue')
hist(s2corr, breaks=seq(0,max(s2corr+.2),by=.2), main='Estimateur sans biais')
abline(v=1,col='blue')
```




```r
mean(s2)  #moyenne des estimations biaisées : sous-estime la vraie valeur 1
```

```
## [1] 0.9058436
```

```r
mean(s2corr)  #moyenne des estimations non-biaisées : proche de la vraie valeur 1
```

```
## [1] 1.006493
```

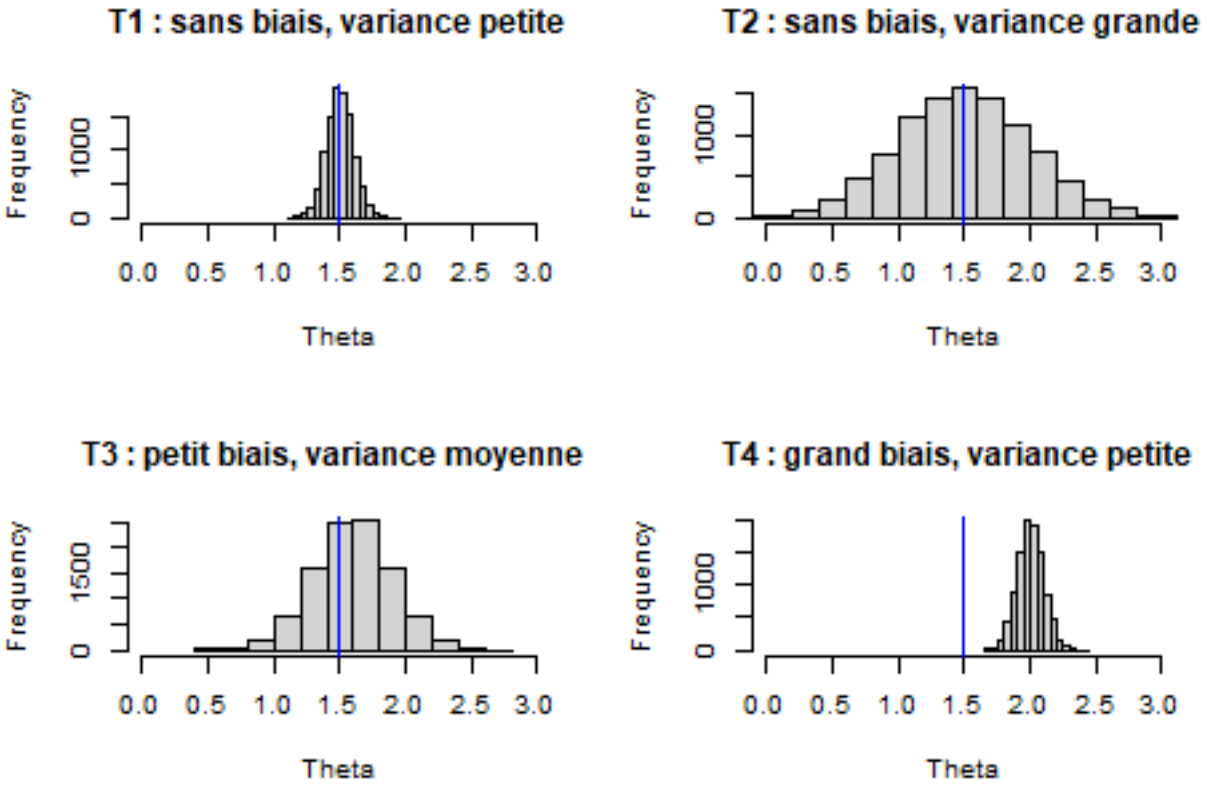
2.4.2 Erreur quadratique d'un estimateur

La figure ci-dessous montre la distribution de 4 estimateurs du paramètre inconnu matérialisé par la ligne bleue.

Les estimateurs T_1 et T_2 sont sans biais. L'estimateur T_1 semble 'plus proche' en général de la vraie valeur que l'estimateur T_2 car sa variance (qui mesure la dispersion autour de son espérance) est plus faible.

Il est plus difficile de comparer les estimateurs T_2, T_3 et T_4 :

- T_2 est sans biais, mais sa variance est importante,
- T_3 est légèrement biaisé (légère surestimation en moyenne) mais sa variance est plus faible de T_2 et il semble en général plus proche de la vraie valeur que T_2,
- T_3 est fortement biaisé (les estimations obtenues sont systématiquement au-dessus de la vraie valeur) mais sa variance est plus petite que les autres estimateurs.



Le risque quadratique d'un estimateur permet de prendre en compte à la fois son biais et sa variance.

Définition 4 *L'**erreur quadratique moyenne** (**EQM**) de l'estimateur T est définie par*

$$EQM(T) = E[(T-\theta)^2].$$

On dira que l'estimateur T_1 est plus précis que l'estimateur T_2 si $EQM(T_1) < EQM(T_2)$.

Remarque 2 *On peut vérifier que*

$$E[(T-\theta)^2] = var(T) + E[(T-\theta)]^2$$

c'est à dire que l'erreur quadratique moyenne est égale à la variance de l'estimateur plus le biais de l'estimateur au carré. Lorsque l'estimateur est non-biaisé, l'EQM coïncide avec la variance : parmi deux estimateurs sans biais, le plus précis est donc celui de variance minimale. Cette formule est également utile en pratique pour calculer l'EQM des estimateurs usuels.

Remarque 3 *Sur l'exemple de la figire précédente, on peut vérifier que*

$$EQM(T_1) < EQM(T_3) < EQM(T_2) < EQM(T_4).$$

En particulier, l'estimateur biaisé T_3 est plus précis que l'estimateur sans biais T_2 .

Proposition 2 *Si $(X_1, ..., X_n)$ est un n-échantillon d'une loi d'espérance μ et de variance $\sigma^2 < \infty$, alors*

$$EQM(\bar{X}) = var(\bar{X}) = \frac{\sigma^2}{n}.$$

En particulier, si $(X_1, ..., X_n)$ est un échantillon de Bernoulli de paramètre θ alors

$$EQM(F) = var(F) = \frac{\theta(1-\theta)}{n}.$$

Si de plus $\mu_4 = E[(X_i - \mu)^4] < \infty$, alors

$$EQM(S^2_{corr}) = var(S^2_{corr}) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4.$$

Preuve. (partielle)
Soit $(X_1, ..., X_n)$ un échantillon d'une loi d'espérance μ et de variance $\sigma < \infty$. On a vu que $\bar{X}$ est un estimateur sans biais de μ et donc

$$\begin{aligned} EQM(\bar{X}) &= var(\bar{X}) \\ &= var(\frac{X_1 + ... + X_n}{n}) \\ &= \frac{var(X_1) + ... + var(X_n)}{n^2} \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Le calcul de $var(S^2_{corr})$ est plus délicat... ■

2.4.3 Propriétés asymptotiques

Un bon estimateur doit avoir de bonnes "propriétés asymptotiques", c'est à dire des propriétés de convergence lorsque la taille de l'échantillon $n \to \infty$.

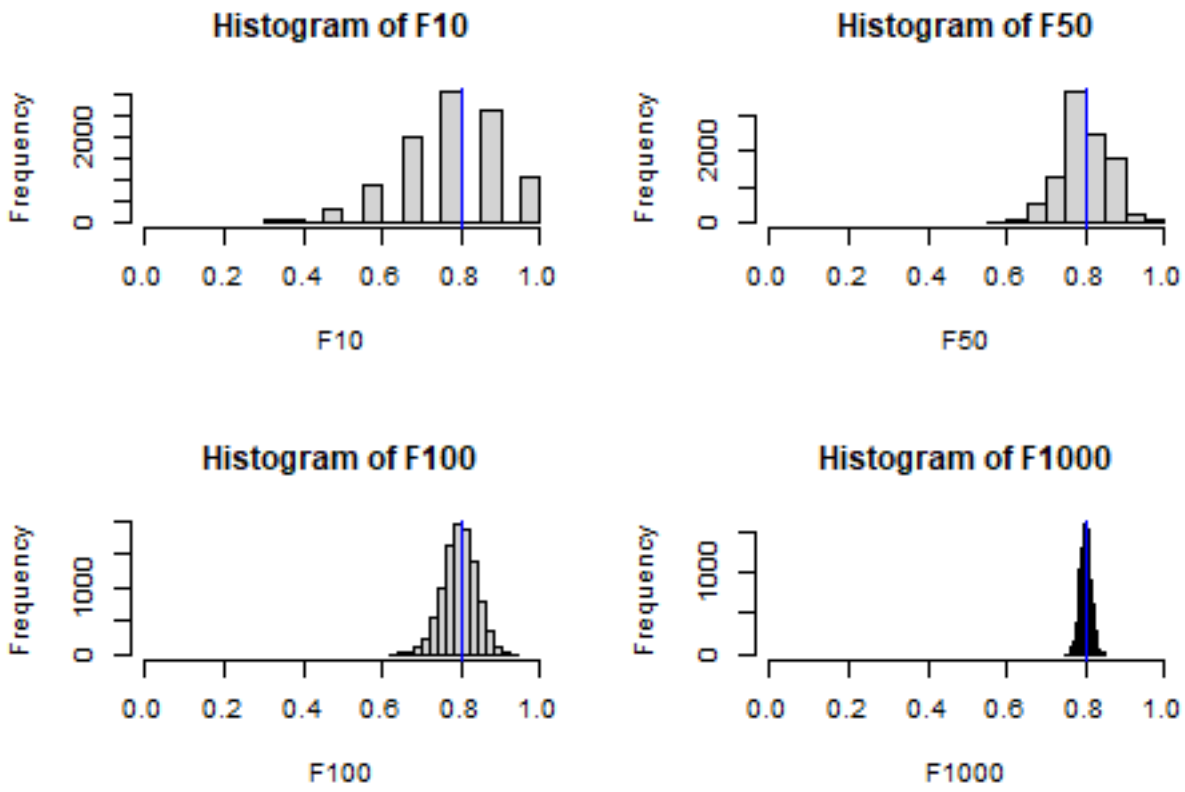
Les codes et la figure ci-dessous illustrent la convergence de l'estimateur de θ dans le cas des échantillons de loi Bernoulli $(X_1, ..., X_n) \sim_{iid} Ber(\theta)$. On note

$$F_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} card\{i \in \{1, ...n, \} | X_i = 1\}$$

la fréquence empirique calculée sur un échantillon de taille n. Lorsque n augmente, F_n se 'concentre' de plus en plus autour de θ (LGN) et la forme de l'histogramme se 'rapproche' d'une gaussienne (TCL).

```
theta=.8
F10=NULL
F50=NULL
F100=NULL
F1000=NULL
for (i in 1:10^4){
  x=rbinom(10,size=1,prob=theta)  #simulation échantillon de Bernoulli de taille n=10
  F10[i]=mean(x)
  x=rbinom(50,size=1,prob=theta)  #simulation échantillon de Bernoulli de taille n=50
  F50[i]=mean(x)
  x=rbinom(100,size=1,prob=theta) #simulation échantillon de Bernoulli de taille n=100
  F100[i]=mean(x)
    x=rbinom(1000,size=1,prob=theta) #simulation échantillon de Bernoulli de taille n=1000
  F1000[i]=mean(x)
}
```

```
par(mfrow=c(2,2))
hist(F10,xlim=c(0,1))  #Histogramme
abline(v=theta,col='blue')  #vraie valeur
hist(F50,xlim=c(0,1))
abline(v=theta,col='blue')
hist(F100,xlim=c(0,1))
abline(v=theta,col='blue')
hist(F1000,xlim=c(0,1))
abline(v=theta,col='blue')
```



Proposition 3 *(Rappel : loi des grands nombres et théorème central limite) Soit (X_i) une suite de variables aléatoires i.i.d d'une loi d'espérance μ et de variance $\sigma^2 < \infty$. Alors (loi des grands nombres)*

$$\bar{X}_n = \frac{X_1 + ... + X_n}{n}$$

converge presque sûrement (p.s.) vers μ et (théorème central limite)

$$\sqrt{n}(\bar{X}_n - \mu)$$

converge en loi vers une loi $\mathcal{N}(0, \sigma^2)$.

Définition 5 *T_n est un estimateur **convergent** de θ lorsque T_n converge p.s. vers θ lorsque $n \to \infty$.*

Proposition 4 *Si $(X_1, ..., X_n)$ est un n-échantillon d'une loi d'espérance μ et de variance $\sigma^2 < \infty$ alors*

$$\bar{X}_n = \frac{X_1 + ... + X_n}{n}$$

est un estimateur convergent de μ. En particulier, si $(X_1, ..., X_n)$ est un échantillon de Bernoulli de paramètre θ alors

$$F_n = \frac{X_1 + ... + X_n}{n}$$

est un estimateur convergent de θ.

Si de plus $\mu_4 = E[(X_i - \mu)^4] < \infty$ alors

$$S_n^2 = \frac{X_1^2 + ... + X_n^2}{n} - \bar{X}_n^2$$

et

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

sont des estimateurs convergents de σ^2.

Preuve. Applications directes de la loi des grands nombres (LGN). ■

Si T_n est un estimateur **convergent** de θ lorsque $T_n - \theta$ converge p.s. vers 0 lorsque $n \to \infty$. Il est alors naturel de s'intéresser à la vitesse de cette convergence. De nombreux estimateurs vérifient un TCL, c'est à dire sont tels que

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta))$$

lorsque $n \to \infty$. On parle alors de "**normalité asymptotique**". Intuitivement, cela signifie que les estimateurs usuels convergent avec une vitesse en $\frac{1}{\sqrt{n}}$ et qu'on peut utiliser l'approximation

$$\sqrt{n}(T_n - \theta) \approx \mathcal{N}(0, \sigma^2(\theta))$$

lorsque n est "suffisamment grand". Ce type de comportement asymptotique est couramment utilisé pour construire des intervalles de confiance ou réaliser des tests (cf paragraphes suivants) et est donc particulièrement souhaitable.

Proposition 5 *Soit $(X_1, ..., X_n)$ est un échantillon d'une loi d'espérance μ et de variance $\sigma^2 < \infty$ alors*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

lorsque $n \to \infty$. En particulier, si $(X_1, ..., X_n)$ est un échantillon de la loi de Bernoulli de paramètre θ et $F_n = \frac{X_1 + ... + X_n}{n}$, alors

$$\sqrt{n}(F_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$$

lorsque $n \to \infty$.

Preuve. Utilisation directe du théorème central limite (TCL). ■

2.4.4 Lemme de Slutsky et Delta méthode

Les deux propositions ci-dessous sont utiles pour étudier les propriétés asymptotiques de certains estimateurs.

Proposition 6 *(Lemme de Slutsky) Soient $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ deux suites de v.a.r., X une v.a.r. et a un nombre réel tels que*

$$X_n \xrightarrow{\mathcal{L}} X \text{ et } Y_n \xrightarrow{\mathcal{L}} a.$$

Alors

$$X_n + Y_n \xrightarrow{\mathcal{L}} X + a,\ X_n Y_n \xrightarrow{\mathcal{L}} Xa \text{ et } \frac{X_n}{Y_n} \xrightarrow{\mathcal{L}} \frac{X}{a} \text{ si } a \neq 0.$$

Remarque 4 *Attention : les résultats de la proposition précédente ne sont plus vraies si Y_n converge vers une variable aléatoire Y non-dégénérée. On peut vérifier que si a est un nombre réel alors*

$$Y_n \xrightarrow{\mathcal{L}} a \Leftrightarrow Y_n \xrightarrow{P} a.$$

On peut utiliser le lemme de Slutsky pour démontrer la proposition suivante qui sera utilisée dans la suite du cours pour construire des intervalles de confiance et des tests d'hypothèses. La proposition est assez proche de la proposition 5, mais la loi limite ne dépend plus des paramètres inconnus.

Proposition 7 *1. Soit $(X_1, ..., X_n)$ est un échantillon d'une loi d'espérance μ et de variance $\sigma^2 < \infty$ alors*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

2. Soit $(X_1, ..., X_n)$ un échantillon de la loi de Bernoulli de paramètre θ et $F_n = \frac{X_1+...+X_n}{n}$ alors

$$\sqrt{n}\frac{F_n - \theta}{\sqrt{F_n(1-F_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Preuve.

1. On a vu que sous les conditions de la proposition on a, lorsque $n \to \infty$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} Z$$

avec $Z \sim \mathcal{N}(0, \sigma^2)$, $\mu = E[X_i]$, $\sigma^2 = var(X_i)$ et

$$S_n \xrightarrow{P} \sigma.$$

En utilisant le lemme de Slutsky, on en déduit que $\sqrt{n}\frac{\bar{X}_n-\mu}{S_n} \xrightarrow{\mathcal{L}} \frac{Z}{\sigma}$ c'est à dire que

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

2. On a vu que lorsque $n \to \infty$

$$\sqrt{n}(F_n - \theta) \xrightarrow{\mathcal{L}} Z \text{ et } F_n \xrightarrow{P} \theta$$

avec $Z \sim \mathcal{N}(0, \theta(1-\theta))$. On déduit du lemme de Slutsky que $\sqrt{n}\frac{F_n-\theta}{\sqrt{F_n(1-F_n)}} \xrightarrow{\mathcal{L}} \frac{Z}{\sqrt{\theta(1-\theta)}}$ c'est à dire que

$$\sqrt{n}\frac{F_n - \theta}{\sqrt{F_n(1-F_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

■

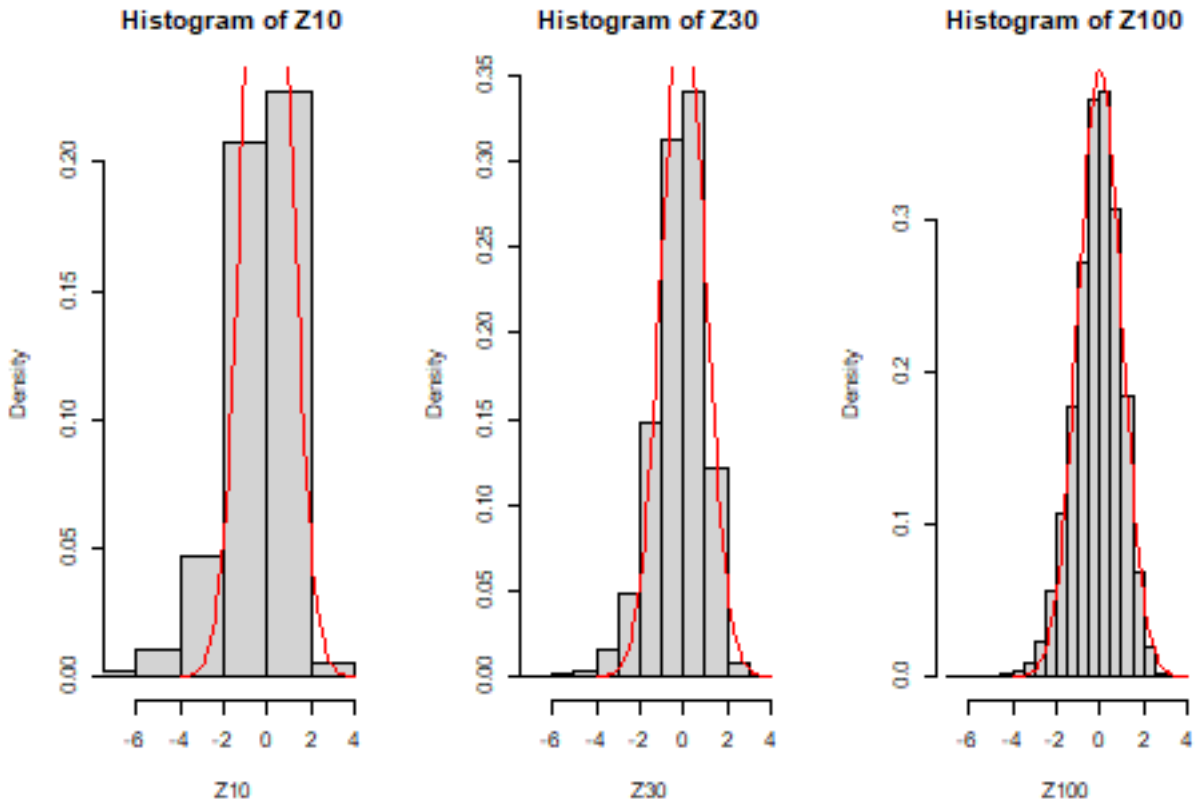
A noter que la quantité $\sqrt{n}\frac{\bar{X}_n-\mu}{S_n}$ dépend uniquement de μ et de l'échantillon $(X_1, ..., X_n)$ et que la loi limite ne dépend pas de la loi des X_i. Ce résultat sera utilisé dans la suite du cours pour construire des intervalles de confiance et des tests d'hypothèses. En pratique, on admet généralement que l'approximation

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \approx \mathcal{N}(0,1)$$

est valide lorsque $n \geq 30$. Ceci est illustré sur des simulations ci-dessous.

```
Z10=NULL
Z30=NULL
Z100=NULL
mu=1
for (i in 1:10^4){
  x=rexp(10)  #on peut remplacer la loi Exponentielle par une autre loi
  Z10[i]=sqrt(10)*(mean(x)-mu)/sd(x)
  x=rexp(30)
```

```
  Z30[i]=sqrt(30)*(mean(x)-mu)/sd(x)
  x=rexp(100)
  Z100[i]=sqrt(100)*(mean(x)-mu)/sd(x)
}
xx=seq(-4,4,by=.1)
par(mfrow=c(1,3))
hist(Z10,xlim=c(-7,4),freq=FALSE)  #freq=FALSE permet de 'normaliser' l'histogramme
lines(xx,dnorm(xx),col='red')  #ajout de la densité de la loi N(0,1)
hist(Z30,xlim=c(-7,4),freq=FALSE)
lines(xx,dnorm(xx),col='red')
hist(Z100,xlim=c(-7,4),freq=FALSE)
lines(xx,dnorm(xx),col='red')
```



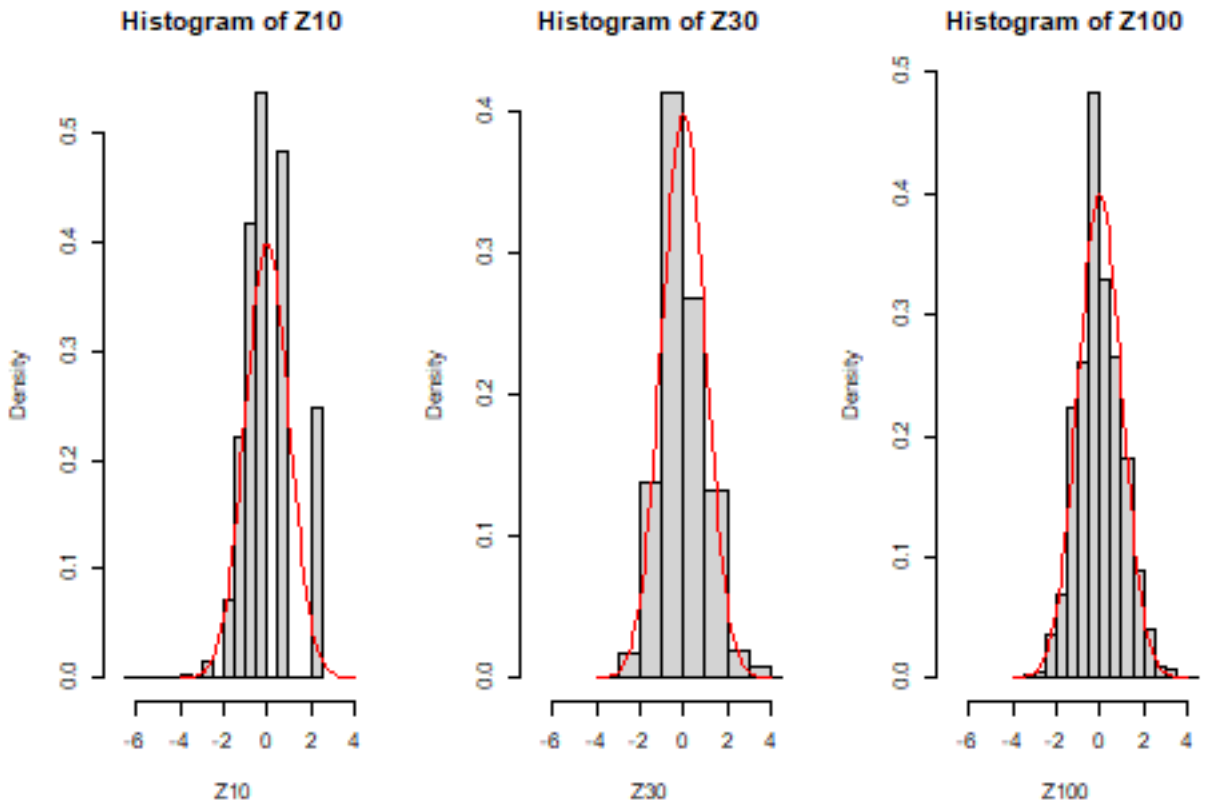
En pratique, dans le cas des échantillons de Bernoulli, on admet généralement que l'approximation

$$\sqrt{n}\frac{F_n-\theta}{\sqrt{F_n(1-F_n)}} \approx \mathcal{N}(0,1)$$

est valide lorsque $n\theta \geq 5$ et $n(1-\theta) \geq 5$. Ceci est illustré sur les simulations ci-dessous.

```
Z10=NULL
Z30=NULL
Z100=NULL
theta=.7
for (i in 1:10^4){
  x=rbinom(10,size=1,prob=theta)
  F=mean(x)
  Z10[i]=sqrt(10)*(F-theta)/sqrt(F*(1-F))
  x=rbinom(30,size=1,prob=theta)
  F=mean(x)
```

```
  Z30[i]=sqrt(30)*(F-theta)/sqrt(F*(1-F))
  x=rbinom(100,size=1,prob=theta)
  F=mean(x)
  Z100[i]=sqrt(100)*(F-theta)/sqrt(F*(1-F))
}
xx=seq(-4,4,by=.1)
par(mfrow=c(1,3))
hist(Z10,xlim=c(-7,4),freq=FALSE)  #freq=FALSE permet de 'normaliser' l'histogramme
lines(xx,dnorm(xx),col='red')  #ajout de la densité de la loi N(0,1)
hist(Z30,xlim=c(-7,4),freq=FALSE)
lines(xx,dnorm(xx),col='red')
hist(Z100,xlim=c(-7,4),freq=FALSE)
lines(xx,dnorm(xx),col='red')
```



Proposition 8 *(Delta méthode) Si T_n est un estimateur convergent de θ et g est continue en θ alors $g(T_n)$ est un estimateur convergent de $g(\theta)$. Si de plus*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

et g est dérivable en θ avec $g'(\theta) \neq 0$ alors

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (g'(\theta))^2\sigma^2).$$

Preuve. (idée de la preuve) : comme g est dérivable en θ et $T_n \approx \theta$, on peut écrire un développement limité de la forme $g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + R_n$. On a donc $\sqrt{n}(g(T_n) - g(\theta)) \approx \sqrt{n}g'(\theta)(T_n - \theta)$ avec

$$\sqrt{n}g'(\theta)(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (g'(\theta))^2\sigma^2).$$

■

Exemple d'application de la proposition précédente. On a vu que sous des conditions générales $\bar{X}_n$ est un estimateur convergent de μ et $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ lorsque $n \to \infty$. Appliquons la proposition précédente avec $g(x) = exp(x)$. On en déduit que $exp(\bar{X}_n)$ est un estimateur convergent de $exp(\mu)$ et

$$\sqrt{n}(exp(\bar{X}_n) - exp(\mu)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, exp(2\mu)\sigma^2).$$

2.5 Quelques méthodes d'estimation

Dans ce paragraphe, $(X_1, ..., X_n)$ est un n-échantillon d'une loi P_θ avec $\theta \in \Theta \subset \mathbb{R}^k$. Ce paragraphe introduit deux méthodes générales pour construire un estimateur de θ.

2.5.1 Méthode des moments

2.5.2 Méthode des moments

On suppose que $\theta \in \Theta \subset \mathbb{R}^k$ (k paramètres à estimer). On note $m_d(\theta) = E[X_i^d]$ le moment (théorique) d'ordre d de la loi P_θ et $M_d = \frac{1}{n}\sum_{i=1}^n X_i^d$ son estimateur usuel (moment empirique). La méthode des moments dans sa version la plus classique consiste à

1. exprimer les paramètres du modèle en fonction des p premiers moments théoriques de la loi, c'est à dire trouver une fonction $G : \mathbb{R}^p \to \Theta$ telle que $\theta = G(m_1(\theta), ..., m_p(\theta))$
2. remplacer les moments théoriques par les moments empiriques. L'estimateur des moments de θ est alors défini par $T = G(M_1, ..., M_p)$.

On peut vérifier que les estimateurs introduits pour les trois exemples du cours sont des estimateurs obtenus par la méthode des moments.

- **Modèle de Bernoulli :** $X_1, ..., X_n \sim_{iid} Ber(\theta)$. On a $k = 1$ et $\theta = G(E[X_i])$ avec $G(x) = x$. Un estimateur de θ par la méthode des moments est $T = G(M_1) = \bar{X}$.
- **Modèle de Poisson :** $X_1, ..., X_n \sim_{iid} Pois(\theta)$. On a $k = 1$ et $\theta = G(E[X_i])$ avec $G(x) = x$. Un estimateur de θ par la méthode des moments est $T = G(M_1) = \bar{X}$. On pourrait aussi utiliser que $\theta = var(X_i) = E[X_i^2] - E[X_i]^2 = G_2(E[X_i], E[X_i^2])$ avec $G_2(x, y) = y - x^2$ et un autre estimateur de θ par la méthode des moments est $T_2 = G_2(M_1, M_2) = M_2 - M_1^2 = S^2$. En pratique, on favorise généralement les estimateurs basés sur les moments d'ordre les plus faibles qui sont généralement les plus 'faciles' à estimer.
- **Modèle Gaussien :** $X_1, ..., X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$. On a $k = 2$ $\theta = (\mu, \sigma) = G(E[X_i], E[X_i^2])$ avec $G(u, v) = (u, \sqrt{v - u^2})$. Un estimateur de θ par la méthode des moments est $T = G(M_1, M_2) = (\bar{X}, S)$.

Sous des conditions générales, d'après la loi des grands nombres et le théorème centrale limite, les estimateurs des moments sont convergents et asymptotiquement gaussiens. On peut alors en déduire, en utilisant la delta-méthode, que si la fonction G est suffisamment régulière alors les estimateurs basés sur la méthode des moments sont convergents et asymptotiquement gaussiens (cf TD).

2.5.3 Méthode du maximum de vraisemblance

La fonction de vraisemblance définie ci-dessous joue un rôle fondamental en statistique. On suppose dans la première partie du paragraphe que les variables aléatoires X_i sont à valeurs dans un ensemble discret $\mathcal{X}$. Le cas continu est discuté dans un deuxième temps.

Définition 6 *On suppose que $\mathcal{X} \subset \mathbb{N}$. On appelle alors fonction de vraisemblance la fonction*

$$L(\theta; x_1, ..., x_n) = P_\theta[X_1 = x_1, ..., X_n = x_n]$$

la probabilité d'observer les données $(x_1, ..., x_n)$ si θ est la vraie valeur des paramètres. On appelle estimation du maximum de vraisemblance la valeur de θ qui maximise cette fonction, c'est à dire

$$t = h(x_1, ..., x_n) = argmax_{\theta \in \Theta} L(\theta; x_1, ..., x_n).$$

*L'**estimateur du maximum de vraisemblance (EMV)** est alors l'estimateur $T = h(X_1, ..., X_n)$. Dans le cadre de ce cours, on suppose que les variables aléatoires sont i.i.d. On a alors*

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} P_\theta[X_i = x_i].$$

*Cette fonction s'écrit comme un produit, et il est généralement plus simple de prendre le logarithme pour transformer ce produit en somme. On travaille alors avec la **fonction de log-vraisemblance***

$$l(\theta; x_1, ..., x_n) = ln(L(\theta; x_1, ..., x_n)) = \sum_{i=1}^{n} ln(P_\theta[X_i = x_i])$$

Revenons sur les deux premiers exemples du cours.

- **Modèle de Bernoulli :** $X_1, ..., X_n \sim_{iid} Ber(\theta)$. On a

$$\mathbb{P}_\theta(X_i = x_i) = \begin{cases} 1 - \theta \text{ si } x_i = 0 \\ \theta \text{ si } x_i = 1 \end{cases}$$

  Ceci se réécrit sous la forme

$$\mathbb{P}_\theta(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i} \text{ pour } x_i \in \{0, 1\}.$$

  La fonction de vraisemblance est donc donnée par

$$\begin{aligned} L(\theta; x_1, ..., x_n) &= \prod_{i=1}^{n} P_\theta(X_i = x_i) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

  On en déduit la fonction de log-vraisemblance

$$l(\theta; x_1, ..., x_n) = ln(\theta) \sum_{i=1}^{n} x_i + ln(1-\theta)(n - \sum_{i=1}^{n} x_i).$$

  Afin de trouver le maximum de cette fonction, on peut calculer la dérivée (par rapport à θ)

$$\frac{\partial l(\theta; x_1, ..., x_n)}{\partial \theta} = \frac{\sum_{i=1}^n x_i}{\theta(1-\theta)} - \frac{n}{1-\theta}.$$

  Puis en étudiant le signe de la dérivée, on en déduit que la fonction de vraisemblance atteint son maximum en $\frac{1}{n}\sum_{i=1}^n x_i$. L'EMV est obtenu en remplaçant les observations $(x_1, ..., x_n)$ par les variables aléatoires correspondantes dans l'expression précédente. On retrouve l'estimateur défini précédemment $F = \frac{1}{n}\sum_{i=1}^n X_i$.

- **Modèle de Poisson :** $X_1, ..., X_n \sim_{iid} Pois(\theta)$. On a $P_\theta[X_i = x_i] = exp(-\theta)\frac{\theta^{x_i}}{x_i!}$. On en déduit que

$$L(\theta; x_1, ..., x_n) = exp(-n\theta)\frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

  puis que

$$l(\theta; x_1, ..., x_n) = -n\theta + ln(\theta) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} ln(x_i!).$$

En dérivant cette fonction, on obtient

$$\frac{\partial l(\theta; x_1, ..., x_n)}{\partial \theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i.$$

En étudiant le signe de la dérivée, on peut vérifier que l'EMV de θ est $\bar{X}$.

La vraisemblance se définit de manière analogue dans le cas des variables aléatoires continues à partir de la densité.

Définition 7 *On suppose que $(X_1, ..., X_n) \sim_{iid} \mathbb{P}_\theta$ avec $\mathbb{P}_\theta$ un loi qui admet une densité $f_\theta(x)$ par rapport à la mesure de Lebesgue. On appelle alors* ***fonction de vraisemblance*** *la fonction de θ définie par*

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

On appelle ***fonction de log-vraisemblance*** *la quantité*

$$l(\theta; x_1, ..., x_n) = ln(L(\theta; x_1, ..., x_n)).$$

et estimateur du maximum de vraisemblance

$$T = argmax_{\theta \in \Theta} L(\theta; X_1, ..., X_n).$$

Revenons sur le **modèle gaussien**. On suppose alors $X_1, ..., X_n \sim_{iid} \mathcal{N}(\mu, \sigma^2)$ et la densité de la loi $\mathcal{N}(\mu, \sigma^2)$ est donnée par

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

La fonction de vraisemblance est donc

$$L(\mu, \sigma; x_1, ..., x_n) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} exp(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}).$$

On en déduit que

$$l(\mu, \sigma; x_1, ..., x_n) = -nln(\sigma) - \frac{n}{2}ln(2\pi) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

puis les dérivées partielles

$$\frac{\partial l(\mu,\sigma;x_1,...,x_n)}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} \text{ et } \frac{\partial l(\mu,\sigma;x_1,...,x_n)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}.$$

En résolvant les équations $\frac{\partial l(\mu,\sigma;x_1,...,x_n)}{\partial \mu} = \frac{\partial l(\mu,\sigma;x_1,...,x_n)}{\partial \sigma} = 0$, on obtient que $(\bar{x}, s)$ est un point critique de la fonction de vraisemblance puis que $(\bar{X}, S)$ est l'estimateur du maximum de vraisemblance des paramètres.

2.6 Information de Fisher

Dans ce paragraphe, on suppose que $\Theta \subset \mathbb{R}$ (paramètre inconnu scalaire).

Définition 8 *On appelle* ***quantité d'information de Fisher*** *apportée par un n-échantillon sur le paramètre θ la quantité suivante (si elle existe)*

$$I_n(\theta) = E[\left(\frac{\partial l(\theta; X_1, ..., X_n)}{\partial \theta}\right)^2]$$

En pratique, il est souvent plus facile d'utiliser l'une des deux formules données dans la proposition suivante pour calculer l'information de Fisher.

Proposition 9 *Si l'information de Fisher existe, sous des conditions générales (cf remarque ci-dessous), on a*

$$I_n(\theta) = var(\Big(\frac{\partial l(\theta;X_1,...,X_n)}{\partial \theta}\Big)) \text{ et } I_n(\theta) = -E[\Big(\frac{\partial^2 l(\theta;X_1,...,X_n)}{\partial \theta^2}\Big)].$$

Remarque 5

1. *On admettra que la proposition précédente s'applique pour les lois usuelles dès que le support de la loi, c'est à dire l'ensemble $A_\theta = \{x | f_\theta(x) > 0\}$, ne dépend pas de θ. Un exemple classique pour lequel le support de la loi dépend de θ est le cas où les X_i suivent une loi uniforme sur $[0, \theta]$. On vérifie alors que la fonction de vraisemblance n'est pas dérivable et que la proposition précédente ne s'applique pas (cf TD?).*

2. *En utilisant la définition de la log-vraisemblance dans le cas des échantillons i.i.d., il est facile de vérifier que $l(\theta; x_1, ..., x_n) = \sum_{i=1}^n l(\theta; x_i)$. On en déduit aisément que, si la proposition précédente s'applique, alors $I_n(\theta) = nI_1(\theta)$.*

3. *L'information de Fisher peut s'interpréter comme la quantité d'information apportée par l'échantillon pour estimer le paramètre inconnu. On a $I_n(\theta) \geq 0$. Si $I_n(\theta) = 0$ alors $f_\theta(x) = f(x)$ et la loi des observations ne dépend pas du paramètre θ. Plus $I_n(\theta)$ est grand, plus il est "facile" d'identifier le paramètre inconnu.*

Prenons l'exemple du **modèle de Bernoulli** $(X_1, ..., X_n) \sim_{iid} Ber(\theta)$ et vérifions que la proposition précédente s'applique. On a vu que

$$l(\theta; X_1, ..., X_n) = ln(\theta) \sum_{i=1}^n X_i + ln(1-\theta)(n - \sum_{i=1}^n X_i)$$

donc

$$\begin{aligned} \frac{\partial l(\theta; X_1, ..., X_n)}{\partial \theta} &= \frac{\sum_{i=1}^n X_i}{v} + \frac{\sum_{i=1}^n X_i - n}{1-\theta} \\ &= \frac{\sum_{i=1}^n X_i}{\theta(1-\theta)} - \frac{n}{1-\theta}. \end{aligned}$$

On en déduit que $E[\Big(\frac{\partial l(\theta;X_1,...,X_n)}{\partial \theta}\Big)] = 0$ et donc que $E[\Big(\frac{\partial l(\theta;X_1,...,X_n)}{\partial \theta}\Big)] = var(\Big(\frac{\partial l(\theta;X_1,...,X_n)}{\partial \theta}\Big))$, puis que l'information de Fisher est donnée par

$$\begin{aligned} I_n(\theta) &= var(\frac{\sum_{i=1}^n X_i}{\theta(1-\theta)} - \frac{n}{1-\theta}) \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

En dérivant un seconde fois, on obtient

$$\frac{\partial^2}{\partial \theta^2} l(\theta; X_1, ..., X_n) = \sum_{i=1}^n X_i \frac{1-2\theta}{\theta^2(1-\theta)^2} + \frac{n}{(1-\theta)^2}$$

puis

$$E[\frac{\partial^2}{\partial \theta^2} l(\theta; X_1, ..., X_n)] = -\frac{n}{\theta(1-\theta)}.$$

On retrouve bien que $I_n(\theta) = -E[\Big(\frac{\partial^2 l(\theta;X_1,...,X_n)}{\partial \theta^2}\Big)]$. On peut aussi vérifier que $I_n(\theta) = nI_1(\theta)$.

On peut vérifier que la proposition précédente s'applique également aux deux autres exemples du cours.

Le théorème suivant est fondamental en théorie de l'estimation.

Théorème 1 ***(Borne de Fréchet-Darmois-Cramer-Rao (FDCR))***
Sous des conditions générales, si T est une estimateur sans biais de θ alors

$$var(T) \geq \frac{1}{I_n(\theta)}$$

Preuve. On admettra que la proposition précédente s'applique pour les lois usuelles dès que le support de la loi ne dépend pas de θ. ■

Le théorème de FDCR donne une borne inférieure pour la variance d'un estimateur sans biais. Plus la quantité d'information apportée par l'échantillon est grande, plus la borne de FDCR est petite.

Définition 9 *On dira qu'un estimateur est* ***efficace*** *si il est* sans biais *et que sa* variance est égale à la borne de FDCR*.*

Par ailleurs, si il existe un estimateur efficace, alors il est unique p.s. En effet, soient T_1 et T_2 deux estimateurs efficaces de θ. T_1 et T_2 sont donc sans biais et leurs variances sont égales à la borne de FDCR V. Considérons l'estimateur $T_3 = \frac{T_1+T_2}{2}$. T_3 est un estimateur sans biais de θ de variance $var(T_3) = \frac{V}{2}(1 + cor(T_1, T_2))$. Comme $var(T_3) \geq V$, on en déduit que $cor(T_1, T_2) = 1$ puis que $T_1 = T_2$ *p.s.*.

Vérifions si les estimateurs introduits pour les exemples du cours sont efficaces.

- **Modèle de Bernoulli** $(X_1, ..., X_n) \sim_{iid} Ber(\theta)$. L'information de Fisher est donnée par

$$I_n(\theta) = \frac{n}{\theta(1-\theta)}.$$

  Le théorème de FDCR nous dit tout estimateur *sans biais* aura une variance supérieure à $\frac{\theta(1-\theta)}{n}$. Or, nous avons vu que $F = \frac{1}{n}(X_1 + ... + X_n)$ est un estimateur sans biais de θ et que sa variance est égale $I_n(\theta)^{-1}$. On en déduit qu'il s'agit de l'unique estimateur efficace de θ.

- **Modèle de Poisson** $X_1, ..., X_n \sim_{iid} Pois(\theta)$. On a vu que

$$\frac{\partial l(\theta; x_1, ..., x_n)}{\partial \theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i.$$

  On en déduit donc que

$$I_n(\theta) = var(\frac{1}{\theta}\sum_{i=1}^{n} X_i) = \frac{n}{\theta}$$

  puisque $var(X_i) = \theta$. Par ailleurs, on a vu que $\bar{X}_n$ est un estimateur sans biais de θ et que $var(\bar{X}_n) = \frac{\theta}{n} = 1/I_n(\mu)$. On en déduit que $\bar{X}_n$ est un estimateur efficace de θ.

- **Modèle gaussien** $(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$. On suppose que seul le paramètre μ est inconnu. On a vu que

$$\frac{\partial l(\mu, \sigma; x_1, ..., x_n)}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2}.$$

  On en déduit donc que

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

  Par ailleurs on a vu que $\bar{X}_n$ est un estimateur sans biais de μ tel que $var(\bar{X}_n) = \frac{\sigma^2}{n} = 1/I_n(\mu)$. On en déduit que $\bar{X}_n$ est un estimateur efficace de μ.

La proposition suivante établit que sous des conditions générales, l'EMV a de bonnes propriétés asymptotiques.

Proposition 10 *Sous des hypothèses générales (cf remarque ci-dessous), l'EMV est convergent et asymptotiquement gaussien et la variance asymptotique est donnée par l'inverse de l'information de Fisher*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{I_1(\theta)}).$$

Preuve. On admettra que la proposition précédente s'applique pour les lois usuelles dès que le support de la loi ne dépend pas de θ. ■

- **Modèle de Bernoulli** $(X_1, ..., X_n) \sim_{iid} Ber(\theta)$. En utilisant le théorème central limite, on a montré que

$$\sqrt{n}(F_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta)).$$

  Par ailleurs, on a vu que $I_1(\theta) = \frac{1}{\theta(1-\theta)}$ et le résultat s'applique bien.

- **Modèle de Poisson** $X_1, ..., X_n \sim_{iid} Pois(\theta)$. On a $E(X_i) = var(X_i) = \theta$, donc d'après le théorème central limite, on a

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta).$$

  Par ailleurs, on a vu que $I_1(\theta) = \frac{1}{\theta}$ et donc le résultat s'applique bien.

- **Modèle gaussien** $(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$. On suppose que seul le paramètre μ est inconnu. D'après le théorème central limite, on a

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

  Par ailleurs, on a vu que $I_1(\mu) = \frac{1}{\sigma^2}$ et donc le résultat s'applique bien.

Remarque 6
1. *On peut donc en déduire, sous certaines réserves, que pour n grand* $E[T_n] \approx \theta$ *et* $var(T_n) \approx \frac{1}{I_n(\theta)}$. *Un tel estimateur est dit "asymptotiquement efficace". Toutes ces bonnes propriétés (convergence, normalité asymptotique avec variance asymptotique connue, efficacité asymptotique) justifient l'utilisation de la méthode du maximum de vraisemblance comme méthode d'estimation par défaut pour les grands échantillons en statistique.*
2. *Pour que le théorème précédent s'applique, il faut pouvoir dériver la vraisemblance trois fois par rapport à* θ *(pour tout x), pouvoir intervertir les signes* ∂ *et* $\int$ *et que* Θ *soit un ensemble ouvert. Ces conditions sont généralement vérifiées lorsque le support de la loi ne dépend pas de* θ.

3 Estimation par intervalle de confiance

Dans les paragraphes précédents, des méthodes permettant d'estimer la valeur d'un paramètre inconnu θ à partir d'observations ont été proposées. Ces méthodes fournissent seulement une valeur ("**estimation ponctuelle**"), mais ne permettent pas de quantifier la précision de cette estimation. Pour cela, on utilise généralement des intervalles de confiance qui peuvent s'interpréter comme des marges d'erreur.

3.1 Construction d'intervalles de confiance pour la moyenne d'un échantillon Gaussien lorsque la variance est connue

Rappels sur la loi normale

- $Z \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$ si et seulement si sa densité est donnée par $f(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ pour $z \in \mathbb{R}$.
- Si $Z \sim \mathcal{N}(\mu, \sigma^2)$ alors $E[Z] = \mu$ et $var(Z) = \sigma^2$.
- Si $Z \sim \mathcal{N}(\mu, \sigma^2)$ alors $U = \frac{Z-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- On note u_α le quantile d'ordre α de la loi $\mathcal{N}(0, 1)$ de telle manière que si $U \sim \mathcal{N}(0, 1)$ alors $P(U \leq u_\alpha) = \alpha$. On a $u_\alpha = \Phi^{-1}(\alpha)$ avec $\Phi(x) = P(U \leq x)$ la fonction de répartition de la loi $\mathcal{N}(0, 1)$.
- On a $P(u_{\alpha/2} \leq U \leq u_{1-\alpha/2}) = 1 - \alpha$. Par symétrie de la loi $\mathcal{N}(0, 1)$, on a $u_{\alpha/2} = -u_{1-\alpha/2}$ et donc $P(-u_{1-\alpha/2} \leq U \leq u_{1-\alpha/2}) = 1 - \alpha$.
- De manière générale, si $Z \sim \mathcal{N}(\mu, \sigma^2)$ alors $P(\mu - u_{1-\alpha/2}\sigma \leq Z \leq \mu + u_{1-\alpha/2}\sigma) = 1 - \alpha$.
- En pratique les quantiles de la loi $\mathcal{N}(0, 1)$ peuvent être obtenus en utilisant des tables statistiques ou des logiciels adaptés (R, Matlab, SAS, Excel...). En particulier, on a $u_{0.975} \approx 1.96$ et donc

$$P(\mu - 1.96\sigma \leq Z \leq \mu + 1.96\sigma) \approx 0.95.$$

- **Stabilité de la loi normale par sommation.** Si $(Z_1, ..., Z_n)$ est une suite de variables gaussiennes **indépendantes** et $(\lambda_1, ..., \lambda_n) \in \mathbb{R}^n$, alors $\sum_{i=1}^{n} \lambda_i Z_i$ est une variable aléatoire gaussienne.

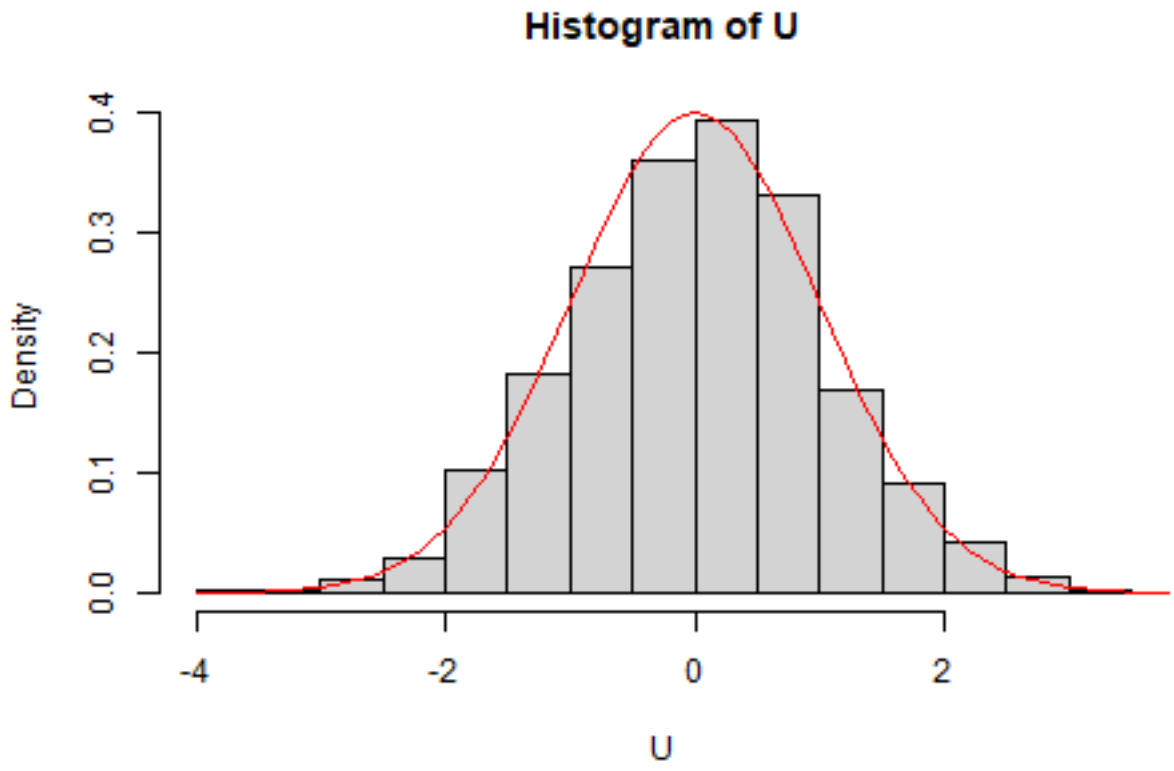
```
qnorm(0.975,mean=0,sd=1) #quantile à 97.5% de la loi N(0,1)
```

```
## [1] 1.959964
```

```
qnorm(0.995,mean=0,sd=1) #quantile à 99.5% de la loi N(0,1)
```

```
## [1] 2.575829
```

```
n=10^3
U=rnorm(10^3,mean=0,sd=1)  #simulation d'un échantillon gaussien
hist(U,freq=FALSE)  #freq=FALSE normalise l'histogramme pour que ce soit une densité (aire des rectangles é
xx=seq(-4,4,by=.1)
lines(xx,dnorm(xx,mean=0,sd=1),type='l',col='red')  #densité proche de l'histogramme
```




```
q=qnorm(0.975,mean=0,sd=1) #quantile à 97.5% de la loi N(0,1)
sum(U<q)/length(U) #environ 97.5% des simulations inférieures à q
```

```
## [1] 0.969
```

```
sum(U<q & U>-q)/length(U)  #environ 97.5% dans l'intervalle [-q,q]
```

```
## [1] 0.941
```

On se place sous les hypothèses du **modèle gaussien** et on suppose donc que $(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$. On cherche à estimer μ, supposé inconnu, mais on suppose que l'écart-type σ est connu. Ceci est rarement le cas en pratique, et ce cas particulier a donc principalement un objectif pédagogique. Nous reviendrons sur la construction d'intervalles de confiance pour la moyenne d'un échantillon sous des hypothèses plus réalistes dans la suite de ce cours.

Avec les hypothèses ci-dessus, on peut montrer que $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ puis que $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ et donc

$$P[u_{\alpha/2} \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq u_{1-\alpha/2}] = 1-\alpha$$

avec u_α le quantile d'ordre α de la loi $\mathcal{N}(0,1)$, ce qui se récrit

$$P[\bar{X} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

L'intervalle $[\bar{X}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$ est un intervalle aléatoire (puisque les bornes dépendent des variables aléatoires $X_1, ..., X_n$) qui contient la vraie valeur du paramètre μ avec une probabilité $1-\alpha$. Un tel intervalle est appelé intervalle de confiance au niveau de confiance $1-\alpha$ pour μ.

Définition : l'intervalle aléatoire $[a(X_1, ..., X_n); b(X_1, ..., X_n)]$ est appelé **intervalle de confiance** au niveau de confiance $1-\alpha$ pour θ si $P[a(X_1, ..., X_n) \leq \theta \leq b(X_1, ..., X_n)]] = 1-\alpha$. L'intervalle $[a(x_1, ..., x_n); b(x_1, ..., x_n)]$ est appelé **fourchette d'estimation** au niveau de confiance $1-\alpha$.

```
moy=1.5  #vraie valeur de mu
sig=2  #vraie valeur de sigma
n=30  #taille échantillon
X=rnorm(n,mean=moy,sd=sig)  #simulation
mi=mean(X)-1.96*sig/sqrt(n) #borne inf de la fourchette d'estimation à 95%
ma=mean(X)+1.96*sig/sqrt(n) #borne sup de la fourchette d'estimation à 95%
c(mi,ma)  #fourchette d'estimation à 95% pour mu
```

```
## [1] 1.120041 2.551422
```

```
#vérifions si la vraie valeur est dans l'IC à 95% avec un proba de 95%
N=10^3
for (i in 2:N){
  X=rnorm(n,mean=moy,sd=sig)  #simulation
  mi[i]=mean(X)-1.96*sig/sqrt(n)
  ma[i]=mean(X)+1.96*sig/sqrt(n)
}
sum(moy>mi & moy<ma)/N  #compte le nombre de fois que la vraie valeur est dans l'intervalle
```

```
## [1] 0.948
```

```
#doit être proche de 95%
```

3.2 Construction d'intervalles de confiance asymptotique pour la moyenne d'un échantillon quelconque de grande taille

Dans ce paragraphe, on suppose que $(X_1, ..., X_n)$ un n-échantillon d'une loi vérifiant $var(X_i) = \sigma^2 < +\infty$ et on note $\mu = E[X_i]$.
Lorsque la taille de l'échantillon n est suffisamment grande, on peut construire des intervalles de confiance pour la moyenne μ en utilisant les propriétés asymptotiques de $\bar{X}$ et S^2. On a montré, en utilisant le lemme de Slutsky, que

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

et donc que pour n "grand"

$$\sqrt{n}\frac{\bar{X}-\mu}{S} \approx \mathcal{N}(0,1).$$

A noter que cette approximation est valable même si l'échantillon n'est pas gaussien. En pratique, on suppose généralement que cette approximation est valide dès que $n \geq 30$. On a alors :

$$P[u_{\alpha/2} \leq \sqrt{n}\frac{\bar{X}-\mu}{S} \leq u_{1-\alpha/2}] \approx 1-\alpha$$

puis

$$P[\bar{X} + u_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2}\frac{S}{\sqrt{n}}] \approx 1-\alpha$$

L'intervalle $[\bar{X}+u_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X}+u_{1-\alpha/2}\frac{S}{\sqrt{n}}]$ est appelé "**intervalle de confiance asymptotique**" au niveau de confiance $1-\alpha$ pour μ.

Remarque. L'intervalle de confiance est centré sur la moyenne empirique $\bar{X}$. La largeur de l'intervalle de confiance dépend de plusieurs facteurs

- plus le niveau de confiance $1-\alpha$ est proche de 1, plus l'intervalle est large;
- plus l'écart-type empirique S est grand, plus l'intervalle est large;
- plus le nombre d'observations n est grand, plus l'intervalle est étroit.

On reprend les **montants des sinistres Cat Nat** et on se place sous les hypothèses du modèle gaussien. On peut alors faire l'application numérique en utilisant R.

```
X=datasin$Montant  #Montant des sinsitres
mean(X)  #estimation ponctuelle
```

```
## [1] 608.7838
```

```
n=length(X)  #taille de l'échantillon n
n
```

```
## [1] 37
```

```
s=sqrt(n-1)/sqrt(n)*sd(X)  #estimation biaisée de sigma
s
```

```
## [1] 494.7131
```

```
mi=mean(X)-1.96*s/sqrt(n)  #borne inférieure de l'IC
ma=mean(X)+1.96*s/sqrt(n)   #borne supérieure de l'IC
c(mi,ma)  #affichage des résultats
```

```
## [1] 449.3763 768.1912
```

```
#Fourchette d'estimation asymptotique à 95% pour le montant moyen des sinistre
```

3.3 Construction d'intervalles de confiance pour une proportion

On se place sous les hypothèses du **modèle de Bernoulli** et on suppose donc que $(X_1, ..., X_n) \sim_{iid} Ber(\theta)$. Il est également possible de construire des intervalles de confiance pour la proportion θ lorsque n est grand. En utilisant le TCL et le lemme de Slutsky, on a montré que pour n "grand' '

$$\sqrt{n}\frac{F-\theta}{\sqrt{F(1-F)}} \approx N(0,1) \tag{1}$$

Si on suppose que cette approximation est valide, on en déduit que

$$P[u_{\alpha/2} \leq \sqrt{n}\frac{F-\theta}{\sqrt{F(1-F)}} \leq u_{1-\alpha/2}] \approx 1-\alpha$$

puis que

$$P[F + u_{\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}} \leq \theta \leq F + u_{1-\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}}] \approx 1-\alpha$$

Donc $[F+u_{\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}}; F+u_{1-\alpha/2}\frac{\sqrt{F(1-F)}}{\sqrt{n}}]$ est un intervalle de confiance asymptotique au niveau de confiance $1-\alpha$ pour θ.

Remarque 7 *En pratique, on suppose généralement que l'approximation (1) est valable dès que* $n\theta \geq 5$ *et* $n(1-\theta) \geq 5$. *Comme* θ *est inconnu en pratique, on vérifie a posteriori si les conditions sont vérifiées pour les bornes de l'intervalle de confiance, c'est à dire* $n(f + u_{\alpha/2}\frac{\sqrt{f(1-f)}}{\sqrt{n}}) \geq 5$ *et* $n(1 - f - u_{1-\alpha/2}\frac{\sqrt{f(1-f)}}{\sqrt{n}}) \geq 5$. *Si ces conditions ne sont pas vérifiées, il est possible de construire des intervalles de confiance en utilisant la loi exacte de* F *(on sait que* nF *suit une loi Binomiale).*

Pour **l'exemple du sondage**, une fourchette d'estimation (asymptotique) pour θ est $[0.45; 0.51]$ (cf code ci-dessous). On obtient donc une marge d'erreur d'environ 6%. On vérifie que les conditions $n(f - u_{\alpha/2}\frac{\sqrt{f(1-f)}}{\sqrt{n}}) \geq 5$ et $n(1 - f - u_{1-\alpha/2}\frac{\sqrt{f(1-f)}}{\sqrt{n}}) \geq 5$ sont (largement) vérifiées.

```
f=.48
n=1000
c(f-1.96*sqrt(f*(1-f))/sqrt(n),f+1.96*sqrt(f*(1-f))/sqrt(n))
```

```
## [1] 0.4490345 0.5109655
```

3.4 Autres modèles paramétriques

Prenons l'exemple du modèle de Poisson. On a vu que

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta)$$

lorsque $n \to \infty$ et on est à nouveau dans une situation où la variance asymptotique dépend de θ. Deux approches sont couramment utilisés pour construire un intervalle de confiance asymptotique dans cette situation.

- **Utilisation du lemme de Slutsky** : le lemme de Slutsky justifie le remplacement des paramètres inconnus dans l'expression de la variance asymptotique par des estimateurs convergents. Pour l'exemple du modèle de Poisson, on obtient

  $$\sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

  lorsque $n \to \infty$. En utilisant cette expression, on peut vérifier que $[\bar{X}_n - 1.96\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}; \bar{X}_n + 1.96\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}]$ est un intervalle de confiance asymptotique à 95% pour θ.

- **Stabilisation de la variance** : dans cette approche, on cherche une fonction g telle que

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

lorsque $n \to \infty$. En utilisant la delta-méthode, on a

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,(g'(\theta))^2\theta)$$

lorsque $n \to \infty$ et on doit donc prendre g de telle manière que $g'(\theta) = \frac{1}{\sqrt{\theta}}$, c'est à dire $g(\theta) = 2\sqrt{\theta}$. On a donc

$$2\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

et on en déduit que $[\left(\sqrt{\bar{X}_n} - \frac{\sqrt{1.96}}{2\sqrt{n}}\right)^2 ; \left(\sqrt{\bar{X}_n} + \frac{\sqrt{1.96}}{2\sqrt{n}}\right)^2]$ est un intervalle de confiance asymptotique à 95% pour θ.

En faisant un développement limité à l'ordre 1, on peut vérifier que les deux expressions sont équivalentes lorsque $n \to +\infty$. On peut utiliser ces résultats pour calculer une fourchette d'estimation pour le paramètre de la loi de Poisson sur l'exemple des données Cat Nat (en supposant que l'approximation asymptotique est raisonnable).

```
n=length(datasin$Nombre)
c(mean(datasin$Nombre)-1.96*sqrt(mean(datasin$Nombre))/sqrt(n),
  mean(datasin$Nombre)+1.96*sqrt(mean(datasin$Nombre))/sqrt(n))  #première formule
```

```
## [1] 2612.344 2645.386
```

```
c((sqrt(mean(datasin$Nombre))-sqrt(1.96)/2/sqrt(n))^2,
  (sqrt(mean(datasin$Nombre))+sqrt(1.96)/2/sqrt(n))^2)  #deuxième formule
```

```
## [1] 2617.077 2640.679
```

4 Tests paramétriques

4.1 Généralités sur les tests

Un test statistique permet de vérifier si certaines hypothèses faites sur la valeur des paramètres sont réalistes ou non. Plus précisément, dans le cadre de ce paragraphe, nous nous intéresserons à tester des hypothèses de la forme

$$H_0 : \theta = \theta_0 \text{ contre l'hypothèse alternative } H_1 : \theta \neq \theta_0$$

avec $\theta_0 \in \Theta$ une valeur fixée. On supposera donc que H_0 est une hypothèse simple, qui spécifie la valeur des paramètres, et que H_1 est l'hypothèse complémentaire à l'hypothèse nulle (c'est à dire que H_1 peut se réécrire sous la forme "H_0 fausse").

On distingue usuellement deux types d'erreurs.

- L'**erreur de première espèce** qui consiste à rejeter H_0 alors que H_0 est vraie. On appelle **risque de première espèce** α la probabilité de refuser H_0 alors que H_0 est vraie.
- L'**erreur de deuxième espèce** qui consiste à accepter H_0 alors que H_0 est fausse. On appelle **risque de deuxième espèce** β la probabilité de choisir H_0 alors que H_0 est fausse.

En pratique, on fixe généralement le risque de première espèce α (valeurs courantes : 10%, 5% ou 1%) et H_0 joue un rôle plus important que H_1. **Le but du test est de vérifier si l'hypothèse H_0 est crédible pour les données étudiées**.

Par contre, il est généralement impossible de calculer le risque de deuxième espèce si l'hypothèse H_1 est une hypothèse complexe de la forme $H_1 : \theta \neq \theta_0$. En effet, la loi de l'échantillon est généralement inconnue sous H_1. H_1 est choisie uniquement par défaut si H_0 n'est pas considérée comme crédible pour les données étudiées.

$P = 1 - \beta$ est appelé la **puissance du test**. Lorsque c'est possible, pour un risque de première espèce α fixé, il est naturel de chercher à construire le test dont la puissance est la plus grande.

Les différentes erreurs sont résumées dans le tableau 2.

Vérité \Décision	On accepte H_0	On refuse H_0
H_0 est vraie	Bonne décision, Probabilité $1-\alpha$	Mauvaise décision, Risque de première espèce α
H_0 est fausse	Risque de deuxième espèce β	Bonne décision, Probabilité $P = 1-\beta$

Table 2: Risques de première et deuxième espèce

4.2 Tests pour une moyenne

On dispose d'un n-échantillon $(X_1, ..., X_n)$ d'une loi d'espérance inconnue $\mu = E[X_i]$ et on veut tester l'hypothèse simple de la forme

$$H_0 : \mu = \mu_0 \text{ contre l'hypothèse alternative } H_1 : \mu \neq \mu_0$$

avec μ_0 une valeur fixée (par exemple $\mu_0 = 0$). En pratique, il semble naturel d'accepter H_0 si $\bar{X}$ est "suffisamment proche" de μ_0. Ceci est formalisé plus précisément ci-dessous.

Premier cas : supposons que $(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$ avec σ connu (cf paragraphe sur les intervalles de confiance). On a alors

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

Donc, si H_0 est vraie, on a $\mu = \mu_0$ et

$$\mathbb{P}_{H_0}[u_{\alpha/2} \leq \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \leq u_{1-\alpha/2}] = 1 - \alpha. \tag{2}$$

La notation $P_{H_0}(A)$ désigne la probabilité de l'évènement A lorsqu'on suppose que l'hypothèse H_0 est vraie. On adopte alors la **règle de décision** suivante :

- on accepte H_0 si $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} \in [u_{\alpha/2}, u_{1-\alpha/2}]$;
- on refuse H_0 sinon.

On accepte donc H_0 lorsque ,

$$\bar{X} \in [\mu_0 + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

c'est à dire lorsque la moyenne empirique $\bar{X}$ est "suffisamment proche" de μ_0. Cette règle de décision semble donc relativement intuitive. De plus, d'après (2), elle est construite de telle manière que la probabilité d'accepter H_0 alors que H_0 est vraie est égale à $1 - \alpha$: le risque de première espèce est donc bien égal à α.

Dans le vocabulaire des tests statistiques, on appelle

- $T = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ la **statistique de test** : c'est la statistique qui est utilisée pour prendre la décision.
- **zone d'acceptation** l'ensemble des valeurs de la statistique de test T pour lesquelles on accepte H_0, ici $[u_{\alpha/2}, u_{1-\alpha/2}[$.

- **zone de rejet** l'ensemble complémentaire des valeurs de la statistique de test T pour lesquelles on rejette H_0, ici $]-\infty, u_{\alpha/2}[\cup]u_{1-\alpha/2}, +\infty[$.

Exemple. Un client commande à son fournisseur un lot de thermomètres. Afin de tester la qualité des thermomètres, le client en choisit $n = 20$ au hasard et les plonge dans un liquide à 20 degrés. Il obtient les résultats suivants : 20.2, 20.4, 20.1, 19.8, 20.1, 20, 20.5, 20.2, 20.3, 20.1, 20.4, 20.6, 20, 19.9, 20.3, 20.4, 20.1, 20.1, 20.3, 20. On suppose que les températures données par le thermomètre suivent une loi normale 'espérance μ avec un écart-type $\sigma = 0.5$ (valeur donnée par le fournisseur). La moyenne empirique de l'échantillon est $\bar{x} = 20.19$. On va utiliser le test avec $\mu_0 = 20$ pour tester

$$H_0 : \mu = 20 \text{ contre } H_1 : \mu \neq 20.$$

Si H_0 est vraie, alors cela signifie que les thermomètres donnent la bonne température en moyenne. La statistique de test prend la valeur $t = \sqrt{n}\frac{\bar{x}-\mu}{\sigma} = \sqrt{20}\frac{20.19-20}{0.5} = 1.70$. Si on fait un test avec un risque de première espèce $\alpha = 5\%$, alors on accepte H_0 car $1.70 < u_{1-\alpha/2} = 1.96$. Par contre, si on prend $\alpha = 10\%$ alors on obtient $u_{1-\alpha/2} = 1.64$ et on refuse H_0 car $1.70 > u_{1-\alpha/2}$.

Lorsqu'on fait un test avec un logiciel de statistique (R, SAS, Excel, Python...), le résultat est donné sous la forme d'une "**p-value**" (ou "**degré de signification**"). Pour le test précédent, cette p-value est définie par

$$p_v = \mathbb{P}[|U| > |t|]$$

avec $U \sim \mathcal{N}(0, 1)$ et t la valeur prise par la statistique de test sur les données. On vérifie qu'on accepte H_0 avec un risque de première espèce α si et seulement si $p_v > \alpha$. La p-value est souvent interprétée comme une indication de la crédibilité de l'hypothèse H_0 : une p-value faible indique que l'hypothèse H_0 est peu crédible. Sur l'exemple des thermomètres, on obtient

$$p_v = P[|U| > |t|] = 2 * P[U > |t|] = 2 * P[U > 1.7] = 2 * (1 - P[U < 1.7])$$

avec $P[U < 1.7] = \phi(1.7)$ où ϕ désigne la fonction de répartition de la loi $\mathcal{N}(0, 1)$ qui est disponible dans les tables ou logiciels statistiques (fonction *pnorm* dans R). On obtient $p_v = 0.089$: on retrouve donc que H_0 est acceptée pour $\alpha = 5\%$ mais est refusée pour $\alpha = 10\%$ (la p-value est la valeur de α à partir de laquelle on refuse H_0).

Deuxième cas : on ne suppose plus que l'échantillon est gaussien ni que la variance σ^2 est connue. Par contre, on suppose que n est suffisamment grand ($n \geq 30$?) pour que l'approximation

$$\sqrt{n}\frac{\bar{X} - \mu}{S} \approx N(0, 1)$$

soit valable. Si H_0 est vraie, on a $\mu = \mu_0$ et

$$\mathbb{P}_{H_0}[u_{\alpha/2} \leq \sqrt{n}\frac{\bar{X} - \mu_0}{S} \leq u_{1-\alpha/2}] \approx 1 - \alpha$$

On adopte alors la règle de décision suivante:

- On accepte H_0 si $\sqrt{n}\frac{\bar{X}-\mu_0}{S} \in [u_{\alpha/2}, u_{1-\alpha/2}]$.
- On refuse H_0 sinon.

Ici la statistique de test est $T = \sqrt{n}\frac{\bar{X}-\mu}{S}$ et la région de rejet est $]-\infty, u_{\alpha/2}[\cup]u_{1-\alpha/2}, +\infty[$. La p-value du test est donnée par

$$p_v = \mathbb{P}[|U| > |t|]$$

avec $U \sim \mathcal{N}(0, 1)$ et t la valeur prise par la statistique de test sur les données.

Exemple. Afin de calculer une prime d'assurance, un actuaire part de l'hypothèse que l'espérance du montant annuel des sinistres Cat Nat est égal 800 millions d'euros. Afin de vérifier si c'est raisonnable, on peut réaliser le test de l'hypothèse

$$H_0 : \mu = 800 \text{ contre } H_1 : \mu \neq 800$$

avec μ l'espérance du montant des sinistres Cat Nat. La statistique de test prend la valeur $t = \sqrt{37}\frac{608.78-800}{510.37} = -2.39$ (cf code ci-dessous) et on refuse donc H_0 pour $\alpha = 5\%$ car $t < -1.96$. La p-value du test est égale à $p_v = 0.02$. On retrouve qu'on refuse H_0 pour $\alpha = 5\%$, mais par contre H_0 est acceptée pour $\alpha = 1\%$.

```
X=datasin$Montant  #Montant des sinsitres
t=sqrt(length(X))*(mean(X)-800)/sd(X)
t  #valeur prise par la statistique de test
```

```
## [1] -2.319117
```

```
2*(1-pnorm(abs(t)))  #p-value du test
```

```
## [1] 0.02038872
```

4.3 Test pour une proportion

On dispose d'un n-échantillon $(X_1, ..., X_n)$ d'une loi de Bernoulli de paramètre π inconnu, et on veut tester l'hypothèse simple

$$H_0 : \pi = \pi_0 \text{ contre l'hypothèse alternative } H_1 : \pi \neq \pi_0.$$

On a vu que pour n "grand" (cf paragraphe sur les intervalles de confiance, on suppose généralement que cette approximation est valable lorsque $n\pi \geq 5$ et $n(1-\pi) \geq 5$), on a

$$\sqrt{n}\frac{F-\pi}{\sqrt{\pi(1-\pi)}} \approx N(0,1)$$

Donc, si H_0 est vraie, on a $\pi = \pi_0$ et

$$\mathbb{P}_{H_0}[u_{\alpha/2} \leq \sqrt{n}\frac{F-\pi_0}{\sqrt{\pi_0(1-\pi_0)}} \leq u_{1-\alpha/2}] \approx 1-\alpha$$

On adopte alors la règle de décision suivante:

- On accepte H_0 si $\sqrt{n}\frac{F-\pi_0}{\sqrt{\pi_0(1-\pi_0)}} \in [u_{\alpha/2}, u_{1-\alpha/2}]$.
- On refuse H_0 sinon.

Ici la statistique de test est $T = \sqrt{n}\frac{F-\pi_0}{\sqrt{\pi_0(1-\pi_0)}}$ et la région de rejet est $]-\infty, u_{\alpha/2}[\cup]u_{1-\alpha/2}, +\infty[$. La p-value du test est donnée par

$$p_v = \mathbb{P}[|U| > |t|]$$

avec $U \sim \mathcal{N}(0,1)$ et t la valeur prise par la statistique de test sur les données.

Exemple. On reprend l'exemple du sondage. Etant donnés les résultats de ce sondage, peut-on exclure que les deux candidats sont à égalité? Pour répondre à cette question, on va réaliser un test de l'hypothèse

$$H_0 : \pi = \tfrac{1}{2} \text{ contre l'hypothèse alternative } H_1 : \pi \neq \tfrac{1}{2}$$

La statistique de test prend la valeur $t = \sqrt{1000}\frac{0.52-0.5}{\sqrt{0.5(1-0.5)}} = 1.26$. On accepte donc H_0 pour $\alpha = 5\%$. La p-value du test est égale $p_v = 0.21$. Le résultat du sondage ne permet pas d'écarter que les deux candidats sont à égalité.

5 Intervalles de confiance et tests pour les échantillons gaussiens

5.1 Résultat probabiliste

La proposition suivante est couramment utilisée pour faire des tests et des intervalles de confiance pour l'espérance et la variance d'échantillons gaussiens.

Proposition 11 *Soit* $(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$. *Alors*

1. $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
2. $n\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
3. $\bar{X}$ *et* S^2 *sont indépendantes*
4. $\sqrt{n-1}\frac{\bar{X}-\mu}{S} \sim \mathcal{T}_{n-1}$

Preuve.

On a déjà démontré 1.

On va démontrer les points 2., 3. et 4. en utilisant le théorème de Cochran vu dans le cours de modèle linéaire.

On note $X =^t (X_1, ..., X_n)$ (^tX est la transposée du vecteur X). X est un vecteur gaussien, et plus précisément $X \sim \mathcal{N}(\tilde{\mu}, \sigma^2 I_n)$ avec $\tilde{\mu} =^t (\mu, ..., \mu)$.

Posons $u = \frac{1}{\sqrt{n}}{}^t(1, ..., 1)$, notons E le sous espace engendré par u et $\pi_E(x)$ le projeté orthogonal du vecteur x sur E. On vérifie que u est un vecteur normé ($^tuu = \|u\|^2 = 1$) puis que

$$\pi_E(x) = (^tux)u =^t (\bar{x}, ..., \bar{x})$$

avec $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

On en déduit que

$$\pi_{E^\perp}(x) = x - \pi_E(x) =^t (x_1 - \bar{x}, ..., x_n - \bar{x}).$$

On vérifie en particulier que $\pi_{E^\perp}(\tilde{\mu}) = 0$ (attendu car $\tilde{\mu} \in E$).

D'après le théorème de Cochran, $\pi_E(X)$ et $\pi_{E^\perp}(X)$ sont des vecteurs aléatoires indépendants. On en déduit que $\bar{X}$ est indépendant de $S^2 = \frac{\|\pi_{E^\perp}(X)\|^2}{n}$ (preuve : $\bar{X} = (1, 0, ..., 0)\pi_E(X)$ est une fonction de $\pi_E(X)$ et $S^2 = \frac{\|\pi_{E^\perp}(X)\|^2}{n}$ est une fonction de $\pi_{E^\perp}(X)$).

Toujours d'après le théorème de Cochran, on a

$$\frac{\|\pi_{E^\perp}(X) - \pi_{E^\perp}(\tilde{\mu})\|^2}{\sigma^2} \propto \chi^2_{dim(E^\perp)}$$

avec $\pi_{E^\perp}(\tilde{\mu}) = 0$, $\|\pi_{E^\perp}(X)\|^2 = nS^2$, et $dim(E^\perp) = n - dim(E) = n - 1$.

D'après 1., on a $U = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. D'après 2. et 3., $n\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ et est indépendant de U. On en déduit 4. en utilisant la définition de la loi de Student. ∎

5.2 Intervalle de confiance et test pour l'espérance d'un échantillon gaussien

La proposition 11 permet en particulier de construire des intervalles de confiance et des tests pour l'espérance d'un échantillon gaussien de taille quelconque et de variance inconnue. Dans ce paragraphe, on suppose que

$$(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$$

avec μ et σ des paramètres inconnus (modèle gaussien). D'après la proposition précédente, on a

$$\sqrt{n-1}\frac{\bar{X} - \mu}{S} \sim \mathcal{T}_{n-1}$$

On peut utiliser ce résultat pour construire un intervalle de confiance pour μ. En effet, on en déduit que

$$P\left(t_{n-1,\alpha/2} \leq \sqrt{n-1}\frac{\bar{X}-\mu}{S} \leq t_{n-1,1-\alpha/2}\right) = 1-\alpha,$$

avec $t_{n,\alpha}$ le quantile d'ordre α de la loi T_n, puis que

$$P\left(\bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n-1}}\right) = 1-\alpha$$

Finalement, $\left[\bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n-1}}; \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n-1}}\right]$ est un intervalle de confiance au niveau de confiance $1-\alpha$ pour μ.

On peut également utiliser la proposition 11 pour construire un test sur l'espérance d'un échantillon gaussien. On veut tester

$$H_0 : \mu = \mu_0 \text{ contre l'hypothèse alternative } H_1 : \mu \neq \mu_0$$

avec μ_0 une valeur fixée. On utilise alors la statistique de test

$$T = \sqrt{n-1}\frac{\bar{X}-\mu_0}{S}$$

Si H_0 est vraie, alors $T \sim \mathcal{T}_{n-1}$ et

$$P_{H_0}[t_{n-1,\alpha/2} \leq T \leq t_{n-1,1-\alpha/2}] = 1-\alpha$$

On adopte alors la **règle de décision** suivante:

- on accepte H_0 si $T = \sqrt{n-1}\frac{\bar{X}-\mu_0}{S} \in [t_{n-1,\alpha/2}, t_{n-1,1-\alpha/2}]$;
- on refuse H_0 sinon.

La p-value du test est donnée par

$$p_v = \mathbb{P}[|U| > |t|]$$

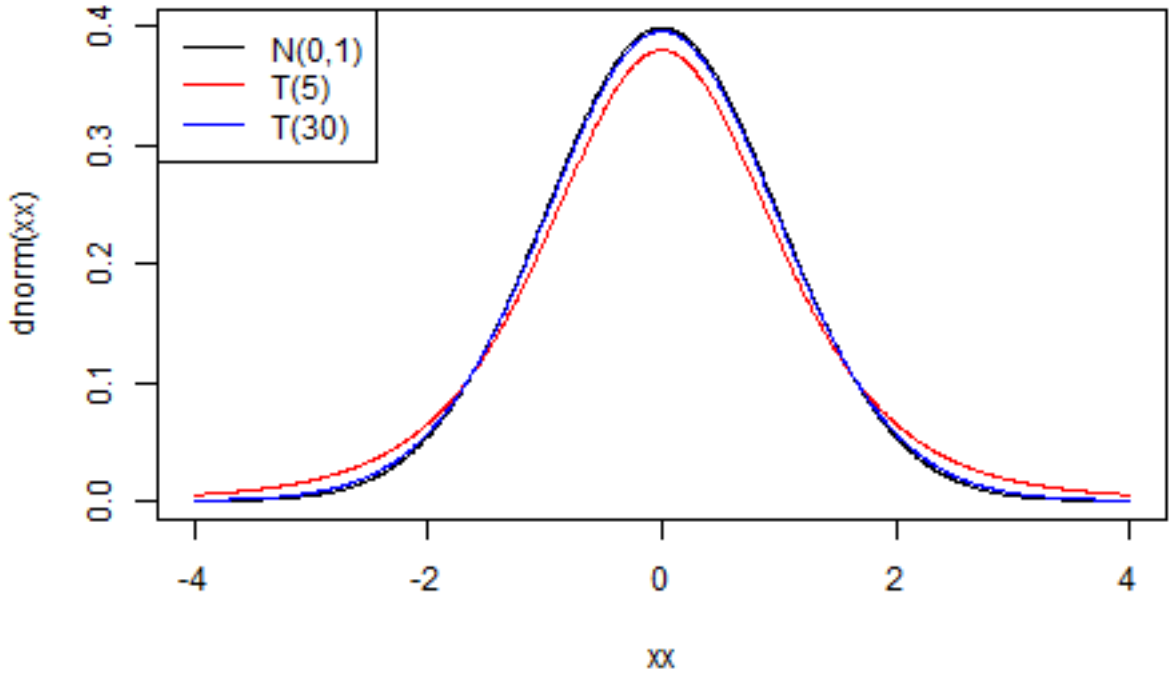
avec $U \sim \mathcal{T}_{n-1}$ et t la valeur prise par la statistique de test sur les données.

Remarque 8 *On retrouve des formules relativement similaires à celles des paragraphes 3.2 (pour les intervalles de confiance) et 4.2 (pour les tests) qui étaient basées sur l'approximation*

$$\sqrt{n}\frac{\bar{X}-\mu}{S} \approx \mathcal{N}(0,1)$$

qui est valide lorsque n est grand (même si l'échantillon $(X_1, ..., X_n)$ n'est pas gaussien). On peut vérifier que pour n "grand", la loi $\mathcal{T}_n$ est proche d'une loi $\mathcal{N}(0,1)$ (cf figure ci-dessous) et donc on obtient des résultats numériques similaires si on utilise les formules des paragraphes 3.2, 4.2 ou 5.2.

```
xx=seq(-4,4,by=.01)
plot(xx,dnorm(xx),type='l')  #densité loi normale
lines(xx,dt(xx,df=5),col='red')  #densité loi de Student à 10 ddl
lines(xx,dt(xx,df=30),col='blue')  #densité loi de Student à 30 ddl
legend('topleft', legend=c("N(0,1)", "T(5)" , "T(30)"),
       col=c("black", "red", "blue"), lty=1)
```




```
#la loi de Student à 30 ddl (bleu) est très proche de la loi N(0,1) (noir)


#calcul d'un IC à 95% pour le montant moyen des sinistres Cat Nat
X=datasin$Montant #Données
mean(X) #estimation ponctuelle de la moyenne
```

```
## [1] 608.7838
```

```
n=length(X)
#la fonction qt donne les quantiles de la lois de Student
mi=mean(X)+qt(0.025,df=n-1)*sd(X)/sqrt(n)  #borne inférieure de la fourchette d'estimation pour mu
ma=mean(X)+qt(0.975,df=n-1)*sd(X)/sqrt(n)   #borne supérieure de la fourchette d'estimation pour mu
c(mi,ma)  #affichage des résultats
```

```
## [1] 441.5630 776.0046
```

```
t.test(x=X,mu=800) #test hypothèse mu=800
```

```
##
##  One Sample t-test
##
## data:  X
## t = -2.3191, df = 36, p-value = 0.02618
## alternative hypothesis: true mean is not equal to 800
## 95 percent confidence interval:
##  441.5630 776.0046
```

```
## sample estimates:
## mean of x
##  608.7838
```

5.3 Intervalle de confiance et test pour la variance d'un échantillon gaussien

On peut également utiliser le résultat du paragraphe 5.1 pour construire des intervalles de confiance et des tests pour la variance d'un échantillon gaussien de taille quelconque et de moyenne inconnue. On suppose toujours dans ce paragraphe que

$$(X_1, ..., X_n) \sim_{iid} \mathcal{N}(\mu, \sigma^2)$$

avec μ et σ des paramètres inconnus. On a montré que

$$n\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

On peut utiliser ce résultat pour construire un intervalle de confiance pour σ^2. En effet, on en déduit que

$$P\left(\chi_{n-1,\alpha/2} \leq n\frac{S^2}{\sigma^2} \leq \chi_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

avec $\chi_{n,\alpha}$ le quantile d'ordre α de la loi χ^2_n puis que

$$P\left(n\frac{S^2}{\chi_{n-1,1-\alpha/2}} \leq \sigma^2 \leq n\frac{S^2}{\chi_{n-1,\alpha/2}}\right) = 1 - \alpha$$

Finalement, $\left[n\frac{S^2}{\chi_{n-1,1-\alpha/2}}; n\frac{S^2}{\chi_{n-1,\alpha/2}}\right]$ est un intervalle de confiance au niveau de confiance $1 - \alpha$ pour σ^2.

On peut aussi tester l'hypothèse

$$H_0 : \sigma = \sigma_0 \text{ contre l'hypothèse alternative } H_1 : \sigma \neq \sigma_0.$$

On utilise la statistique de test

$$X = n\frac{S^2}{\sigma_0^2}.$$

Si H_0 est vraie, alors $X \sim \chi^2_{n-1}$ et

$$P_{H_0}\left(\chi_{n-1,\alpha/2} \leq X \leq \chi_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

On adopte alors la règle de décision suivante:

- on accepte H_0 si $X = n\frac{S^2}{\sigma_0^2} \in [\chi_{n-1,\alpha/2}, \chi_{n-1,1-\alpha/2}]$;
- on refuse H_0 sinon.

Remarque 9 *Les résultats sont énoncés avec l'estimateur biaisé $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ mais on peut facilement les adapter pour l'estimateur sans biais*

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

Calculons un intervalle de confiance à 95% pour la variance des sinistres Cat Nat sous l'hypothèse que les sinistres proviennent d'un échantillon gaussien.

```r
var(X)  #estimation (sans biais) ponctuelle de la variance du montant des sinistres
```

```
## [1] 251539.4
```

```r
#la fonction qchisq donne les quantiles de la lois du chi2
mi=(n-1)*var(X)/qchisq(0.975,df=n-1) #borne inférieure de la fourchette d'estimation pour sigma
ma=(n-1)*var(X)/qchisq(0.025,df=n-1) #borne supérieure de la fourchette d'estimation pour sigma
c(mi,ma)  #affichage des résultats
```

```
## [1] 166345.9 424422.0
```

5.4 Test de comparaison de la variance de deux échantillons gaussiens indépendants

Définition 10 *On appelle loi de Fisher à m et n degrés de liberté la loi de*

$$F = \frac{U/m}{V/n}$$

avec $U \sim \chi^2_m$ et $V \sim \chi^2_n$ une v.a. indépendante de U. On notera $F \sim \mathcal{F}_{m,n}$ et $f_{m,n,\alpha}$ le quantile d'ordre α de la loi $F_{m,n}$.

Dans ce paragraphe, on suppose qu'on dispose de deux échantillons gaussiens indépendants

$$(X_1, ..., X_{n_X}) \sim_{iid} \mathcal{N}(\mu_X, \sigma^2_X)$$
$$(Y_1, ..., Y_{n_Y}) \sim_{iid} \mathcal{N}(\mu_Y, \sigma^2_Y)$$

On note $\bar{X}$ et $S^2_{X,corr}$ [resp. $\bar{Y}$ et $S^2_{Y,corr}$] la moyenne et la variance empirique (estimateur sans biais) de l'échantillon $(X_1, ..., X_{n_X})$ [resp. $(Y_1, ..., Y_{n_Y})$].

Remarque 10 *Ici on suppose que les échantillons $(X_1, ..., X_{n_X})$ et $(Y_1, ..., Y_{n_Y})$ sont **indépendants**, et donc que toutes les variables aléatoires sont indépendantes. Cette hypothèse n'est pas vérifiée pour les échantillons **appariés**. Par exemple, si on mesure une certaine quantité (poids par exemple) d'un patient avant et après un traitement : il n'est pas raisonnable de supposer que le poids après traitement est indépendant du poids avant traitement. Une manière de faire des tests pour comparer des échantillons appariés consiste à calculer les différences (la différence de poids avant/après traitement dans l'exemple ci-dessus); tester l'égalité des espérances revient alors à tester si on peut supposer que l'espérance des différences est nulle.*

On veut tester

$$H_0 : \sigma^2_X = \sigma^2_Y \text{ contre l'hypothèse alternative } H_1 : \sigma^2_X \neq \sigma^2_Y$$

On note $S_X^2 D'après la proposition 11, on a $(n_X - 1)\frac{S^2_{X,corr}}{\sigma^2_X} \sim \chi^2_{n_X-1}$ et $(n_Y - 1)\frac{S^2_{Y,corr}}{\sigma^2_Y} \sim \chi^2_{n_Y-1}$ et les deux variables aléatoires sont indépendantes car les deux échantillons sont indépendants. On en déduit donc, d'après la définition de la loi de Fisher, que

$$\frac{\sigma^2_Y}{\sigma^2_X}\frac{S^2_{X,corr}}{S^2_{Y,corr}} \sim \mathcal{F}_{n_X-1,n_Y-1}.$$

En particulier, si H_0 est vraie, on a $\frac{\sigma^2_Y}{\sigma^2_X} = 1$ puis $F = \frac{S^2_{X,corr}}{S^2_{Y,corr}} \sim \mathcal{F}_{n_X-1,n_Y-1}$. F est la statistique de test, et on accepte H_0 si et seulement si $F \in [f_{n_X-1,n_Y-1,\alpha/2}, f_{n_X-1,n_Y-1,1-\alpha/2}]$ avec $f_{n_X-1,n_Y-1,\alpha}$ le quantile d'ordre α de la loi F_{n_X-1,n_Y-1}.

Exemple. Le régime d'indemnisation Cat-Nat ayant été modifié à plusieurs reprises depuis sa création, un actuaire se demande si il est raisonnable de supposer que les montants des sinitres proviennent d'un échantillon identiquement distribué ou si la loi a évolué au cours du temps. Il se demande en particulier si il est raisonnable de supposer que la loi des sinistres est la même sur la période 1984-2003 que sur la période 2004-2020. On peut commencer par comparer les variances des deux échantillons en utilisant le test précédent.

```r
x1=X[datasin$An<2004]  #données avant 2004
x2=X[datasin$An>=2004]  #données après 2004
var(x1)/var(x2)  #statistique de test
```

```
## [1] 1.74361
```

```r
var.test(x1,x2)  #réalisation du test avec R
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 1.7436, num df = 19, denom df = 16, p-value = 0.2659
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6462532 4.5170885
## sample estimates:
## ratio of variances
##            1.74361
```

```r
#Conclusion : on accepte l'égalité des variances
```

5.5 Test de comparaison de la moyenne de deux échantillons gaussiens indépendants

On se place sous les mêmes hypothèses que dans le paragraphe précédent. On veut tester

$$H_0 : \mu_X = \mu_Y \text{ contre l'hypothèse alternative } H_1 : \mu_X \neq \mu_Y$$

Pour réaliser le test, on va supposer que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Cette hypothèse peut être vérifiée en utilisant le test du paragraphe précédent. D'après la proposition 11, on a $\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n_X})$ et $\bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y^2}{n_Y})$ et les deux variables aléatoires sont indépendantes. On en déduit donc que

$$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}).$$

Toujours d'après la proposition 11, on a $(n_X - 1)\frac{S_{X,corr}^2}{\sigma_X^2} \sim \chi_{n_X-1}^2$ et $(n_Y - 1)\frac{S_{Y,corr}^2}{\sigma_Y^2} \sim \chi_{n_Y-1}^2$ et les deux variables aléatoires sont indépendantes. On en déduit donc que

$$(n_X - 1)\frac{S_{X,corr}^2}{\sigma_X^2} + (n_Y - 1)\frac{S_{Y,corr}^2}{\sigma_Y^2} \sim \chi_{n_X+n_Y-2}^2.$$

A noter que comme on a supposé que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, l' expression précédente se simplifie et

$$(n_X - 1)\frac{S_{X,corr}^2}{\sigma^2} + (n_Y - 1)\frac{S_{Y,corr}^2}{\sigma^2} \sim \chi_{n_X+n_Y-2}^2.$$

Par ailleurs, on vérifie que les variable aléatoires $\bar{X} - \bar{Y}$ et $(n_X - 1)\frac{S_{X,corr}^2}{\sigma^2} + (n_Y - 1)\frac{S_{Y,corr}^2}{\sigma^2} \sim \chi_{n_X+n_Y-2}^2$ sont indépendantes ($\bar{X}$ est indépendante de S_X^2 d'après la proposition 11 et $\bar{X}$ est indépendant de S_Y^2 d'après l'hypothèse d'indépendance entre les échantillons, idem pour $\bar{Y}$).

Par ailleurs, si on suppose que H_0 est vraie, alors $\mu_X = \mu_Y$ et donc

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \sigma^2(\frac{1}{n_X} + \frac{1}{n_Y})\right).$$

ce qui implique que

$$\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim \mathcal{N}(0,1).$$

On en déduit finalement que, si H_0 est vraie alors

$$\begin{aligned} T &= \sqrt{n_X+n_Y-2}\frac{\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}}{\sqrt{(n_X-1)\frac{S^2_{X,corr}}{\sigma^2}+(n_Y-1)\frac{S^2_{Y,corr}}{\sigma^2}} \sim \chi^2_{n_X+n_Y-2}} \\ &= \frac{\sqrt{n_X+n_Y-2}}{\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}\frac{\bar{X}-\bar{Y}}{\sqrt{(n_X-1)S^2_{X,corr}+(n_Y-1)S^2_{Y,corr}}} \\ &\sim \mathcal{T}_{n_X+n_Y-2} \end{aligned}$$

Si on note $S^2_{X,Y,corr} = \frac{(n_X-1)S^2_{X,corr}+(n_Y-1)S^2_{Y,corr}}{n_X+n_Y-2}$ l'estimateur sans biais de σ^2, alors on peut réécrire

$$T = \frac{1}{\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}\frac{\bar{X}-\bar{Y}}{S_{X,Y,corr}}.$$

On adopte alors la règle de décision suivante : on accepte H_0 si et seulement si $T \in [t_{n_X+n_Y-2,\alpha/2}, t_{n_X+n_Y-2,1-\alpha/2}]$. La p-value du test est donnée par

$$p_v = \mathbb{P}[|U| > |t|]$$

avec $U \sim \mathcal{T}_{n_X+n_Y-2}$ et t la valeur prise par la statistique de test sur les données.

Exemple. Vérifions si il est raisonnable de supposer que l'espérance des sinistres est la même sur la période 1984-2003 que sur la période 2004-2020.

```
t.test(x1,x2,var.equal=TRUE)  #réalisation du test avec R
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 1.0468, df = 35, p-value = 0.3024
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -162.4781  508.3898
## sample estimates:
## mean of x mean of y
##  688.2500  515.2941
```

```
#Conclusion : on accepte H0
#Exercice : faire le test sans utiliser la fonction t.test
```

6 Tests du χ^2 pour les lois discrètes

6.1 Principe du test d'adéquation du χ^2.

Exemple. Afin de vérifier si un dé est équilibré, on réalise 90 lancers successifs d'un même dé. Les résultats sont reportés dans le tableau 3. Est-il raisonnable, à partir de cette expérience, de supposer que le dé est équilibré?

Dans ce paragraphe $(X_1, ..., X_n)$ désigne un échantillon à valeurs dans $\{1, ..., k\}$. On note, pour $j \in \{1, ..., k\}$,

- $\pi_j = P[X_i = j]$ la probabilité théorique d'obtenir la valeur j,

Face	1	2	3	4	5	6
Effectifs	9	16	23	10	13	19
Fréquences	0.10	0.18	0.26	0.11	0.14	0.21

Table 3: Résultats obtenus en lançant 90 fois un dé.

- $A_n(j) = card\{i \in \{1,...n\} | X_i = j\} = \sum_{i=1}^n \mathbb{1}(X_i = j)$ le nombre aléatoire de fois que la valeur j est présente dans l'échantillon aléatoire,
- $F_n(j) = \frac{A_n(j)}{n}$ la fréquence empirique d'apparition de la valeur j.

Le test d'adéquation du χ^2 permet de tester une hypothèse de la forme

$$H_0 : \forall j \in \{1, ..., k\}, \pi_j = \pi_{j,0} \text{ contre } H_1 : \exists j \in \{1, ..., k\} \ \pi_j \neq \pi_{j,0}$$

avec $\pi_{j,0}$ des valeurs fixées qui décrivent la loi sous H_0.

Le test du χ^2 est basé sur la proposition 12.

Proposition 12 *Sous les hypothèses ci-dessus, on a*

$$\sum_{j=1}^k \frac{(A_n(j)-n\pi_j)^2}{n\pi_j} \xrightarrow{\mathcal{L}} \chi^2_{k-1} \text{ lorsque } n \to +\infty.$$

Preuve. La preuve de ce résultat repose sur le théorème limite central multivarié et des arguments d'algèbre linéaire pour étudier la matrice de covariance asymptotique. Elle est admise dans le cadre de cours. On peut quand même remarquer que la statistique s'écrit comme la somme de k termes et chacun des termes suit une loi normale d'après le TCL. Ces termes ne sont pas indépendants puisque $\sum_{j=1}^k A_n(j) = n$. Il suffit donc d'en connaître $k-1$; c'est le d.d.l. de la loi du χ^2. ■

Exercice 6.1

1. *Calculer $E[A_n(j)]$ et $var(A_n(j))$.*
2. *Ecrire la loi des grands nombres et le théorème central limite pour la suite $Z_i = \mathbb{1}(X_i = j)$.*
3. *On suppose dans la suite que $k = 2$. Montrer que $\sum_{j=1}^k \frac{(A_n(j)-n\pi_j)^2}{n\pi_j} = \frac{(A_n(1)-n\pi_1)^2}{n\pi_1(1-\pi_1)}$.*
4. *En déduire que la proposition est vraie dans le cas particulier $k = 2$.*

La proposition précédente implique que si H_0 est vraie et n "grand" alors

$$D_n = \sum_{j=1}^k \frac{(A_n(j) - n\pi_{j,0})^2}{n\pi_{j,0}} = n \sum_{j=1}^k \frac{(F_n(j) - \pi_{j,0})^2}{\pi_{j,0}} \approx \chi^2_{k-1}.$$

C'est le point de départ pour réaliser le test du χ^2. La **statistique de test** D_n s'interprète comme une distance (on l'appelle généralement **la distance du χ^2**) entre les proportions observées $F_n(j) = \frac{A_n(j)}{n}$ et les proportions théoriques $\pi_{j,0}$ sous H_0. On s'attend donc à ce que D_n soit faible sous H_0.

Supposons que n est suffisamment grand pour que

$$P_{H_0}(D_n \leq \chi^2_{k-1,1-\alpha}) \approx 1 - \alpha.$$

On adopte alors la règle de décision suivante pour un risque de première espèce (asymptotique) α :

- on accepte H_0 si $D_n \leq \chi^2_{k-1,1-\alpha}$
- on refuse H_0 si $D_n > \chi^2_{k-1,1-\alpha}$

La p-value du test est $p_v = P(X > d_n)$ avec d_n la valeur observée pour la statistique de test sur l'échantillon et $X \sim \chi^2_{k-1}$.

Retour sur l'exemple du dé. On note π_j la probabilité théorique que le dé tombe sur la face j. Le dé est équilibré si l'hypothèse H_0 ci-dessous est vérifiée

$$H_0 : \pi_1 = ... = \pi_6 = 1/6 \text{ contre } H_1 : H_0 \text{ fausse}$$

On note $\pi_{j,0} = 1/6$ pour $j \in \{1, ..., 6\}$ les probabilités d'apparition sous H_0. On a $n = 90$, $k = 6$, et les effectifs $a_n(j)$ sont donnés dans le tableau 3. On peut présenter le calcul de la statistique du χ^2 sous la forme du tableau 4.

Face (j)	1	2	3	4	5	6	Total
Effectifs observés ($a_n(j)$)	9	16	23	10	13	19	90
Effectifs espérés ($n\pi_{j,0}$)	15	15	15	15	15	15	90
Contributions ($\frac{(a_n(j)-n\pi_{j,0})^2}{n\pi_{j,0}}$)	2.40	0.07	4.27	1.67	0.27	1.07	9.73

Table 4: Test du χ^2 sur l'exemple du dé.

On obtient donc la valeur suivante $d_n = 9.73$ pour la statistique de test. Choisissons le risque de première espèce $\alpha = 5\%$. Le quantile de la loi du χ^2 peut se lire dans une table ou être calculé en utilisant la commande *R qchisq* : on obtient $\chi^2_{k-1,1-\alpha} = \chi^2_{5,0.95} = 11.0705$. On a $d_n < \chi^2_{k-1,1-\alpha}$ donc on accepte H_0. La p-value du test est $p_v = P(X > 9.73)$ avec $X \sim \chi^2_{k-1}$. On peut utiliser R pour calculer cette valeur (cf code ci-dessous), et on obtient $p_v \approx 0.08$ On accepte donc H_0 pour $\alpha = 5\%$, mais on refuse H_0 pour $\alpha = 10\%$.

Utilisation de R. Les commandes suivantes permettent de réaliser le test avec *R* sur l'exemple du dé.

```
Nobs=c(9,16,23,10,13,19) #effectifs observés
p=rep(1,6)/6 #probabilités sous H0
chisq.test(Nobs,p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Nobs
## X-squared = 9.7333, df = 5, p-value = 0.08315
```

Validité asymptotique. Le test repose sur un argument asymptotique et est donc valide lorsque n est "grand". En pratique on admet généralement que l'approximation par la loi du χ^2 est bonne lorsque les effectifs espérés sont supérieurs à 5 (c'est à dire $n\pi_{j,0} \geq 5$ pour $j \in \{1, ..., k\}$). Lorsque cette condition n'est pas vérifiée, on peut soit regrouper des classes pour augmenter les effectifs espérés soit utiliser des simulations (méthode de Monte Carlo) pour approcher la loi de D_n sous H_0. Cette dernière solution est disponible sous R avec la commande suivante. On voit que la p-value est légérement modifiée par rapport au test basé sur la loi du χ^2.

```
chisq.test(Nobs,p=p,simulate.p.value=TRUE)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  Nobs
## X-squared = 9.7333, df = NA, p-value = 0.08246
```

Prise en compte de paramètres estimés. On cherche souvent à tester l'adéquation à une loi qui dépend de paramètres inconnus (par exemple une loi de Poisson ou une loi binomiale). Si on estime les paramètres par maximum de vraisemblance, alors le degré de liberté de la loi asymptotique devient $k - l - 1$ avec l le nombre de paramètres à estimer (par exemple, $l = 1$ pour une loi de Poisson).

6.2 Test d'homogénéité du χ^2

On peut étendre le test précédent à la comparaison de plusieurs échantillons décrits par une variable qualitative.

Exemple. Les résultats de l'évolution d'une maladie sur 1000 personnes ayant suivi l'un ou l'autre des traitements A et B sont résumés dans le tableau 5. Peut-on conclure de cette expérience que les traitements A et B ont des effets différents?

	Guérison	Amélioration	Stationnaire	Totaux
Traitement A	280	210	110	600
Traitement B	220	90	90	400
Totaux	500	300	200	

Table 5: Evolution d'une maladie sur 1000 personnes.

Dans ce paragraphe, on suppose que les observations proviennent de p échantillons **indépendants** de tailles respectives $n_1, ..., n_p$ et à valeurs dans le même ensemble $\{1, ..., k\}$. Pour $i \in \{1, ..., p\}$ et $j = \{1, ..., k\}$, $A_n(i, j)$ désignera le nombre (aléatoire) d'individus du i^{eme} échantillon qui prennent la valeur j. Nous noterons également $n = \sum_{i=1}^{p} n_i$ le nombre total d'observations et $A_n(j) = \sum i = 1^p A_n(i, j)$ le nombre total d'individus qui prennent la valeur j. On dispose donc d'un tableau de la forme du tableau 6.

	Classe 1	Classe 2	...	Classe k	Total
Echantillon 1	$A_n(1,1)$	$A_n(1,2)$	...	$A_n(1,k)$	n_1
Echantillon 2	$A_n(2,1)$	$A_n(2,2)$	...	$A_n(2,k)$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Echantillon p	$A_n(p,1)$	$A_n(p,2)$	...	$A_n(p,k)$	n_p
Total	$A_n(1)$	$A_n(2)$	...	$A_n(k)$	n

Table 6: Test d'homogénéité du χ^2

On souhaite tester l'hypothèse

$$H_0 \text{ : } \textit{les p échantillons suivent la même loi}$$

contre l'hypothèse alternative

$$H_1 \text{ : } \textit{les p échantillons ne suivent pas la même loi}$$

Sous l'hypothèse H_0, nous notons π_i la probabilité associée à la modalité $i \in \{1, ..., k\}$. D'après la proposition 12, pour $i \in \{1, ...p\}$,

$$\sum_{j=1}^{k} \frac{(A_n(i,j) - n_i\pi_j)^2}{n_i\pi_j} \approx \chi^2_{k-1}$$

si n_i est "grand". Les p échantillons étant indépendants, on en déduit que

$$\sum_{i=1}^{p}\sum_{j=1}^{k} \frac{(A_n(i,j) - n_i\pi_j)^2}{n_i\pi_j} \approx \chi^2_{p(k-1)}.$$

Comme on ne connaît pas $\pi_1, ..., \pi_k$, on estime ces paramètres par $F_j = \frac{\sum_{i=1}^{p} A_n(i,j)}{n} = \frac{A_n(j)}{n}$ la fréquence empirique de la modalité j dans la réunion des p échantillons. Finalement, la statistique du test est

$$D_n = \sum_{i=1}^{p}\sum_{j=1}^{k} \frac{(A_n(i,j) - n_iF_j)^2}{n_iF_j}$$

On a estimé $k-1$ paramètres (puisque $\pi_1+...+\pi_k=1$) et on en "déduit" que pour n grand $D_n \approx \chi^2_{(p-1)(k-1)}$ puisque $p(k-1)-(k-1)=(p-1)(k-1)$. On accepte finalement H_0 si et seulement si $D_n \leq \chi^2_{(p-1)(k-1),1-\alpha}$.

Utilisation de R. Les codes R ci-dessous réalisent le test sur l'exemple introduit au début du paragraphe. L'hypothèse H_0 est refusée : l'évolution du la maladie n'est pas la même pour les patients qui suivent le traitement A et le traitement B. **Exercice :** vérifiez que vous retrouver les valeurs numériques données par R en réalisant le test "à la main".

```
tab=matrix(c(280,220,210,90,110,90),nrow=2)
tab
```

```
##      [,1] [,2] [,3]
## [1,]  280  210  110
## [2,]  220   90   90
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 17.917, df = 2, p-value = 0.0001287
```

Remarque. Lorsque $p=k=2$, le test du χ^2 permet de comparer les probabilités dans deux échantillons de Bernoulli indépendants.

6.3 Test du χ^2 d'indépendance de deux variables

On peut également utiliser le test du χ^2 pour vérifier l'indépendance de deux variables aléatoires.

Exemple. On veut savoir si le temps écoulé depuis la vaccination contre une maladie donnée a une influence sur le degré de gravité de la maladie lorsqu'elle apparaît. Pour simplifier, nous ne distinguons que trois degrés de gravité. Parmi les malades, nous comparons les vaccinées depuis moins de 25 ans et ceux vaccinés depuis plus de 25 ans :

Degré de gravité	*Légère*	*Moyenne*	*Forte*	*Total*
vaccin <25 *ans*	43	324	347	714
vaccin >25 *ans*	120	230	510	860
Total	163	554	857	1574

Existe-t-il une dépendance entre la date de vaccination et le degré de gravité de la maladie?

Dans ce paragraphe, on suppose avoir un échantillon de taille n pour lequel on observe 2 variables qualitatives pour chaque individu

- $(X_1,...,X_n)$ à valeurs dans $\{1,...,k_1\}$
- $(Y_1,...,Y_n)$ à valeurs dans $\{1,...,k_2\}$

Notons

$$A_n(i,j)=card\{l\in\{1,...,n\}|X_l=i \text{ et } Y_l=j\}$$

le nombre de fois que la modalité (i,j) est observée pour $i=1,...,k_1$ et $j=1,...,k_2$ et $n=\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}A_n(i,j)$. On dispose donc d'un **tableau de contingence** de la forme du tableau 7.

On souhaite tester **l'hypothèse**

$$H_0 : \textit{les 2 variables sont indépendantes}$$

contre l'hypothèse alternative

	$j=1$	j=2	...	$j=k_2$	Total
$i=1$	$A_n(1,1)$	$A_n(1,2)$	$\dots$	$A_n(1,k_2)$	$A_n(1,.)$
$i=2$	$A_n(2,1)$	$A_n(2,2)$	$\dots$	$A_n(2,k_2)$	$A_n(2,.)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=k_1$	$A_n(k_1,1)$	$A_n(k_1,2)$	$\dots$	$A_n(k_1,k_2)$	$A_n(k_1,.)$
Total	$A_n(.,1)$	$A_n(.,2)$	$\dots$	$A_n(.,k_2)$	n

Table 7: Test d'indépendance du χ^2

H_1 : les 2 variables ne sont pas indépendantes.

Par définition, l'hypothèse H_0 est vérifiée si et seulement si

$$P[X_l = i, Y_l = j] = P[X_l = i]P[Y_l = j]$$

$\forall (i,j) \in \{1,...,k_1\} \times \{1,...,k_2\}$. Notons alors $\pi_i = P[X_l = i]$, $\pi'_j = P[Y_l = j]$ et $\pi_{i,j} = P[X_l = i, Y_l = j]$ pour $i = 1,...,k_1$ et $j = 1,...,k_2$.

D'après la proposition 12,

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(A_n(i,j) - n\pi_{i,j})^2}{n\pi_{i,j}} \approx \chi^2_{k-1}$$

pour n "grand" avec $k = k_1k_2$.

Sous l'hypothèse H_0, on a donc

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(A_n(i,j) - n\pi_i\pi'_j)^2}{n\pi_i\pi'_j} \approx \chi^2_{k-1}$$

pour n "grand".

On estime

- π_i par $F_n(i) = A_n(i,.)/n$ avec $A_n(i,.) = \sum_{j=1}^{k_2} A_n(i,j) = card\{l \in \{1,...,n\} | X_l = i\}$
- π'_j par $F'_n(j) = A_n(.,j)/n$ avec $A_n(.,j) = \sum_{i=1}^{k_1} A_n(i,j) = card\{l \in \{1,...,n\} | Y_l = j\}$

La statistique du test est

$$D_n = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(A_n(i,j) - \frac{A_n(i,.)A_n(.,j)}{n})^2}{\frac{A_n(i,.)A_n(.,j)}{n}}$$

On a estimé $k_1 + k_2 - 2$ paramètres (puisque $\pi_1 + ... + \pi_{k_1} = \pi'_1 + ... + \pi'_{k_2} = 1$) et on en "déduit" que pour n grand $D_n \approx \chi^2_{(k_1-1)(k_2-1)}$ puisque $k - 1 - (k_1 + k_2 - 2) = (k_1 - 1)(k_2 - 1)$.

On vérifie aisément que

$$D_n = n\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{(F_n(i,j) - F_n(i)F'_n(j))^2}{F_n(i)F'_n(j)}$$

avec $F_n(i,j) = \frac{A_n(i,j)}{n}$. D_n s'interprète donc comme une distance entre les fréquences observées et celles attendues sous l'hypothèse d'indépendance.

La règle de décision est la suivante : on accepte H_0 si et seulement si $D_n \leq \chi^2_{(k_1-1)(k_2-1),1-\alpha}$.

Utilisation de R.

Les commandes R ci-dessous permettent de réaliser le test d'indépendance du χ^2 sur l'exemple donné au début du paragraphe. Conclusion : on ne peut pas supposer que la date de vaccination et la gravité de la maladie sont indépendantes (p-value très faible). Exercice : vérifier que vous retrouver les valeurs numériques données par R en réalisant le test "à la main".

```r
tab=matrix(c(43,120,324,230,347,510),nrow=2)
tab
```

```
##      [,1] [,2] [,3]
## [1,]   43  324  347
## [2,]  120  230  510
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 70.389, df = 2, p-value = 5.19e-16
```

Remarque. On peut vérifier que la statistique de test est la même pour le test d'homogénéité et le test d'indépendance alors que le cadre est différent (plusieurs échantillons indépendants dans le premier cas, un seul échantillon avec deux variables dans le deuxième cas).