# [Initiez-vous à la statistique inférentielle](https://openclassrooms.com/fr/courses/4525306-initiez-vous-a-la-statistique-inferentielle)

Table des matières

[Initiez-vous à la statistique inférentielle 1](#_Toc57986722)

[Familiarisez-vous avec deux cas pratiques 5](#_Toc57986723)

[Taux de guérison suite à un nouveau traitement 5](#_Toc57986724)

[Consommation d’essence de cars 5](#_Toc57986725)

[Adoptez la posture du Data Analyst 7](#_Toc57986726)

[Comprenez les enjeux de l’inférence 8](#_Toc57986727)

[Déterminez votre modèle probabiliste 10](#_Toc57986728)

[Quels modèles pour nos exemples introductifs ? 10](#_Toc57986729)

[Le coin R : consommation d’essence 10](#_Toc57986730)

[Découvrez la notion d'estimateur 15](#_Toc57986731)

[Estimez une proportion 16](#_Toc57986732)

[Cas du taux de guérison 16](#_Toc57986733)

[Cas général 16](#_Toc57986734)

[Estimez une moyenne et une variance 17](#_Toc57986735)

[Cas de la consommation d'essence 17](#_Toc57986736)

[Cas général 17](#_Toc57986737)

[Comprenez ce qui fait la qualité d’un estimateur 19](#_Toc57986738)

[**Qualités d'un estimateur** 19](#_Toc57986739)

[**Cas d’une proportion, d’une moyenne (théorique) et d’une variance (théorique)** 20](#_Toc57986740)

[Déterminez la qualité de votre estimateur 21](#_Toc57986741)

[Mathématiquement 21](#_Toc57986742)

[Retour au taux de guérison 22](#_Toc57986743)

[Retour à la consommation d'essence 23](#_Toc57986744)

[Allez plus loin : méthodes des moments et du maximum de vraisemblance 25](#_Toc57986745)

[Méthode des moments 25](#_Toc57986746)

[Estimateur du maximum de vraisemblance 25](#_Toc57986747)

[Application à la loi exponentielle 26](#_Toc57986748)

[Testez-vos connaissances sur les estimateurs 29](#_Toc57986749)

[Compétences évaluées 29](#_Toc57986750)

[ Question 1 29](#_Toc57986751)

[ Question 2 29](#_Toc57986752)

[ Question 3 29](#_Toc57986753)

[ Question 4 30](#_Toc57986754)

[ Question 5 30](#_Toc57986755)

[Découvrez les intervalles de confiance 32](#_Toc57986756)

[L'idée 32](#_Toc57986757)

[Mathématiquement 32](#_Toc57986758)

[**Déterminez un intervalle de confiance sur une proportion** 33](#_Toc57986759)

[**Le coin méthodologique** 33](#_Toc57986760)

[**Le coin R : exemple du taux de guérison** 33](#_Toc57986761)

[Déterminez un intervalle de confiance sur une moyenne 36](#_Toc57986762)

[Le coin méthodologique 36](#_Toc57986763)

[Pour aller plus loin 36](#_Toc57986764)

[Le coin R : consommation d'essence 37](#_Toc57986765)

[Déterminez un intervalle de confiance sur une variance 39](#_Toc57986766)

[Le coin méthodologique 39](#_Toc57986767)

[Pour aller plus loin 39](#_Toc57986768)

[Le coin R : exemple de la consommation d'essence 39](#_Toc57986769)

[Testez vos connaissances sur les intervalles de confiance 42](#_Toc57986770)

[Compétences évaluées 42](#_Toc57986771)

[ Question 1 42](#_Toc57986772)

[ Question 2 42](#_Toc57986773)

[ Question 3 42](#_Toc57986774)

[ Question 4 43](#_Toc57986775)

[ Question 5 43](#_Toc57986776)

[Découvrez les tests statistiques 44](#_Toc57986777)

[Cas du taux de guérison 44](#_Toc57986778)

[Cas de la consommation d'essence 45](#_Toc57986779)

[Formalisez votre problème de test 46](#_Toc57986780)

[Les hypothèse d'un test 46](#_Toc57986781)

[La région critique 46](#_Toc57986782)

[Les risques associés aux décisions 46](#_Toc57986783)

[La p-valeur 47](#_Toc57986784)

[Testez une proportion 48](#_Toc57986785)

[Testez p=p0 versus p>p0 48](#_Toc57986786)

[Testez p=p0  versus p≠p0 51](#_Toc57986787)

[Le coin R : exemple du taux de guérison 52](#_Toc57986788)

[Testez une moyenne ou une variance 54](#_Toc57986789)

[Tester la moyenne (théorique) d'un échantillon 54](#_Toc57986790)

[Tester la variance (théorique) d'un échantillon 56](#_Toc57986791)

[Comparez deux échantillons gaussiens (test de comparaison) 59](#_Toc57986792)

[L'idée 59](#_Toc57986793)

[Le coin méthodologique 59](#_Toc57986794)

[Le coin R : exemple des iris de Fisher 63](#_Toc57986795)

[Pour aller plus loin 64](#_Toc57986796)

[Découvrez les tests d'adéquation : le Khi-deux et Kolmogorov Smirnov 65](#_Toc57986797)

[L'objectif 65](#_Toc57986798)

[Le test du Khi-deux 65](#_Toc57986799)

[Effectif 66](#_Toc57986800)

[Le test de Kolmogorov Smirnov 67](#_Toc57986801)

[Résumons ! 69](#_Toc57986802)

[Entraînez-vous à tester l'équiprobabilité des naissances de femmes et d'hommes 71](#_Toc57986803)

[À vous de jouer 71](#_Toc57986804)

[Savez-vous effectuer un test statistique ? 73](#_Toc57986805)

[Compétences évaluées 73](#_Toc57986806)

[ Question 1 73](#_Toc57986807)

[ Question 2 73](#_Toc57986808)

[ Question 3 73](#_Toc57986809)

[ Question 4 74](#_Toc57986810)

[ Question 5 74](#_Toc57986811)

[Conclusion 75](#_Toc57986812)

Imaginez-vous préparer un marathon : vos relevés aux entraînements indiquent un gain moyen de 10 minutes par rapport à ceux de l’an passé. Pouvez-vous conclure que vous avez vraiment progressé ? Ou cet écart est-il simplement dû à la chance ?

Pour trancher, vous pouvez compter sur la **statistique inférentielle**.

Ce cours établira le lien entre le monde réel, celui des observations, et le monde théorique des probabilités.

Vous apprendrez à **déterminer des marges d’erreur,**lorsque vous passez d’un échantillon  
à la population toute entière. Nous travaillerons également sur la notion de risque, inhérente à toute prise de décision.

Suivez ce cours, pour découvrir les notions d’**estimation**, d’**intervalle de confiance** et de **test statistique**. Promis, vous ne déciderez plus comme avant !

**Objectifs pédagogiques :**

* Comprendre la notion d’inférence.
* Etablir un lien entre des observations et un modèle probabiliste.
* Réaliser une estimation ponctuelle.
* Calculer un intervalle de confiance pour une proportion et une moyenne.
* Tester une hypothèse sur une proportion, une moyenne ou une variance.
* Tester l’adéquation à une loi par un test statistique.

**Prérequis :**

* Pratiquer les [**statistiques descriptives**](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees).
* Connaître les notions de base des [**probabilités**](https://openclassrooms.com/courses/4525296-maitrisez-les-bases-des-probabilites).

Ce cours fait partie d'un parcours de formation Data Analyst réalisé en collaboration avec l'ENSAE-ENSAI.

# Familiarisez-vous avec deux cas pratiques

Le code informatique dans ce cours est donné en langage R. Si vous utilisez Python, la traduction est donnée [**ici**](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525306/python_statistiques_inferentielles.html).

### Taux de guérison suite à un nouveau traitement

Considérons ce premier cas, qu’on peut qualifier de **discret** :

Un laboratoire cherche à savoir si la nouvelle composition du médicament (contre une maladie bénigne) qu’il compte commercialiser améliore le taux de guérison par rapport à un médicament déjà sur le marché.

Des tests cliniques ont été effectués sur **n=216**   individus sur lesquels on a observé la guérison (notée 1) ou la non-guérison (notée 0) :



**xi** désigne l’observation de la guérison ou non pour l’individu **i** .

Le laboratoire observe au total 167 guérisons, soit environ 77,3% de guérisons.

Le laboratoire s’adresse à un data analyst pour répondre à plusieurs de ses interrogations :

1. Il aimerait connaître le taux de guérison (théorique) **p** suite à la prise de son  
   médicament.
2. Le laboratoire étant conscient que le taux de guérison théorique sera délicat à  
   appréhender parfaitement, il souhaiterait disposer d’une fourchette de ce taux  
   de guérison.
3. Enfin, il aimerait vérifier que son nouveau médicament est (significativement) meilleur  
   que celui actuellement sur la marché dont le taux de guérison avéré est **p0=0.75**  
   (75%).

Notons que l’éventuelle autorisation de mise sur le marché de ce médicament repose sur une procédure plus complexe. Si cela vous intéresse, voyez : [**http://ansm.sante.fr/Activites/**](http://ansm.sante.fr/Activites/%20Autorisations-de-Mise-sur-le-Marche-AMM/L-AMM-et-le-parcours-du-medicament/)  
[**Autorisations-de-Mise-sur-le-Marche-AMM/L-AMM-et-le-parcours-du-medicament/**](http://ansm.sante.fr/Activites/%20Autorisations-de-Mise-sur-le-Marche-AMM/L-AMM-et-le-parcours-du-medicament/)

L'échantillon dont vous aurez besoin pour tester les lignes de codes est à télécharger ici : [**guerison.txt**](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525306/guerison.txt)

Le code dans ce cours est donné en langage R, mais une traduction en Python est disponible en haut de ce chapitre.

### Consommation d’essence de cars

Considérons ce second cas, qu’on peut qualifier de **continu, par opposition à "discret"**.

Un constructeur de cars souhaite appréhender la consommation d’essence de son dernier modèle. Pour cela, il lance un protocole d’essais sur 128 cars et recueille leur consommation d’essence en litres après avoir parcouru 100 km (appelée nombre de litres aux 100) :



Pour information, la consommation moyenne observée est égale à 31.45 litres aux 100.

Le constructeur s’adresse à un data analyst pour répondre à plusieurs de ses interrogations :

1. Il aimerait connaître la consommation (théorique) d’essence **μ** de son modèle de car.
2. De manière plus modeste, il souhaiterait également disposer d’un intervalle autour de cette consommation, il l’aura en effet évalué sur un nombre limité de véhicules et de trajets.
3. Enfin, il souhaiterait communiquer auprès de ses clients sur un chiffre ambitieux :  
   une consommation égale à  litres aux 100.

L'échantillon dont vous aurez besoin pour tester les lignes de codes est à télécharger ici  : [**essence.txt**](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525306/essence.txt)

Le code dans ce cours est donné en langage R, mais une traduction en Python est disponible en haut de ce chapitre

# Adoptez la posture du Data Analyst

Le data analyst va pouvoir apporter des réponses à chacune de ces questions, en ayant posé au préalable un **modèle probabiliste** (on reviendra sur les raisons de cette hypothèse dans la suite), discret dans le premier exemple, continu dans le second exemple.

1. A la première question, le data analyst donnera comme réponse une **estimation (ponctuelle)**  
   — du *taux de guérison théorique***p** dans le premier exemple.  
   — de la *consommation d’essence théorique***μ** dans le second exemple (on peut vouloir aussi estimer la dispersion de cette consommation ).
2. La réponse apportée à la seconde question sera un **intervalle de confiance**.
3. Enfin pour répondre à la troisième question, il mettra en place un **test statistique** pour vérifier que :  
   — **p>p0**, avec **p0=0.75**, dans le premier exemple,  
   — **μ=μ0**, avec **μ0=31**, dans le second exemple.

Nous aborderons chacun de ces 3 points dans les parties 2, 3 et 4 de ce cours. Gardez-les donc bien en tête !

Globalement, le data analyst apporte des outils d’**aide à la décision**, enjeu crucial dans de nombreux domaines aujourd’hui.

# Comprenez les enjeux de l’inférence

En [statistique descriptive](https://openclassrooms.com/courses/nettoyez-et-decrivez-votre-jeu-de-donnees), on souhaite, comme son nom l’indique, décrire un échantillon (contenant une ou plusieurs variables). Mais il est impossible d’inférer/d’extrapoler ce qui est constaté sur un échantillon à la population statistique tout entière.

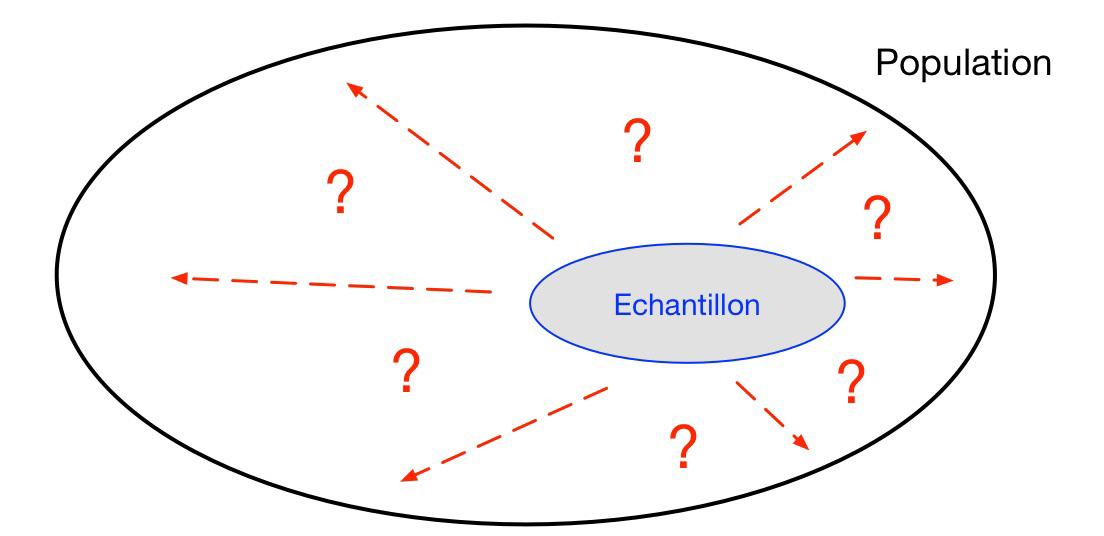
Si l'on revient à l'exemple du médicament, on ne peut que constater que le **taux de guérison observé** sur la population vaut 77,8%. Il est pourtant impossible d’affirmer qu’il s’agisse du **taux de guérison théorique**, il aurait été ainsi possible d’observer un taux de 74% ou 79% sur d’autres échantillons de taille similaire.

De même, dans l'exemple de la consommation d'essence du car, on ne peut que constater que la consommation au 100 moyenne vaut 31.45% sur notre échantillon, il est très probable qu’on observerait d’autres valeurs sur des échantillons différents.

A ce stade on peut distinguer deux sources d’aléas :

1. La **variabilité intrinsèque** du phénomène :
   * Dans le premier exemple, tous les patients présentent des spécificités qui pourront influer sur sa guérison (un patrimoine génétique différent par exemple).
   * Dans le second exemple, les mesures sont effectuées sur des cars qui peuvent présenter de légers écarts de fabrication, et sur des trajets qui présentent forcément des écarts (notamment en fonction du conducteur).
2. La **variabilité due à l’échantillonnage** : on obtiendrait sans nul doute des résultats différents (pas radicalement opposés néanmoins) sur d’autres échantillons.

Introduire un **modèle probabiliste** nous permettra de disposer d’outils théoriques capables de prendre en compte les aléas du problème, et d’étendre les résultats obtenus sur un échantillon à la population toute entière.



On suppose donc ici que chaque observation **xi** est la réalisation d’une variable aléatoire **Xi**, c’est-à-dire le fruit d’un tirage aléatoire.

En outre, on se placera, pour ce cours, dans le cas de figure où toutes nos mesures peuvent être considérées comme **indépendantes**, et **représentatives du même phénomène**.

Mathématiquement, on dit que nos variables aléatoires sont **i.i.d** : indépendantes et identiquement distribuées.

# Déterminez votre modèle probabiliste

### Quels modèles pour nos exemples introductifs ?

1. Dans le premier exemple, on est dans un modèle dit de **Bernoulli** : les variables aléatoires suivent la même loi **B(p)**, et ce de manière indépendante.
2. Dans le second exemple, on va se placer sous hypothèse **gaussienne** : les variables aléatoires suivent toutes la même loi **N(μ,σ2)**, et ce de manière indépendante. Cette hypothèse gaussienne n’est pas anodine, mais on peut l’émettre a priori ici au vu de la forme de l’histogramme ci-dessous.

A la fin de ce cours, nous pourrons vérifier que cette hypothèse est raisonnable, via un test statistique.

### Le coin R : consommation d’essence

Appliquons tout cela avec un ordinateur !

Retrouvez les datasets et les codes en R et en Python à [**cette adresse**](https://s3-eu-west-1.amazonaws.com/course.oc-static.com/courses/4525306/python_statistiques_inferentielles.html).

On importe le fichier contenant les consommations d’essence :

essence <- read.table("essence.txt",header=TRUE)

On calcule la moyenne, la variance et l’écart-type de l’échantillon :

xbar <- mean(essence$conso)

round(xbar,digits=2)

## [1] 31.45

sprime <- sd(essence$conso)

round(sprime,digits=2)

## [1] 2.16

sprime2 <- var(essence$conso)

round(sprime2,digits=2)

## [1] 4.66

Si on souhaite obtenir la variance “biaisée”, on peut l’obtenir ainsi :

n\_essence <- dim(essence)[1]

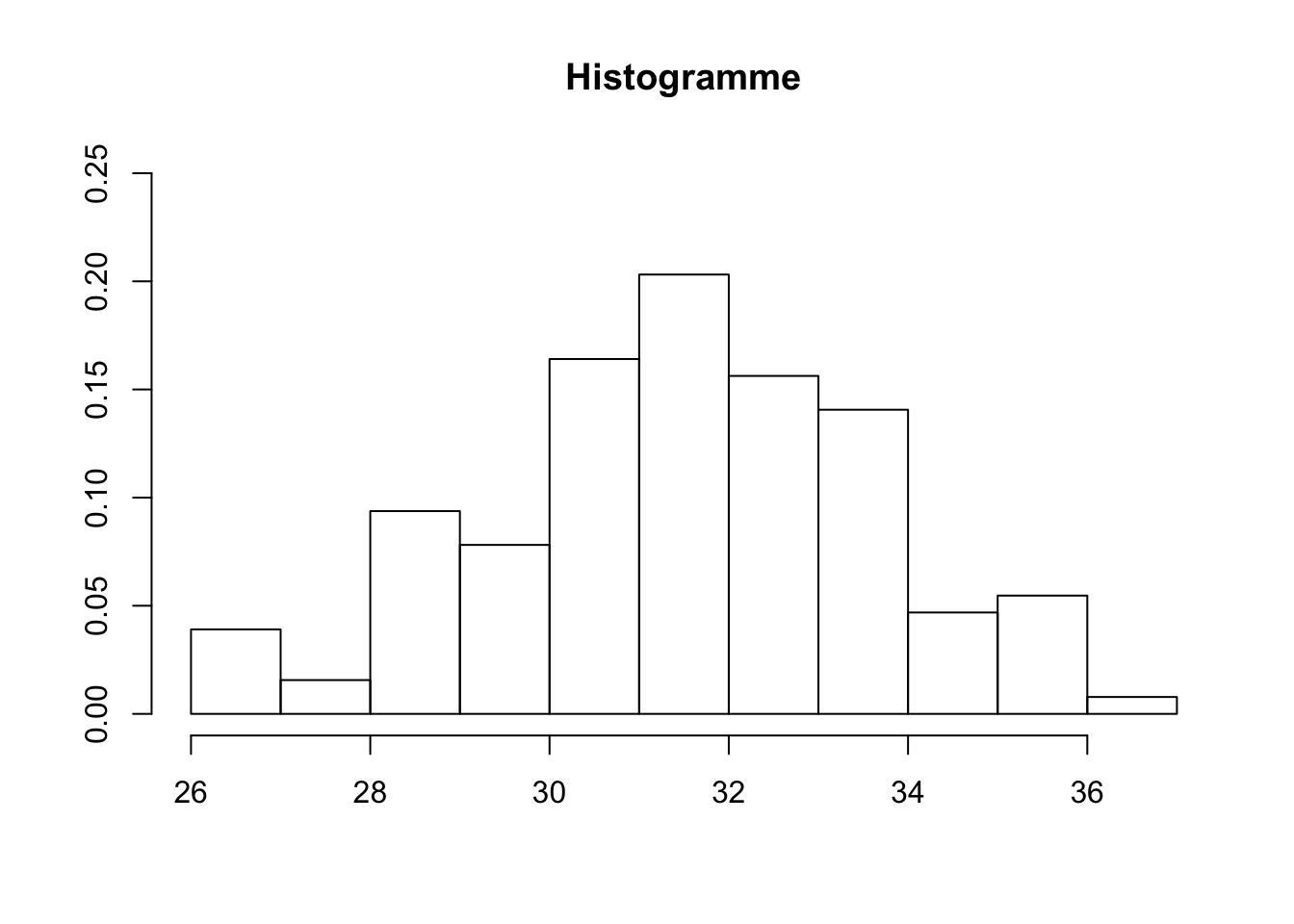
v <- sprime2\*(n\_essence-1)/n\_essence

round(v,digits=2)

## [1] 4.63

On peut visualiser l’histogramme :

hist(essence$conso,prob=TRUE,xlab="",ylab="",ylim=c(0, 0.25),main="Histogramme")

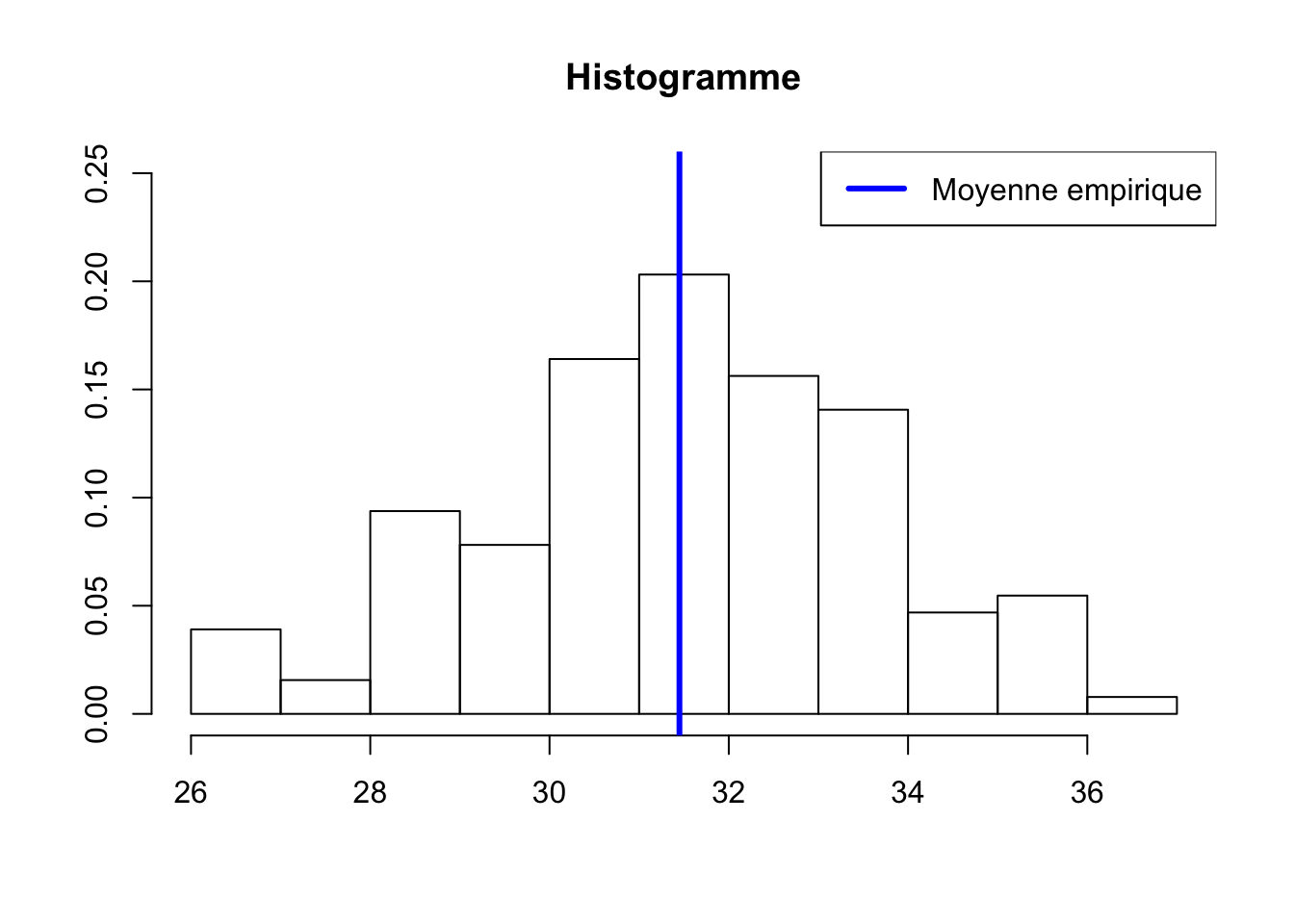


On peut visualiser sur cet histogramme la moyenne empirique :

hist(essence$conso,prob=TRUE,xlab="",ylab="",ylim=c(0, 0.25),main="Histogramme")

abline(v=xbar,col="blue",lwd=3)

legend("topright",legend=("Moyenne empirique"),col="blue",lty=1,lwd=3)



On peut visualiser sur cet histogramme la moyenne empirique et le seuil “métier” (le seuil métier est la valeur sur laquelle souhaite communiquer le constructeur : 31 litres au cent) :

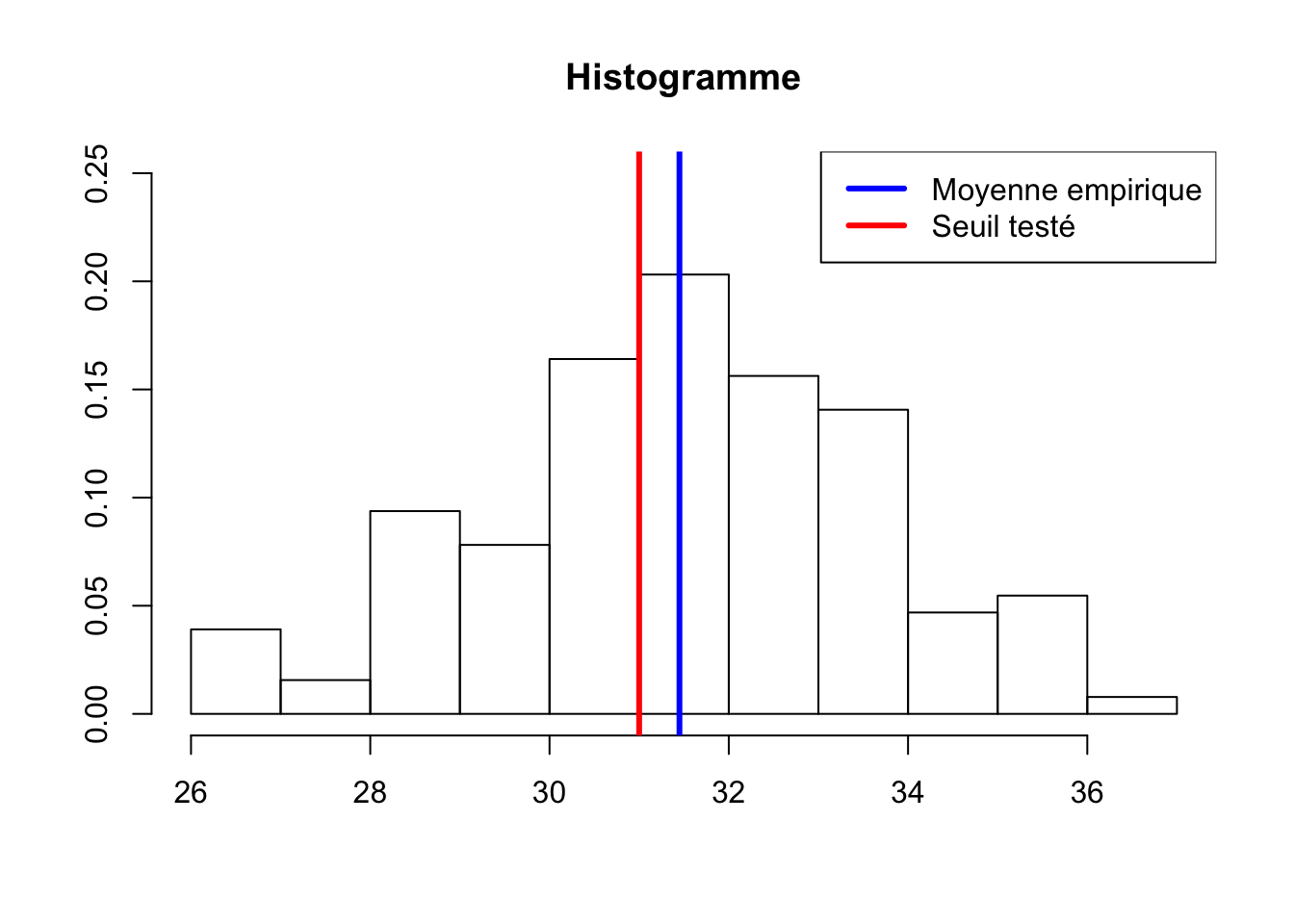
mu0 <- 31

hist(essence$conso,prob=TRUE,xlab="",ylab="",ylim=c(0, 0.25),main="Histogramme")

abline(v=xbar,col="blue",lwd=3)

abline(v=mu0,col="red",lwd=3)

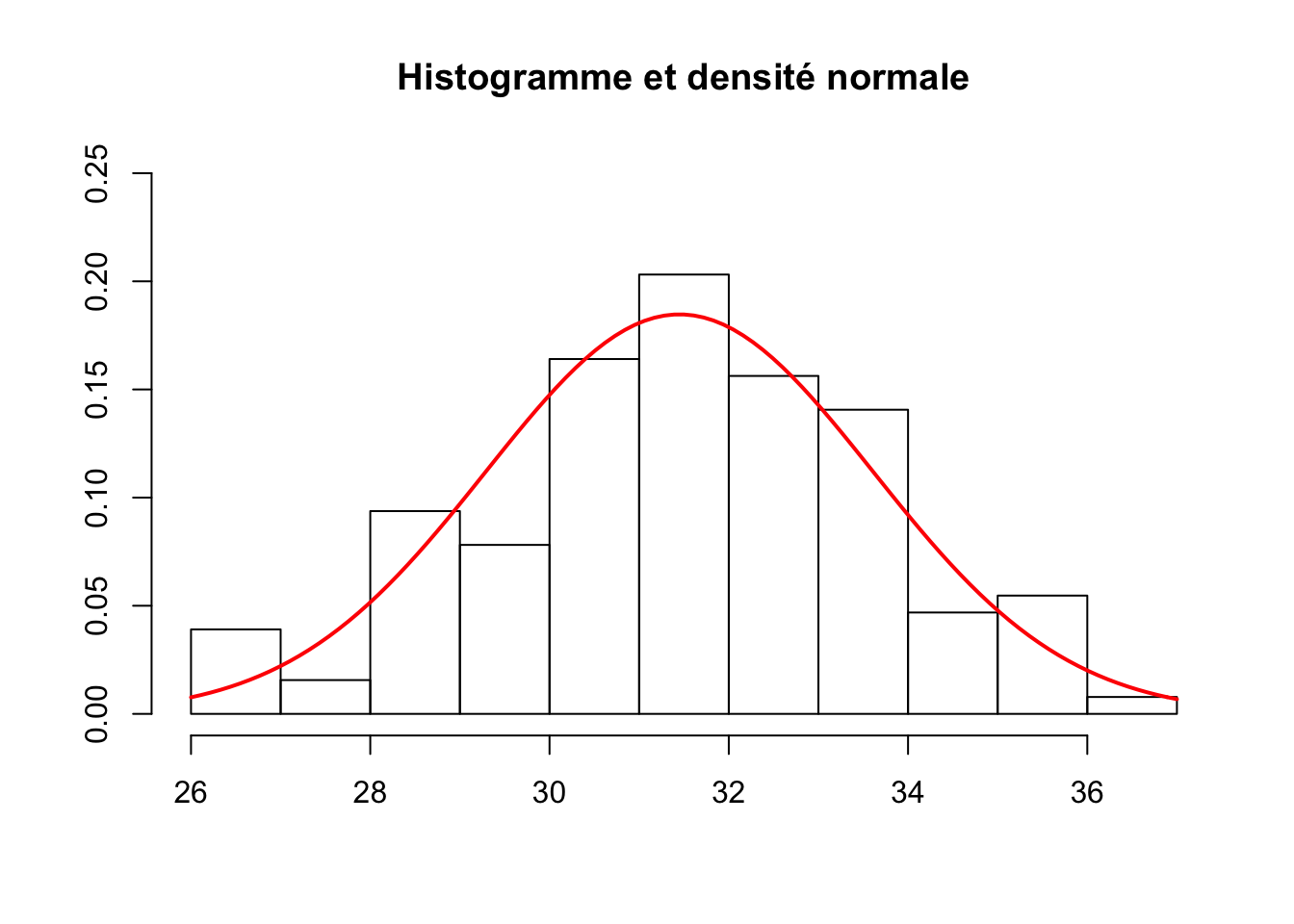
legend("topright",legend=c("Moyenne empirique","Seuil testé"),col=c("blue","red"),lty=1,lwd=3)



On peut enfin superposer sur cet histogramme la densité gaussienne :

hist(essence$conso,prob=TRUE,xlab="",ylab="",ylim=c(0, 0.25),main="Histogramme et densité normale")

curve(dnorm(x,mean=xbar,sd=sprime),col="red",lwd=2,add=TRUE,yaxt="n")



On constate que l’écart entre l’histogramme, primo-estimation de la densité de probabilité, et la densité gaussienne sont plutôt proches (car la courbe rouge et l'histogramme sont quasiment superposables).

# Découvrez la notion d'estimateur

Les deux exemples traités nous conduisent à considérer un échantillon [i.i.d](https://openclassrooms.com/courses/initiez-vous-a-la-statistique-inferentielle/p1c4-inferez)**(X1,…,Xn)** dont la loi de probabilité est connue à un paramètre

**θ** près :

* dans le premier exemple : **θ=p** ,
* dans le second exemple :  **θ=(μ,σ2)**.

Pour appréhender ces paramètres, on ne dispose que des observations **(x1,…,xn)**, réalisations des variables aléatoires i.i.d **(X1,…,Xn).**

**De manière assez simple, un estimateur est une fonction des observations**, qui prend ses valeurs dans le domaine de définition du paramètre. On le note généralement avec un chapeau :



Dans le premier exemple, il faudra par exemple considérer une fonction, l'**estimateur**, qui prenne ses valeurs dans **[0,1]** car le paramètre qu'on cherche à estimer appartient à cet intervalle.

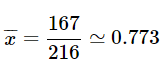
Au-delà de cette notion de base, on définira plus tard ce qu'est un "bon" estimateur. Il en existe toujours plusieurs pour un même paramètre ! L'enjeu sera donc de choisir les meilleurs...

# Estimez une proportion

### Cas du taux de guérison

Pour estimer le taux de guérison de notre premier exemple, **p**, le choix de la moyenne empirique paraît naturel. Cette quantité représente la fréquence de guérisons réellement observées sur les 216 individus de l'échantillon.

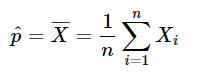
On obtient comme "estimateur" du taux de guérison :



Le taux de guérison estimé vaut donc 77,3%.

### Cas général

Dans le cas où un dispose d'un échantillon i.i.d de loi **B(p)**, on considère comme estimateur de **p** :



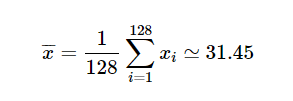
On sait que cet estimateur est doté des bonnes propriétés dont on parle dans la suite du cours.

# Estimez une moyenne et une variance

### Cas de la consommation d'essence

Pour estimer la consommation moyenne (théorique) dans le second exemple, **μ**, le choix de la  moyenne empirique  paraît là encore naturel.

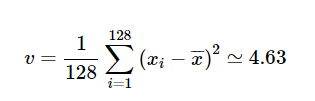
On obtient comme "estimateur" cette consommation moyenne :



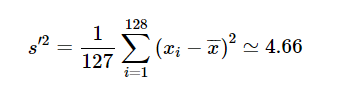
La consommation moyenne (théorique) d'essence au 100 est donc estimée à 31.45 litres.

Pour estimer cette fois la variance de cette consommation, , on choisit de manière analogue la variance... empirique ! On constate ici que les notions vues en statistiques descriptives ont tout leur intérêt dans le domaine de l'inférentiel.

On peut donc considérer comme "estimateur" cette variance de consommation :



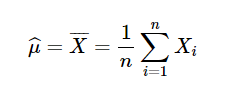
On considère souvent la version dite "non biaisée" de cette variance (on divise par 127 au lieu de 128) :



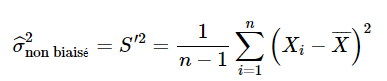
Cette version est souvent celle par défaut dans les logiciels statistiques, on rediscutera plus tard du choix privilégié de  par rapport à **V**.

### Cas général

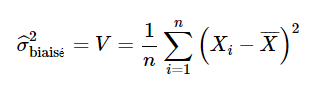
Dans le cas où un dispose d'un échantillon i.i.d dont la loi admet comme moyenne théorique (il s'agit de l'espérance mathématique !) **μ** et comme variance , on considère comme estimateur de **μ** :



et comme estimateurs de  :



ou :



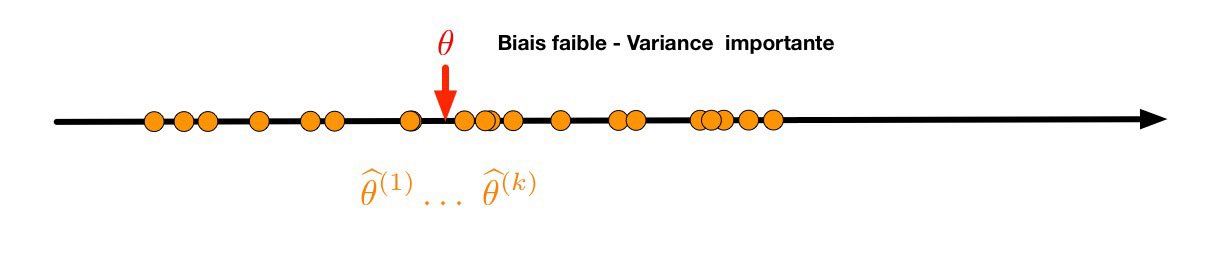
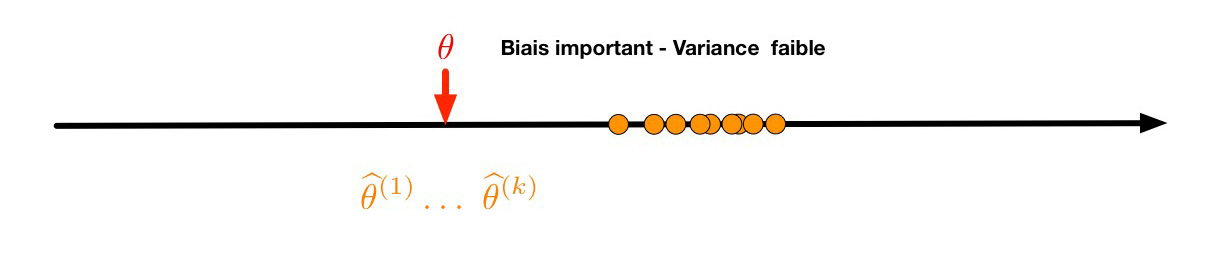
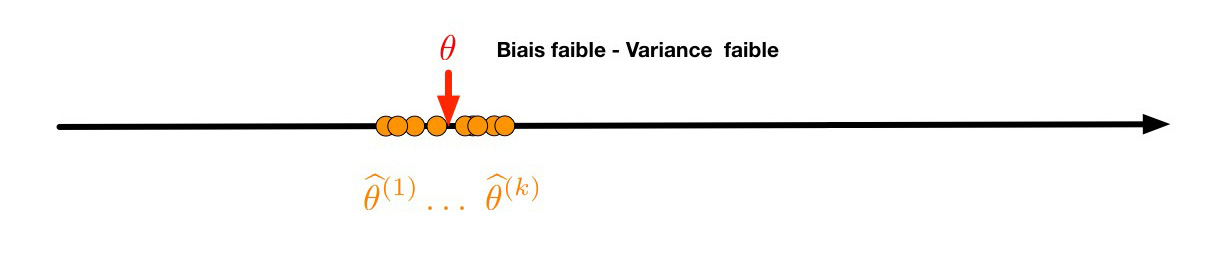
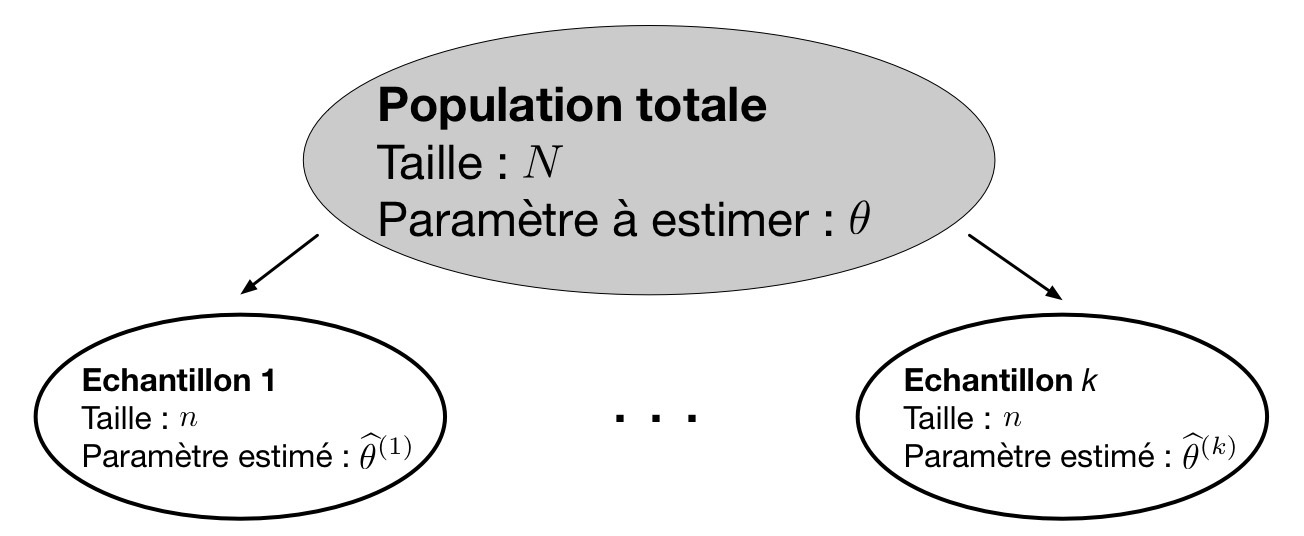
# Comprenez ce qui fait la qualité d’un estimateur

Ce chapitre est un peu théorique, mais la pratique arrive après, patience !

**Qualités d'un estimateur**

On étudie usuellement les propriétés théoriques d’un estimateur pour s’assurer de sa pertinence, au-delà de l’intuition. En pratique on utilisera des estimateurs dont la qualité théorique a été éprouvée dans la littérature, il ne vous sera donc pas utile de redémontrer ces propriétés car d'autres se sont déjà penchés sur la question. ;)

1. On souhaite qu’un estimateur soit **exhaustif**, c’est-à-dire qu’il puisse *capter* toute la connaissance sur le paramètre contenue dans les observations : l’échantillon.
2. On souhaite qu’un est estimateur soit **consistant** (**convergeant**) : l’estimateur calculé sur un échantillon sera d’autant plus fin (proche de la vérité) que la taille de l’échantillon sera importante. Si on pouvait faire croître indéfiniment la taille de l’échantillon, on se rapprocherait asymptotiquement de la vraie valeur du paramètre.
3. On souhaiterait dans l’absolu qu’un estimateur soit **sans biais** (du moins pour des  
   grands échantillons) et de **variance faible**. Pour illustrer ces effets *biais* et *variance*, considérons une population régie par un paramètre **θ** (la moyenne par exemple). Imaginons que l’on puisse tirer **k** échantillons indépendants de cette population, et que l’on dispose ainsi de **k** valeurs de l’estimateur : .

Ces illustrations montrent des estimateurs à biais et variance faibles ou forts (on a considéré pour cette illustration un paramètre en dimension 1 : θ ∈ R). Idéalement on souhaiterait être dans la troisième situation : biais et variance faibles.

Visuellement :

* + L’effet **variance** se traduit par la *dispersion* des estimateurs obtenus sur chacun  
    des échantillons.
  + L’effet **biais** se traduit par l’*écart* existant entre la moyenne des estimateurs obtenus sur chacun des échantillons et la valeur du paramètre qu’on cherche à estimer. Nous reviendrons sur cette notion par la suite.

1. On souhaiterait plus généralement que l’estimateur soit de **risque (quadratique)**  
   faible, ce risque est la somme de la la variance et du biais au carré de l’estimateur.

**Cas d’une proportion, d’une moyenne (théorique) et d’une variance (théorique)**

* Pour estimer une **proportion**, on considère la **moyenne empirique** qui est un estimateur consistant et sans biais.
* Pour estimer une **moyenne (théorique)**, on considère la **moyenne empirique** qui est un estimateur consistant et sans biais.
* Pour estimer une **variance (théorique)**, on considère :
  + de manière privilégiée, la **variance empirique corrigée** , en  (c'est-à-dire ) qui est un estimateur consistant et sans biais,
  + éventuellement, la **variance empirique non-corrigée** , en  (c'est-à-dire **V** ) qui est un estimateur consistant mais biaisé (même s'il est heureusement *asymptotiquement* sans biais).

 Pour plus de précisions mathématiques sur les variances empiriques corrigé et non corrigée, rendez-vous au chapitre suivant ;)

# Déterminez la qualité de votre estimateur

### Mathématiquement

Un estimateur est **exhaustif** si la loi de l'échantillon **(X1,…,Xn)** conditionnellement à l'estimateur  est indépendante du paramètre **θ** .

Un estimateur  de **θ** est **consistant** (**convergeant**) si :



Il s'agit d'une convergence dite **en probabilité**. Point de détail mathématique, la convergence **presque** **sûre**, qu'on retrouve dans la [**loi forte des grands nombres**](https://fr.wikipedia.org/wiki/Loi_forte_des_grands_nombres), entraîne la **convergence** **en en probabilité**.

Le **biais** de l'estimateur  de **θ** vaut :

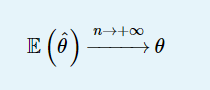


Il s'agit de l'écart entre la moyenne théorique de notre estimateur et le paramètre.

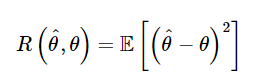
Un estimateur  de **θ** est dit sans biais si son biais est nul :



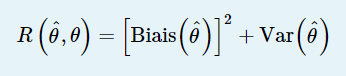
et asymptotiquement sans biais si son biais est asymptotiquement nul :



Le risque quadratique d'un estimateur vaut :



Il peut s'écrire comme la somme du carré du biais et de la variance de l'estimateur :

****

### Retour au taux de guérison

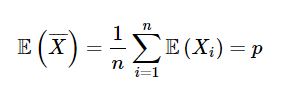
La loi forte des grands nombres (échantillon i.i.d de loi **B(p)**) assure que :



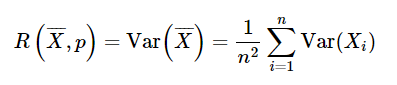
La convergence presque sûre entraînant la convergence en probabilité, l'estimateur est donc consistant.



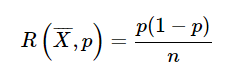
L'estimateur est sans biais, en effet :



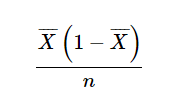
L'estimateur étant sans biais, le risque quadratique est égal à la variance :



car les variables **Xi** sont indépendantes, donc non corrélées, d'où :



Le risque quadratique fournit un indicateur de la dispersion de  autour de la valeur à estimer **p** . Certes on ne connaît pas sa valeur qui dépend de **p**, mais on peut l'estimer par :



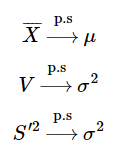
On constate que le risque quadratique décroît logiquement quand la taille de l'échantillon augmente.

Notons que ces résultats sont valables quelle que soit la loi, pas seulement pour une loi 

Tout cela nous conforte dans notre décision intuitive de choisir la moyenne empirique (la fréquence empirique) comme estimateur.

### Retour à la consommation d'essence

La loi forte des grands nombres (et un théorème, celui de [Slutsky](https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_Slutsky)) assure que les estimateurs  sont consistants :

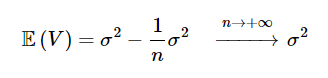


La convergence presque sûre entraîne toujours la convergence en probabilité.

est un estimateur sans biais de **μ** :



**V** est un estimateur biaisé, mais asymptotiquement sans biais, de  :



Si **V** est un estimateur biaisé (mais asymptotiquement sans biais) de , est quant à elle sans biais :



C'est pour cela qu'on retrouve usuellement dans les logiciels la version dite sans biais (unbiased) .

Si la démonstration complète vous intéresse, c'est [**par ici**](https://fr.wikipedia.org/wiki/Variance_(statistiques_et_probabilit%C3%A9s)#Biais)!

## Allez plus loin : méthodes des moments et du maximum de vraisemblance

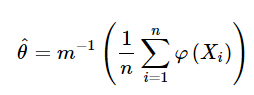
Ce chapitre est facultatif si vous souhaitez vous former au métier de Data Analyst. Par contre, il est obligatoire pour ceux qui visent le métier de Data Scientist.

Notez que, contrairement à ce que nous avons vu dans le chapitre précédent, il n'est pas toujours aussi simple de trouver des estimateurs. Il existe des méthodologies pour imaginer des estimateurs, en sus des idées "naturelles", parmi lesquelles la **méthode des moments** et la **méthode du maximum de vraisemblance**.

### Méthode des moments

La méthode des moments consiste à trouver une fonction **m**, continue et inversible, et une fonction (continue) **φ** telles que .

L'estimateur des moments pour **θ** vaut :

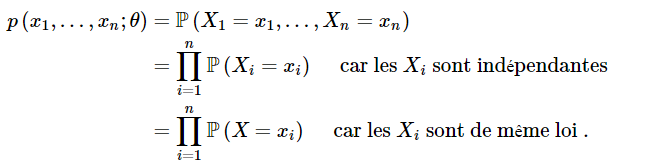


On sait que cet estimateur est consistant.

### Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance, comme son nom l'indique, maximise la vraisemblance définie comme suit :

#### Dans le cas discret i.i.d :



Rappelons que la loi de **X** et donc que   dépend de **θ**.

#### Dans le cas continu i.i.d :

#### 

#### Maximum de vraisemblance

La vraisemblance mesure la probabilité que les observations proviennent effectivement d'un échantillon de loi paramétrée par **θ**. Trouver le maximum de vraisemblance consiste donc à trouver le paramètre le plus vraisemblable pour notre échantillon !

On considère usuellement la log-vraisemblance (qui facilite les calculs pour des lois de probabilité appartenant à la famille dite exponentielle) :



L'intérêt de cette méthode, au-delà de trouver un estimateur dans un cas moins intuitif, est de trouver un estimateur qui est muni d'excellentes qualités dans de très nombreux cas : consistant, asymptotiquement sans biais et de variance minimale. La plupart du temps, cet estimateur est unique et se détermine explicitement.

### Application à la loi exponentielle

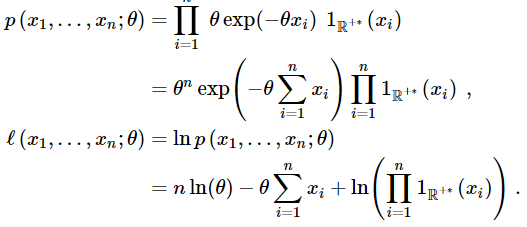
#### Estimateur du maximum de vraisemblance

Soit un échantillon**(X1,…,Xn)** de loi .

Rappelons que la densité de cette loi exponentielle est :

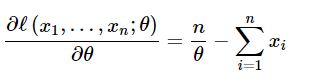


On a :

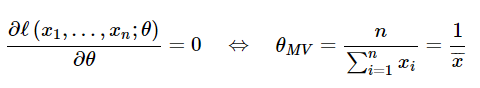


Le dernier terme n'intervient pas en réalité, il est nul si tous les points de l'échantillon sont strictement positifs (et formellement égal à **−∞** sinon).

Il suffit donc de dériver les deux premiers termes par rapport à **θ** pour déterminer l'extremum (et on vérifie qu'il s'agit bien d'un maximum !) :



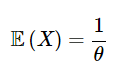
On obtient :



est donc l'estimateur du maximum de vraisemblance de **θ** .

#### Méthode des moments

On aurait également pu obtenir cette solution par la méthode des moments en notant que pour une loi :



Il suffisait de considérer les fonctions :



Notons qu'on aurait également pu se baser sur le résultat suivant :



pour obtenir un autre estimateur, mais celui-ci aurait été moins performant que l'estimateur du maximum de vraisemblance.

# Testez-vos connaissances sur les estimateurs

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Comprendre la notion d’inférence
* Etablir un lien entre des observations et un modèle probabiliste
* Réaliser une estimation ponctuelle

### Question 1

**Si l’on doit modéliser un phénomène du type échec/réussite, on utilisera une loi :**

* + 

de Bernoulli

* + 

de Poisson

* + 

normale

*Comme vu avec le taux de guérison, on utilise une loi de Bernoulli.*

### Question 2

**Un estimateur d’un paramètre d’une loi :**

* + 

est une fonction quelconque de l’échantillon

* + 

est une fonction de l’échantillon qui prend ses valeurs dans le même espace que le paramètre

* + 

ne dépend pas de l’échantillon

*Un estimateur est basé sur des données observées, l’échantillon, et doit prendre ses valeurs dans le même espace que le paramètre.*

### Question 3

**Pour estimer une proportion (avec un échantillon constitué de 0 et de 1), on utilise (plusieurs réponses possibles) :**

*Attention, plusieurs réponses sont possibles.*

* + 

la moyenne empirique de l’échantillon

* + 

la fréquence empirique des "1"

* + 

la fréquence empirique des  "0"

*On comptabilise le nombre de 1 pour estimer la proportion théorique, il s’agit également de la moyenne empirique.*

### Question 4

**L’estimateur sans biais de la variance d’une loi de probabilité est :**

* + 
  + 
  + 



* + 



*La réponse 2 représente la variance dite "biaisée" et la réponse 3 est une autre version biaisée.*

### Question 5

**Un bon estimateur est forcément :**

* + 

sensible

* + 

sans biais

* + 

consistant

*On peut espérer qu’un estimateur soit biaisé mais ce n’est pas une obligation (on attendra simplement qu’il soit asymptotiquement sans biais). Quant à la sensibilité, ce n’est pas une propriété d’un estimateur, et cette notion n'est pas abordée dans ce cours (pour en savoir plus, rendez-vous*[*ici*](https://fr.wikipedia.org/wiki/Sensibilit%C3%A9_et_sp%C3%A9cificit%C3%A9)*).*

## Découvrez les intervalles de confiance

On passe maintenant à la réponse à la [deuxième question](https://openclassrooms.com/courses/initiez-vous-a-la-statistique-inferentielle/p1c3-sachez-quelles-reponses-apporter), grâce aux intervalles de confiance !

### L'idée

On a vu précédemment que l'estimation d'un paramètre **θ** peut différer selon l'échantillon qu'on va considérer. Cet estimateur  est bel et bien une variable aléatoire qui tombe "autour" de **θ** mais rarement sur sa "vraie" valeur.

De la loi de probabilité de notre variable aléatoire, l'estimateur, on va à partir d'un intervalle sur **** déduire un intervalle sur notre paramètre **θ**. Et on pourra énoncer des phrases du type "Il y a 95% de chances pour que **θ** se trouve dans cet intervalle".

### Mathématiquement

Cette fois, on cherche une estimation du paramètre **θ** dans un **intervalle de confiance**, une fourchette dont on connaîtra la probabilité.

On cherche donc à déterminer les bornes d'un intervalle, dépendantes de l'échantillon, notées  et , telles que la probabilité que le paramètre soit à l'intérieur soit dans cet intervalle, soit connue, égale à **1−α** :



**1−α∈]0,1[** désigne le **niveau de confiance** de l'intervalle.

Tel qu'écrit, il s'agit d'un intervalle de confiance **bilatère** (on encadre le paramètre à gauche ET à droite), il est également possible de construire un intervalle **unilatère** (on encadre le paramètre à gauche OU à droite).

On se trouve toujours face à un dilemme : pour garantir le niveau de confiance, l'intervalle ne doit pas être trop étroit mais, pour être pratiquement utilisable, il ne doit pas être trop large. On cherche donc des intervalles aussi étroits que possible, au niveau de confiance **1−α** imposé, et ce uniformément en **θ**, d'où la difficulté du problème.

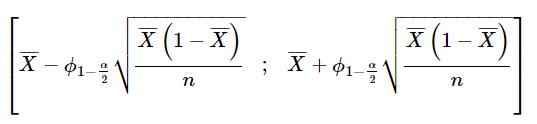
Classiquement on considère des intervalles de confiance de niveaux 90% (**α=10**) ou 95% (**α=5**).

**Déterminez un intervalle de confiance sur une proportion**

**Le coin méthodologique**

Dans le cas de petits échantillons, on se base sur la loi binomiale, discrète, tandis que dans le cas de grands échantillons (en pratique **n≥30** ) on préfère utiliser l'approximation par une loi normale, on parle alors d'un intervalle de confiance asymptotique. Notons que les calculs sont facilités par l'utilisation d'une loi continue.

Sous cette dernière hypothèse gaussienne, l'*intervalle de confiance bilatère (symétrique) de niveau***1−α***pour***p** est :



où **ϕ1−α/2** désigne le quantile d'ordre **1−α/2** de la loi **N(0,1)**.

On remarque que :

* L'encadrement autour de fait intervenir la quantité qui est une estimation de la variance d'une loi de Bernoulli : .
* Pour un niveau de confiance fixé, la largeur de l'intervalle de confiance sera d'autant plus petite que la taille d'échantillon sera élevée : plus on dispose d'informations, plus on est confiant.
* La largeur de l'intervalle sera d'autant plus large que le niveau de confiance sera élevé (le quantile d'ordre , croît lorsque croît).

**Le coin R : exemple du taux de guérison**

On importe le fichier contenant les guérisons ou non-guérisons :

guerison <- read.table("guerison.txt",header=TRUE)

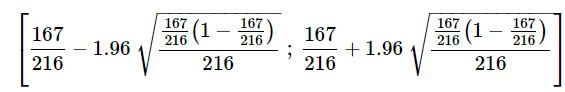
On peut estimer le taux de guérison théorique pp :

n\_guerison <- dim(guerison)[1]

n\_guerison\_gueris <- sum(guerison$taux==1)

p\_estim <- n\_guerison\_gueris/n\_guerison

 Si on souhaite encadrer le taux de guérison p avec une probabilité de **1−α=95%**, on obtient alors comme intervalle de confiance Φ1−α/2=Φ0.975≃1.96



Si on lance “manuellement” les calculs au niveau de test 5% :

alpha <- 0.05

icinf <- p\_estim-qnorm(p=1-alpha/2)\*sqrt(p\_estim\*(1-p\_estim)/n\_guerison)

round(icinf,digits=2)

## [1] 0.72

icsup <- p\_estim+qnorm(p=1-alpha/2)\*sqrt(p\_estim\*(1-p\_estim)/n\_guerison)

round(icsup,digits=2)

## [1] 0.83

On obtient alors :

**[0.72 ; 0.82]=[72% ; 83%]**

On constate que la largeur de l’intervalle n’est pas négligeable, mais n’oublions pas qu’il n’y a que 216 individus dans l’échantillon.

En pratique, le data analyst peut obtenir simplement cet intervalle à l’aide de la commande  prop.test  (cette commande lance d’autres calculs en sus de l’intervalle de confiance, on les verra plus tard) :

alpha <- 0.05

prop.test(n\_guerison\_gueris,n\_guerison,conf.level=1-alpha)

##

## 1-sample proportions test with continuity correction

##

## data: n\_guerison\_gueris out of n\_guerison, null probability 0.5

## X-squared = 63.375, df = 1, p-value = 1.709e-15

## alternative hypothesis: true p is not equal to 0.5

## 95 percent confidence interval:

## 0.7103413 0.8260184

## sample estimates:

## p

## 0.7731481

Les résultats numériques diffèrent légèrement à cause d’une “correction de normalité” effectuée par cette commande.

On aurait également pu obtenir un intervalle de confiance “exact”, basé sur la loi binomiale, à l’aide de la commande  binom.test  :

alpha <- 0.05

binom.test(n\_guerison\_gueris,n\_guerison,conf.level=1-alpha)

##

## Exact binomial test

##

## data: n\_guerison\_gueris and n\_guerison

## number of successes = 167, number of trials = 216, p-value =

## 2.925e-16

## alternative hypothesis: true probability of success is not equal to 0.5

## 95 percent confidence interval:

## 0.7114251 0.8272107

## sample estimates:

## probability of success

## 0.7731481

On constate que les résultats diffèrent très peu de l’intervalle de confiance asymptotique.

Enfin, si on avait choisi un niveau de confiance plus faible, **1−α=90%** par exemple, on aurait obtenu un intervalle de confiance plus étroit :

alpha <- 0.10

prop.test(n\_guerison\_gueris,n\_guerison,conf.level=1-alpha)

##

## 1-sample proportions test with continuity correction

##

## data: n\_guerison\_gueris out of n\_guerison, null probability 0.5

## X-squared = 63.375, df = 1, p-value = 1.709e-15

## alternative hypothesis: true p is not equal to 0.5

## 90 percent confidence interval:

## 0.7206161 0.8185913

## sample estimates:

## p

## 0.7731481

On obtient ici **[0.721 ; 0.819]=[72.1% ; 81.9%]**, la largeur de l’intervalle a bel et bien diminué !

## Déterminez un intervalle de confiance sur une moyenne

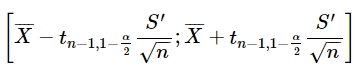
### Le coin méthodologique

Considérons un échantillon i.i.d de loi **N(μ,σ2)**, ou un grand échantillon i.i.d non gaussien (en pratique de taille supérieure à 30).

On va retrouver un intervalle de confiance qui fait intervenir les mêmes types de quantité :

* La taille de l'échantillon.
* La dispersion (empirique) de l'échantillon.
* Des quantiles d'une loi de probabilité : pas la loi gaussienne curieusement ici mais une loi proche, celle de [Student](https://fr.wikipedia.org/wiki/Loi_de_Student) (cela est dû à la non-connaissance de la variance théorique  , pour les plus curieux on donne l'argument ci-après).

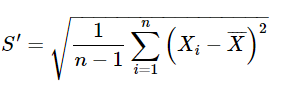
L'intervalle de confiance bilatère de niveau **1−α** pour **μ** est alors :



où **tn−1,1−α/2** désigne le quantile d'ordre **1−α/2** de la loi de Student à **n−1** degrés de liberté :

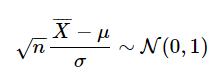


Rappelons que  est l'écart type empirique (dans sa version non biaisée) :



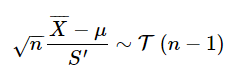
### Pour aller plus loin

Mathématiquement, pour établir l'intervalle de confiance, on aurait pu se baser sur le résultat probabiliste suivant :

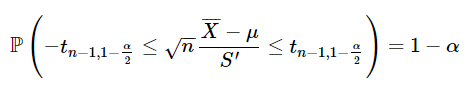


Ce résultat est vrai pour un échantillon i.i.d gaussien, ou asymptotiquement vrai pour un grand échantillon (via le théorème de la limite centrale, le fameux [TCL](https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_central_limite)).

Ce résultat fait intervenir l'écart-type théorique, **σ** , inconnu, l'idée est alors de remplacer cet écart-type théorique inconnu par l'empirique, , on obtient alors une loi de Student (à **n−1** degrés de liberté) :



On peut alors facilement encadrer notre statistique de la manière suivante :

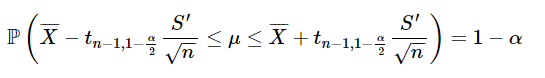


Il faut noter ici qu'on a, pour une loi de Student :



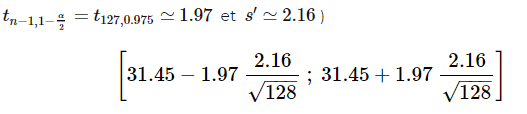
cette loi étant symétrique.

On déduit de cet encadrement le suivant (par simples équivalences) :



d'où l'intervalle de confiance donné précédemment.

### Le coin R : consommation d'essence

Si on souhaite encadrer la consommation d’essence moyenne (théorique) μ avec une probabilité de 95%, on obtient alors comme intervalle de confiance 

Remarquons ici que l’hypothèse gaussienne n’était pas obligatoire ici, en effet l’échantillon est de taille suffisamment importante (supérieure à 30).

Si on lance “manuellement” les calculs au niveau de test 5% :

alpha <- 0.05

icinf <- xbar-qt(p=1-alpha/2,df=n\_essence-1)\*sprime/sqrt(n\_essence)

round(icinf,digits=2)

## [1] 31.07

icsup <- xbar+qt(p=1-alpha/2,df=n\_essence-1)\*sprime/sqrt(n\_essence)

round(icsup,digits=2)

## [1] 31.83

On obtient alors :

**[31.07 ; 31.83]**

En pratique, le data analyst peut utiliser la commande  t.test  pour obtenir cet intervalle de confiance :

alpha <- 0.05

t.test(essence$conso,conf.level=1-alpha)

##

## One Sample t-test

##

## data: essence$conso

## t = 164.74, df = 127, p-value < 2.2e-16

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

## 31.07169 31.82722

## sample estimates:

## mean of x

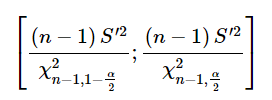
## 31.44945

## Déterminez un intervalle de confiance sur une variance

### Le coin méthodologique

Considérons un échantillon i.i.d de loi **N(μ,σ2)**, ou un grand échantillon i.i.d non gaussien (en pratique de taille supérieure à 30).

L'intervalle de confiance bilatère de niveau 1−α1−α pour σ2σ2 est alors :

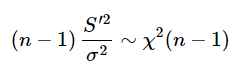


où désignent respectivement les quantiles d'ordres **α/2** et **1−α/2** de la loi  

La distribution n'est pas symétrique, ce qui explique la différence avec les autres intervalles de confiance dans lesquels un seul quantile apparaissait.

### Pour aller plus loin

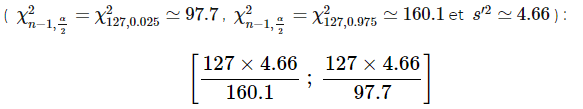
Mathématiquement, pour établir l'intervalle de confiance, on se base sur le résultat probabiliste suivant :



Ce résultat est vrai pour un échantillon i.i.d sous hypothèse normale, ou asymptotiquement vrai pour un grand échantillon i.i.d quelconque.

### Le coin R : exemple de la consommation d'essence

Si on souhaite encadrer la variance (théorique) de la consommation d’essence  avec une probabilité de 95%, on obtient alors comme intervalle de confiance



Si on lance “manuellement” les calculs au niveau de test 5% :

alpha <- 0.05

icinf <- (n\_essence-1)\*sprime2/qchisq(p=1-alpha/2,df=n\_essence-1)

round(icinf,digits=2)

## [1] 3.7

icsup <- (n\_essence-1)\*sprime2/qchisq(p=alpha/2,df=n\_essence-1)

round(icsup,digits=2)

## [1] 6.06

On obtient alors :

Remarquons encore ici que l’hypothèse gaussienne n’était pas obligatoire ici, en effet l’échantillon est de taille suffisamment importante (supérieure à 30).

En pratique, le data analyst pourra utiliser la commande  varTest  issue du package  EnvStats  pour obtenir cet intervalle de confiance :

[3.70 ; 6.06][3.70 ; 6.06]

library(EnvStats)

##

## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':

##

## predict, predict.lm

## The following object is masked from 'package:base':

##

## print.default

alpha <- 0.05

varTest(essence$conso,conf.level=1-alpha)

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: variance = 1

##

## Alternative Hypothesis: True variance is not equal to 1

##

## Test Name: Chi-Squared Test on Variance

##

## Estimated Parameter(s): variance = 4.66481

##

## Data: essence$conso

##

## Test Statistic: Chi-Squared = 592.4309

##

## Test Statistic Parameter: df = 127

##

## P-value: 0

##

## 95% Confidence Interval: LCL = 3.700708

## UCL = 6.063869

# Testez vos connaissances sur les intervalles de confiance

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Comprendre la notion d’inférence
* Etablir un lien entre des observations et un modèle probabiliste
* Calculer un intervalle de confiance pour une proportion et une moyenne

### Question 1

**Toutes choses égales par ailleurs, la largeur d’un intervalle de confiance :**

* + 

décroît avec la taille de l’échantillon

* + 

croît avec la taille de l’échantillon

*Plus l’échantillon est important, plus on est confiant.*

### Question 2

**Toutes choses égales par ailleurs, la largeur d’un intervalle de confiance :**

* + 

décroît avec le niveau de confiance

* + 

croît avec le niveau de confiance

*Plus le niveau de confiance attendu est élevé, plus on doit considérer un intervalle large.*

### Question 3

**On vous donne [12 ; 25] comme intervalle de confiance bilatère sur la moyenne (espérance) d’une loi. Alors :**

* + 

on est certain que la moyenne est comprise entre 12 et 25

* + 

on ne peut rien dire en l’état, il nous manque des informations sur la loi de probabilité

* + 

on ne peut rien dire en l’état, il nous manque des informations sur le niveau de confiance.

*On ne connaît pas le niveau de confiance, il n’est donc pas possible d’interpréter cet intervalle.*

### Question 4

**Dans un intervalle de confiance, on retrouve le quantile d’une loi normale :**

* + 

toujours

* + 

jamais

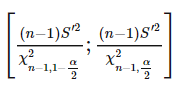
* + 

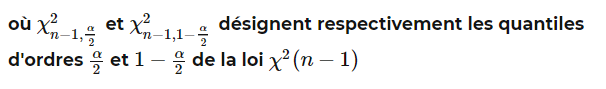
parfois

*On a vu en cours des intervalles de confiance qui faisaient intervenir d’autres lois que la loi normale : la loi de Student et la loi du Khi-deux.*

### Question 5

**Voici un intervalle de confiance de niveau**1−α**:**

****

****

* + 

Cet intervalle est bilatère.

* + 

Cet intervalle est unilatère.

## Découvrez les tests statistiques

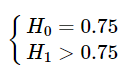
Nous passons maintenant à la réponse à la [troisième question](https://openclassrooms.com/courses/initiez-vous-a-la-statistique-inferentielle/p1c3-sachez-quelles-reponses-apporter), à laquelle nous répondons grâce aux tests statistiques.

### Cas du taux de guérison

Le laboratoire cherche à savoir si la nouvelle composition du médicament présente un taux de guérison meilleur que le précédent. Il considère a priori que son efficacité est similaire au précédent médicament et ne prendra la responsabilité de proposer la nouvelle composition que si le nouveau taux de guérison s'avère être **significativement** supérieur à celui de l'ancien médicament **p0=0.75**.

Son a priori correspond à ce qu'on appellera **l'hypothèse nulle**, notée **H0**. Quant à l'autre hypothèse, l'alternative, notée **H1**, elle permet d'indiquer dans quel cas de figure on rejettera cet a priori.

On considère donc ici le test suivant :



L'hypothèse **p=p0** paraît intuitivement d'autant moins crédible que  proportion de guérison, est plus forte. Quand  sera jugé "suffisamment plus élevé" que **p0**(significativement supérieur à **p0**), le laboratoire pourra rejeter l'hypothèse **p=p0**.

C'est le rejet de **H0** qui, s'il est fait à mauvais escient, sera considéré comme le plus coûteux pour le laboratoire, car ayant des répercussions humaines et économiques néfastes : on parle de **risque de première espèce**.

En pratique on recherche la valeur **c** (**≥0**) telle qu'on rejettera l'hypothèse nulle si :



c'est-à-dire quand la proportion de guérison observée est "vraiment" supérieure à **p0**.

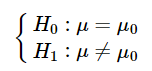
Pour fixer ce seuil **c** qui détermine la région de rejet, le data analyst doit demander au laboratoire de fixer une borne supérieure à la probabilité qu'il juge tolérable pour ce rejet à mauvais escient : le **niveau de test**.

Ensuite, sous cette contrainte, on cherchera à choisir **c** de manière à minimiser la probabilité de non-rejet de **H0** à mauvais escient : le risque de seconde espèce.

On constate que la première préoccupation (sur le risque de première espèce) conduit à souhaiter prendre la frontière **c** assez élevée, et la seconde (sur le risque de seconde espèce) pousse au contraire à l'abaisser.

### Cas de la consommation d'essence

Le fabricant de cars souhaite communiquer sur une consommation moyenne (théorique) d'essence égale à **μ0=31** litres aux 100. Il ne souhaite pas sous-estimer ou sur-estimer cette valeur seuil.  
De la même manière que précédemment, on considère le test :



Ce test est dit bilatère car il existe deux motifs de rejet de son hypothèse de travail.

L'hypothèse **μ=μ0** paraît intuitivement d'autant moins crédible que , consommation moyenne observée sur son échantillon, est jugé "suffisamment différente" de **μ0** (significativement inférieure ou supérieure à **μ0**). C'est toujours ce rejet qui, s'il est fait à mauvais escient, sera considéré comme plus coûteux pour le constructeur...

On pratique on recherche la valeur **c>0** telle qu'on rejette l'hypothèse nulle si :



c'est-à-dire si la consommation d'essence moyenne observée est "vraiment différente" de **μ0**.

Pour fixer ce seuil **c** qui détermine la région de rejet, le data analyst doit là-encore demander au constructeur de fixer une borne supérieure pour son risque de première espèce, celui de rejeter à tort son a priori.

## Formalisez votre problème de test

On cherche donc à tester une hypothèse sur un paramètre **θ** (**p** dans le premier exemple introductif, **μ** dans le second).

Beaucoup de concepts théoriques dans ce chapitre : les applications arrivent par la suite ;)

### Les hypothèse d'un test

On appelle **hypothèse nulle**, notée **H0**, l'hypothèse qu'on ne souhaite pas rejeter trop facilement : c'est celle qu'on teste.

L'**hypothèse alternative**, notée **H1**, indique dans quelles conditions on rejette **H0** (elle fournit les informations de forme sur la région critique, définition à suivre immédiatement !).

### La région critique

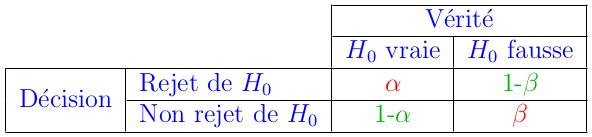
La **région critique W** indique quand on rejette l'hypothèse nulle, elle est basée sur une **statistique de test**. Pour déterminer entièrement la région critique, il faut connaître la loi (éventuellement asymptotique) de la statistique de test sous **H0**.

### Les risques associés aux décisions

On considère plusieurs risques pour la prise de décision, et ce de manière séquentielle :

* le **risque de première espèce** désigne celui de rejeter l'hypothèse nulle **H0** alors que cette hypothèse est vraie. **C'est le risque dont on veut se prémunir en premier lieu.**  
  On se fixe un niveau de test qui représente le risque de première espèce maximum. On note **α** ce niveau. Les valeurs couramment employées sont 5%, 1% et 0.5%.
* Une fois ce premier risque borné par le niveau de test, on minimise le **risque de seconde espèce β** (de manière équivalente, maximise la puissance du test **1−β**).

On peut résumer ces risques dans le tableau suivant :



Les cases en vert désignent les configurations de bonne décision, les cases en rouge les configurations d'erreur dans la décision.

### La p-valeur

Usuellement on travaille avec la **p-valeur** du test : la plus petite valeur du niveau de test conduisant au rejet de **H0**. C'est cette valeur qu'on trouve dans les logiciels statistiques. On construit cette **p-valeur** de manière à avoir équivalence entre le rejet de **H0** et le fait que la **p-valeur** soit inférieure au niveau de test :



Si la p-valeur est faible (selon un critère fixé par le domaine concerné), cela signifie que l’observation est rare (donc peu probable) sous **H0**, ce qui incite à rejeter l'hypothèse **H0**.

## Testez une proportion

### Testez p=p0 versus p>p0

On teste :

#### 

#### Choisir une statistique de test et la région critique

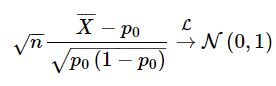
La forme de la zone de rejet (région critique) est : on rejettera d'autant plus facilement **H0** que sera plus élevée que **p0**. La statistique de test utilisée est .

Notons que la structure de la zone de rejet est donnée par l'hypothèse alternative **H1**.

#### Vers un test asymptotique

Afin de déterminer **c**, on préfère considérer la loi asymptotique (continue) qu'est la loi normale (d'où le nom de test asymptotique).

On sait que si **(X1,…,Xn)** est un échantillon i.i.d de loi **B(p0)**, le théorème de la limite centrale indique que :



Cette approximation n'est utilisée en pratique que si .

Sous cette condition, qui nous conduit à considérer ici un test dit asymptotique, on obtient après calculs :

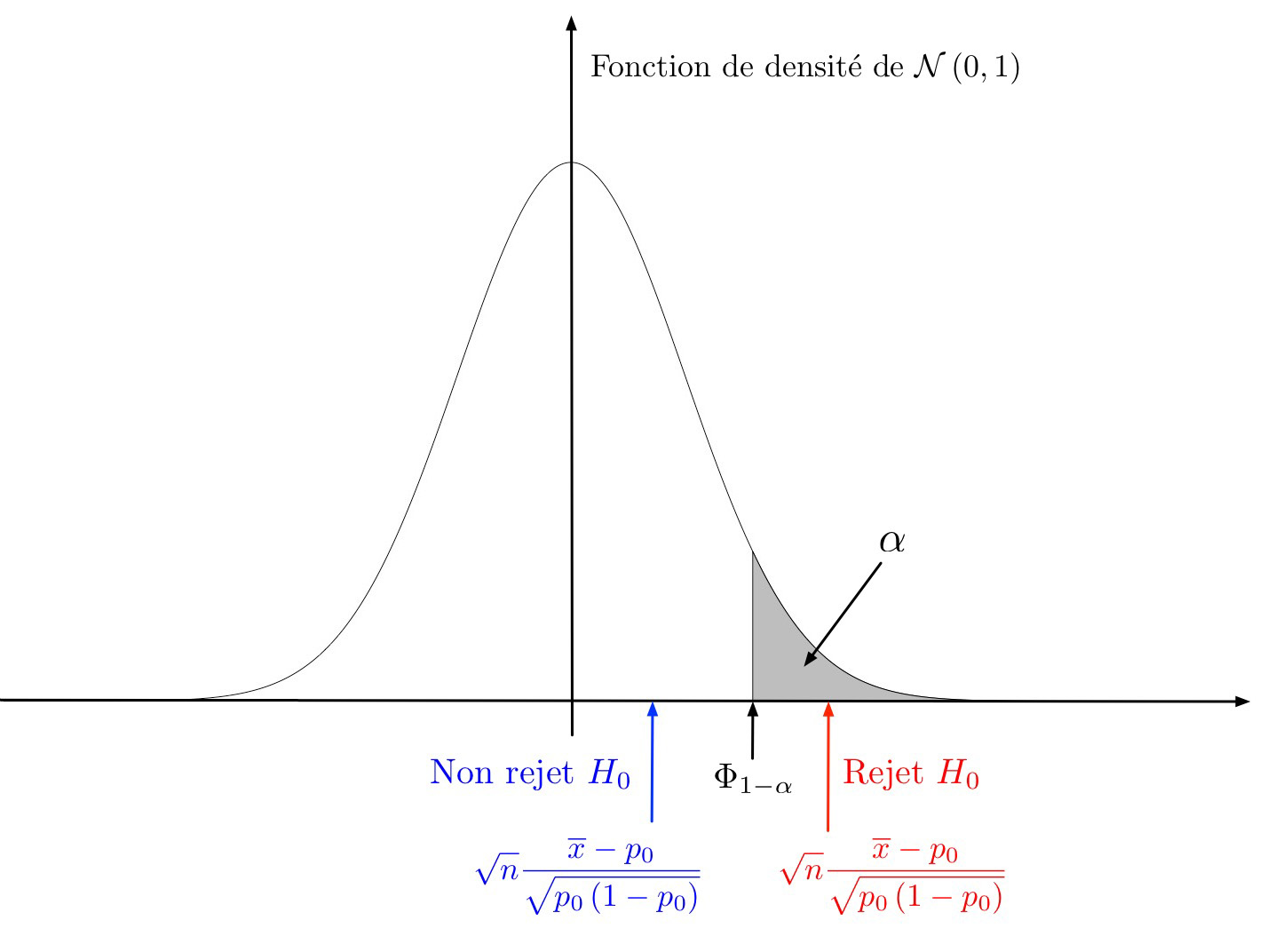
#### 

#### La règle de décision

Les décisions prises sont, avec un niveau de test égal à **α** :

* le rejet de **H0** si ,
* le non-rejet de **H0** si .

On peut résumer cette règle de décision par le schéma suivant :

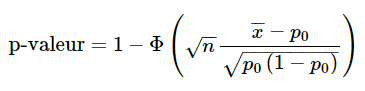


Remarquons que :

* La constante est positive : **c≥0** .  
  On retrouve notre considération introductive selon laquelle si , on ne rejette pas l'hypothèse nulle (et donc l'ancien médicament est conservé). Il en est encore ainsi si est "trop peu supérieur" à **p0**. C'est si  est "significativement supérieur" à **p0** au niveau de test **α**, d'au moins , que l'on rejette l'hypothèse nulle, le nouveau médicament semble alors préférable à l'ancien.
* La constante **c** décroît lorsque la taille de l'échantillon **n** croît .  
  A **α** fixé, une même proportion observée , jugée non significativement supérieure à **p0** si elle est issue d'un échantillon d'assez petite taille, le deviendra nécessairement pour des tailles suffisamment élevées, c'est-à-dire pour .
* La constante **c** décroît lorsque le niveau de test **α** croît  
  On sait que **Φ1−α** décroît lorsque **α** croît.  
  A **n** fixé, une même valeur de  peut conduire à rejeter l'hypothèse si **α** est assez fort (test "peu sévère") et ne pas la rejeter si **α** est assez faible (test "sévère").

#### Utiliser la p-valeur

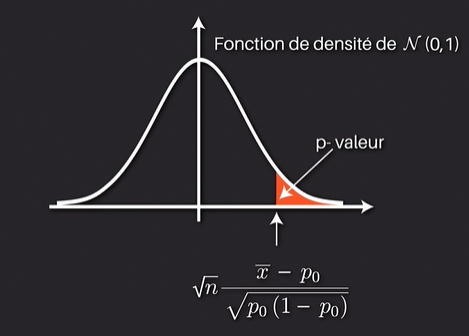
On peut mettre en évidence la p-valeur du test :



 La procédure de décision consiste donc à se fixer un niveau de test **α** et à calculer la p-valeur. Les décisions prises sont alors :

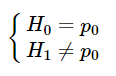
* le rejet de **H0** si **p-valeur<α** ,
* le non-rejet de **H0** si **p-valeur≥α**.

Le schéma suivant nous permet de visualiser la p-valeur :



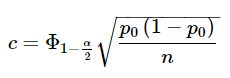
### Testez p=p0  versus p≠p0

Si on considère cette fois le test alternatif suivant :



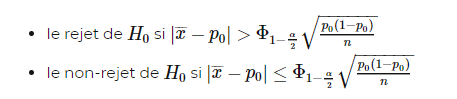
on choisit comme région critique   
On constate que la région critique intègre cette fois une valeur absolue afin de tenir compte du fait que **p≠p0** regroupe les cas **p≥p0** ET **p≤p0**.

La résolution mathématique conduit à :



On trouve ici **Φ1−α/2** à la place de **Φ1−α** dans le test unilatère précédent (celui du premier exemple introductif).

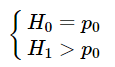
Les décisions prises sont, au niveau de test **α** :



Et il est encore possible de calculer la p-valeur.

### Le coin R : exemple du taux de guérison

On teste :

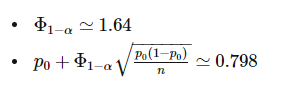


avec **p0=0.75** .

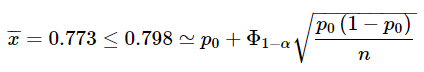
On considère que l’hypothèse gaussienne est acceptable ici, en effet :



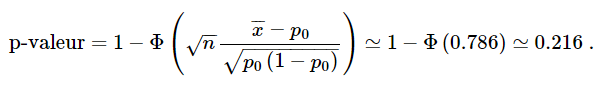
Pour **α=5** , on a :



On ne rejette pas **H0** au niveau de test 5% car :



La p-valeur vaut :



On constate que le niveau de test retenu devrait au moins être égal à 21,6% pour rejeter **H0** !

Accepter de se tromper dans plus de 20% des cas lorsque rejette l’hypothèse nulle n’est pas rencontré en pratique.

Au final, le laboratoire ne peut pas conclure avec un niveau de test raisonnable que le nouveau médicament est meilleur que celui déjà sur le marché.

Dans R, on utilise de nouveau la commande  prop.test  :

prop.test(x=167,n=216,p=0.75,alternative="greater")

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: p = 0.75

##

## Alternative Hypothesis: True p is greater than 0.75

##

## Test Name: 1-sample proportions test with continuity correction

##

## Estimated Parameter(s): p = 0.7731481

##

## Data: 167 out of 216, null probability 0.75

##

## Test Statistic: X-squared = 0.5

##

## Test Statistic Parameter: df = 1

##

## P-value: 0.2397501

##

## 95% Confidence Interval: LCL = 0.7206161

## Testez une moyenne ou une variance

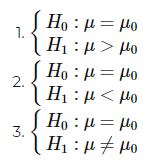
Les tests qui suivent ne seront pas abordés en détails. Les arguments sont identiques au chapitre précédent, seuls les calculs mathématiques diffèrent.

### Tester la moyenne (théorique) d'un échantillon

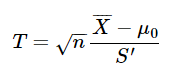
#### Le coin méthodologique

Supposons que nous ayons à disposition un échantillon i.i.d de loi , ou un grand échantillon i.i.d quelconque (en pratique de taille supérieure à 30).

On considère les tests suivants (c'est le test 3 qui est considéré dans le second exemple introductif) :



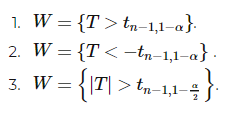
On pourrait utiliser  (ou comme statistique de test, malheureusement sa loi dépend du paramètre **σ** inconnu, c'est pourquoi on utilise :



On a sous **H0** :



Les régions critiques au niveau **α** sont :



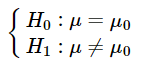
On constate bien ici que la forme de la région critique est donnée par l'hypothèse alternative **H1**.

Les p-valeurs sont :

#### 

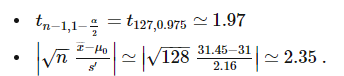
#### Le coin R : exemple de la consommation d'essence

On teste :

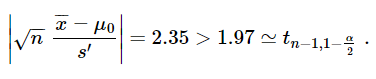


avec  **μ0=31**

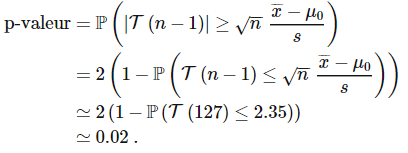
Pour **α=5** , on a :



On rejette **H0** au niveau de test 5% car :



La p-valeur vaut :

On constate qu’on rejette bien l’hypothèse nulle au niveau de test 5% (**0.02<0.05**) mais pas au niveau de test 1% (**0.02≥0.05**)

En pratique, le data analyst peut utiliser la commande  t.test  pour résoudre ce test :

alpha <- 0.05

t.test(essence$conso,mu=31,alternative="two.sided")

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: mean = 31

##

## Alternative Hypothesis: True mean is not equal to 31

##

## Test Name: One Sample t-test

##

## Estimated Parameter(s): mean of x = 31.44945

##

## Data: essence$conso

##

## Test Statistic: t = 2.354358

##

## Test Statistic Parameter: df = 127

##

## P-value: 0.02008833

##

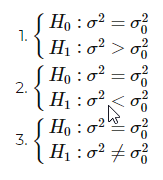
## 95% Confidence Interval: LCL = 31.07169

## UCL = 31.82722

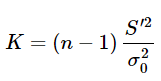
### Tester la variance (théorique) d'un échantillon

Supposons que nous ayons à disposition un échantillon i.i.d de loi , ou un grand échantillon i.i.d quelconque (en pratique de taille supérieure à 30).

On considère les tests suivants :



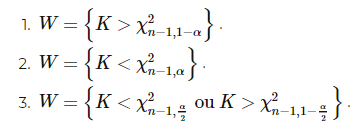
On utilise comme statistique de test :



On a sous **H0** :



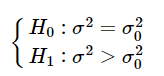
Les régions critiques au niveau **α** sont :



De là on tire les p-valeurs...

#### Le coin R : exemple de la consommation d'essence

Considérons par exemple le test suivant :



avec .

En pratique, le data analyst pourra de nouveau utiliser la commande  varTest  issue du package  EnvStats  pour obtenir cet intervalle de confiance :

library(EnvStats)

alpha <- 0.05

varTest(essence$conso,sigma.squared=4.5,alternative="greater")

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: variance = 4.5

##

## Alternative Hypothesis: True variance is greater than 4.5

##

## Test Name: Chi-Squared Test on Variance

##

## Estimated Parameter(s): variance = 4.66481

##

## Data: essence$conso

##

## Test Statistic: Chi-Squared = 131.6513

##

## Test Statistic Parameter: df = 127

##

## P-value: 0.3706697

##

## 95% Confidence Interval: LCL = 3.839436

## UCL = Inf

On ne rejette donc pas l’hypothèse nulle au niveau de test 5%, la p-valeur vaut en effet environ 0.37 (elle n’est évidemment pas inférieure à 0.05).

## Comparez deux échantillons gaussiens (test de comparaison)

### L'idée

Si on souhaite comparer deux échantillons (i.i.d) gaussiens, il nous suffit en fait de comparer leurs paramètres : leur moyenne **μ1** et **μ2**, et leur variance  et .

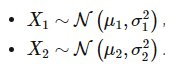
La méthodologie la plus classique est d'effectuer de manière séquentielle :

* Un test d'égalité des variances.
* Un test d'égalité des moyennes.

Si les variances ne sont pas considérées comme égales, les deux échantillons n'ont pas la même loi. Si les variances sont considérées comme égales, il est alors possible d'estimer cette variance sur les deux échantillons à la fois, et de tester l'égalité des moyennes en utilisant cette variance empirique globale.  
Notons qu'il est néanmoins possible d'effectuer un test de comparaison des moyennes sous hypothèse de variances différentes. Il ne s'agit pas d'une comparaison des lois, mais alors d'une comparaison simple des moyennes.

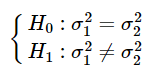
### Le coin méthodologique

Soient deux échantillons i.i.d gaussiens , indépendants entre eux :

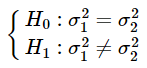


On considère :

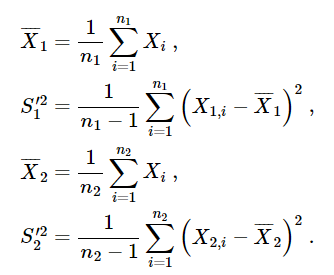
* Le test d'égalité des variances :



* Le test d'égalité des moyennes :

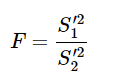


Pour effectuer les tests, on se base encore sur les moyennes et variances empiriques des deux échantillons :



#### Résolution du test d'égalité des variances

On utilise comme statistique de test :

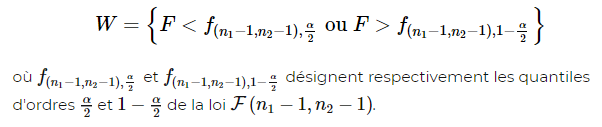


On peut montrer que sous **** :

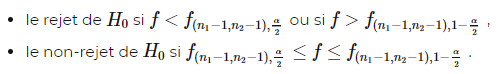


On comprend aisément qu'on rejettera d'autant plus facilement l'égalité des variances que ce ratio s'éloignera de 1.

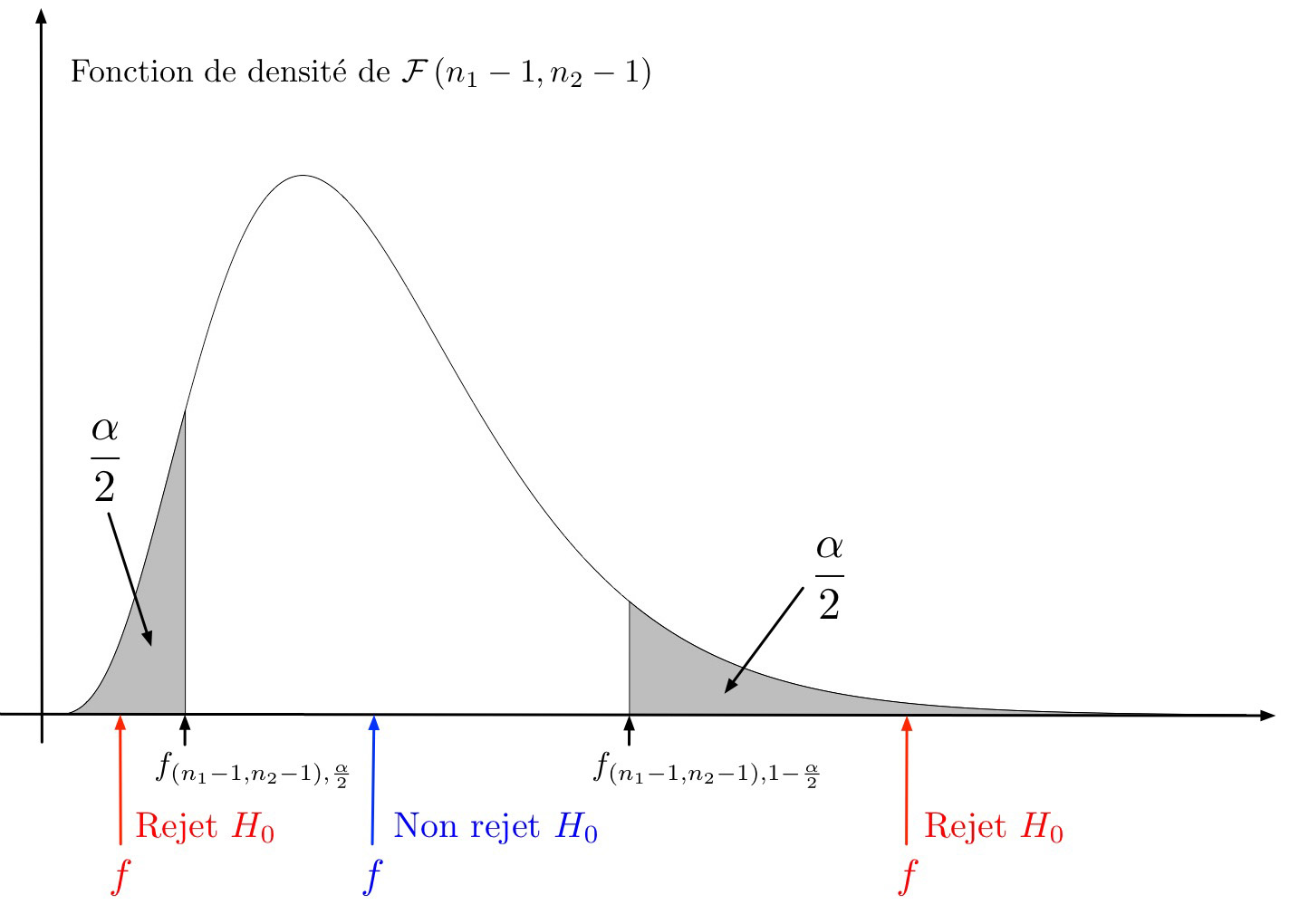
La région critique au niveau **α** est :



En pratique, les décisions prises sont, pour un niveau de test α :



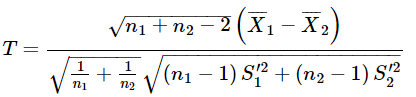
On peut résumer cette règle de décision par le schéma suivant :



#### Résolution du test d'égalité des moyennes

On est dans le cas où .

Pour des arguments du même type (non énoncés ici) que ceux évoqués lors du test sur la moyenne, on considère comme statistique de test :

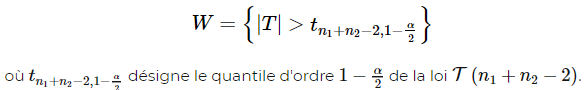


On constate que le numérateur fait apparaître l'écart entre les deux moyennes empiriques. Plus cet écart sera important plus on sera enclin de rejeter l'égalité des moyennes.

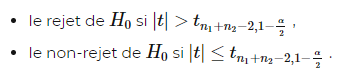
On peut montrer que sous **H0** :



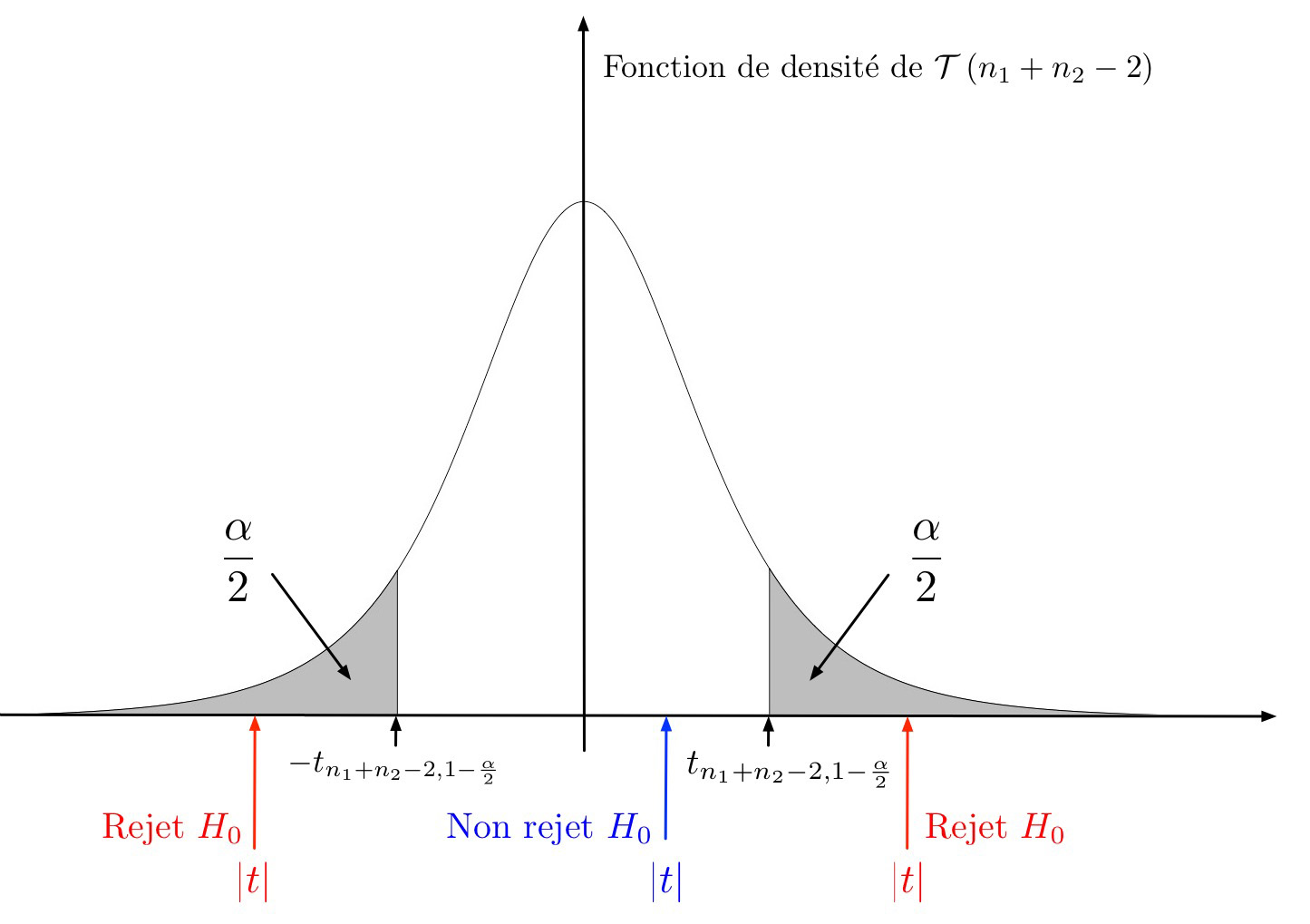
La région critique au niveau α est :



En pratique, les décisions prises sont, pour un niveau de test **α** :



On peut résumer cette règle de décision par le schéma suivant :



La p-valeur vaut :



### Le coin R : exemple des iris de Fisher

 Considérons dans le (fameux) jeu de données historique des [iris de Fisher](https://fr.wikipedia.org/wiki/Iris_de_Fisher) la longueur des pétales d’iris des variétés Versicolor et Virginica, sachant que nous disposons d’échantillons de 50 individus de chaque variété.

versi <- iris[iris$Species=="versicolor",]$Petal.Length

virgi <- iris[iris$Species=="virginica",]$Petal.Length

On teste tout d’abord l’égalité des variances à l’aide de la commande  var.test  :

var.test(versi,virgi)

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: ratio of variances = 1

##

## Alternative Hypothesis: True ratio of variances is not equal to 1

##

## Test Name: F test to compare two variances

##

## Estimated Parameter(s): ratio of variances = 0.7249678

##

## Data: versi and virgi

##

## Test Statistic: F = 0.7249678

##

## Test Statistic Parameters: num df = 49

## denom df = 49

##

## P-value: 0.2637454

##

## 95% Confidence Interval: LCL = 0.411402

## UCL = 1.277530

La p-valeur valant 0.26, on ne rejette donc pas l’égalité des variances au niveau de test 5%.

On teste ensuite l’égalité des moyennes à l’aide de la commande  t.test  :

t.test(versi,virgi,var.equal=TRUE)

##

## Results of Hypothesis Test

## --------------------------

##

## Null Hypothesis: difference in means = 0

##

## Alternative Hypothesis: True difference in means is not equal to 0

##

## Test Name: Two Sample t-test

##

## Estimated Parameter(s): mean of x = 4.260

## mean of y = 5.552

##

## Data: versi and virgi

##

## Test Statistic: t = -12.60378

##

## Test Statistic Parameter: df = 98

##

## P-value: 3.17882e-22

##

## 95% Confidence Interval: LCL = -1.495426

## UCL = -1.088574

On obtient une p-valeur égale à environ . On constate donc que l’hypothèse d’égalité des longueurs moyennes de pétales est (très facilement) rejetée à un niveau de test de 5%.

Au-delà de l’aspect gaussien que nous n’avions pas vérifié rigoureusement (on pourrait l’éprouver via des tests d’adéquation à une loi), on rejette que les longueurs des pétales d’iris des variétés et suivent la même distribution, on a en effet rejeté l’hypothèse d’égalité des moyennes.

### Pour aller plus loin

On distingue ici les échantillons :

* **appariés** : on applique deux traitements sur les mêmes individus et on compare leur effet sur deux échantillons qui ne sont donc pas indépendants.
* **non-appariés** : on compare deux échantillons indépendants.

On considère ici des tests non-paramétriques : on les appelle ainsi car ils ne dépendent pas des distributions des deux échantillons :

* Le **test de Wilcoxon** pour des échantillons appariés.
* Le **test de Mann et Whitney** pour des échantillons non-appariés.

Sous R, on peut utiliser la commande  wilcox.test  avec l'option  paired=TRUE  pour les échantillons appariés ou  paired=FALSE  pour les échantillons non-appariés.

## Découvrez les tests d'adéquation : le Khi-deux et Kolmogorov Smirnov

### L'objectif

Dans les travaux de modélisation que le data analyst sera amené à traiter, il y a aura régulièrement des hypothèses sur des lois de probabilité qu'il lui faudra vérifier.

Il existe de nombreux tests pour vérifier qu'un échantillon suit ou non une loi de probabilité donnée, on en donne ici deux représentants, un dans le cas discret, le test dit du Khi-deux, et un dans le cas continu, le test de Kolmogorov Smirnov.

Si on reprend nos deux exemples introductifs, la loi de probabilité était acquise dans le premier cas (loi de Bernoulli) tandis que dans le second cas on a supposé que l'échantillon était potentiellement gaussien. Cette dernière hypothèse peut tout à fait être testée ; si Kolmogorov-Smirnov est utilisable (et présenté ici car plutôt intuitif), il nous faut reconnaître que l'on utilise souvent d'autres tests comme celui de [Shapiro-Wilk](https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk) (non présenté ici).

### Le test du Khi-deux

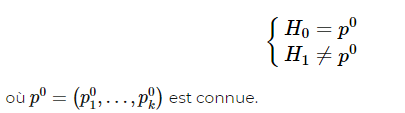
#### L'idée

On considère un échantillon dont la loi de probabilité est discrète, avec un nombre de valeurs potentielles finies (si la loi n'est pas de support borné, il suffit de regrouper les valeurs en queue, de probabilité plus faible).  
Prenons un exemple simple, le lancer d'un dé à 6 faces. Pour savoir s'il est équilibré, il nous faudrait idéalement vérifier que la probabilité d'obtenir la face 1 vaut 1/6, ... , la probabilité d'obtenir la face 6 vaut 1/6. Le test du Khi-deux nous simplifie la tâche car il permet de regrouper ces six tests élémentaires. Pour cela on considère une (pseudo-)distance entre les probabilités estimées (que sont les fréquences empiriques) et les valeurs testées : plus cette quantité sera grande, plus on sera enclin à rejeter l'hypothèse que le dé est équilibré.

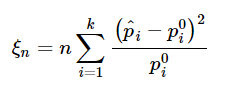
#### Le test

On suppose qu'on dispose de **n** variables aléatoires **Xi** discrètes, à valeurs dans**{a1,…,ak}**, indépendantes et de même loi caractérisée par**p=(p1,…,pk)** où v.

On souhaite tester :

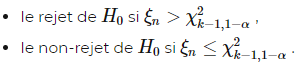


On utilise à cet effet le test du Khi-deux avec comme statistique :



est égal à la fréquence empirique de la valeur **ai**.

La loi asymptotique de **ξn** sous **H0** est connue, il s'agit d'une loi du Khi-deux à **k−1** degrés de liberté. On prend donc comme décision, pour un niveau de test **α** :



#### Le coin R : exemple des dés de Weldon

Weldon a réalisé   n=26306n=26306 lancers de 12 dés à 6 faces et a comptabilisé, à chaque lancer, le nombre de dés indiquant soit 5 soit 6. Il a obtenu les résultats suivants :

| Nb de 5 ou 6 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Effectif | 185 | 1149 | 3265 | 5475 | 6114 | 5194 | 3067 | 1331 | 403 | 105 | 14 | 4 | 0 |

Soit **Xi** la v.a.r. à valeurs dans  {0,…,12} désignant le nombre de faces comportant un cinq ou un six lors du i-ème lancer.

Dans le cas où les dés sont équilibrés : Xi∼B(1/3).

On teste si les dés sont équilibrés :

n <- c(185,1149,3265,5475,6114,5194,3067,1331,403,105,14,4,0)

p0 <- dbinom(0:12,12,1/3)

chisq.test(n,p=p0)

## Warning in chisq.test(n, p = p0): Chi-squared approximation may be

## incorrect

##

## Results of Hypothesis Test

## --------------------------

##

## Alternative Hypothesis:

##

## Test Name: Chi-squared test for given probabilities

##

## Data: n

##

## Test Statistic: X-squared = 41.31222

##

## Test Statistic Parameter: df = 12

##

## P-value: 4.344864e-05

Le test n’est pas valide car les effectifs ne sont pas suffisants pour les valeurs 10, 11 et 12. On agrège ces modalités et on relance les calculs :

n <- c(185,1149,3265,5475,6114,5194,3067,1331,403,105,18)

p0 <- c(dbinom(0:9,12,1/3),sum(dbinom(10:12,12,1/3)))

chisq.test(n,p=p0)

##

## Results of Hypothesis Test

## --------------------------

##

## Alternative Hypothesis:

##

## Test Name: Chi-squared test for given probabilities

##

## Data: n

##

## Test Statistic: X-squared = 35.4943

##

## Test Statistic Parameter: df = 10

##

## P-value: 0.0001027878

On rejette très largement l’hypothèse de dés équilibrés au niveau de test 5%.

### Le test de Kolmogorov Smirnov

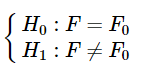
#### L'idée

On considère un échantillon dont la loi de probabilité est continue.  
Revenons à notre exemple de la consommation d'essence. Pour évaluer si cet échantillon peut être considéré comme gaussien, on peut étudier l'écart entre la fonction de répartition d'une loi normale et celle estimée de notre échantillon : la fonction de répartition empirique !  
Plus cette quantité est grande, plus on est enclin à rejeter l'hypothèse comme quoi l'échantillon est gaussien. Le test du Kolmogorov-Smirnov, qui permet de tester toute loi de probabilité, repose sur l'écart maximum observé entre les deux fonctions de répartition, l'empirique et la théorique testée. La loi de cette statistique est toujours la même quelle que soit la loi testée, et est connue asymptotiquement : on peut donc réaliser sans souci un test d'adéquation.

#### Le test

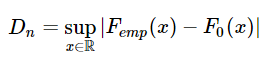
Soit (X1,…,Xn) un échantillon i.i.d de même loi que X admettant F comme fonction de répartition.

On souhaite tester :

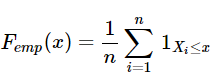


où **F0** est connue.

On utilise à cet effet le test de Kolmogorov-Smirnov avec comme statistique :



où **Femp** est la fonction de répartition empirique de l'échantillon :



La statistique de test **Dn** est basée sur la distance maximale entre la fonction de répartition empirique de l'échantillon, estimation de sa fonction de répartition, et **F0**.

La loi asymptotique de **Dn** sous **H0** est connue. Son équation, pas très explicite, n'est pas donnée ici ; elle est néanmoins implémentée dans les logiciels et permet de déterminer la p-valeur, utilisée en pratique pour la prise de décision.

#### Le coin R : exemple de la consommation d'essence

On peut tester l’adéquation de la consommation d’essence à une loi normale à l’aide de Kolmogorov-Smirnov :

ks.test(essence$conso,"pnorm",mean=mean(essence$conso),sd=sd(essence$conso))

## Warning in ks.test(essence$conso, "pnorm", mean = mean(essence$conso), sd

## = sd(essence$conso)): ties should not be present for the Kolmogorov-Smirnov

## test

##

## Results of Hypothesis Test

## --------------------------

##

## Alternative Hypothesis: two-sided

##

## Test Name: One-sample Kolmogorov-Smirnov test

##

## Data: essence$conso

##

## Test Statistic: D = 0.04830966

##

## P-value: 0.9262283

On ne peut donc pas rejeter l’hypothèse de normalité au niveau de test 5%.

On aurait également pu lancer un test de Shapiro-Wilk :

shapiro.test(essence$conso)

##

## Results of Hypothesis Test

## --------------------------

##

## Alternative Hypothesis:

##

## Test Name: Shapiro-Wilk normality test

##

## Data: essence$conso

##

## Test Statistic: W = 0.9871523

##

## P-value: 0.2743285

### Résumons !

Au cours de cette partie 4, nous avons vu beaucoup de tests statistiques différents. Voici un petit résumé des différents types de tests que nous avons vu :

* Tests sur les paramètres de la loi d'un échantillon :
  + sur une proportion,
  + sur une moyenne,
  + sur une variance.
* Test de comparaison de deux échantillons (gaussiens).
* Tests d'adéquation à une loi :
  + dans le cas discret : test du Khi-deux.
  + dans le cas continu : test de Kolmogorov-Smirnov.

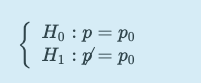
## Entraînez-vous à tester l'équiprobabilité des naissances de femmes et d'hommes

### À vous de jouer

Pour vous entraîner, réalisez cet exercice étape par étape. Une fois terminé, vous pouvez comparer votre travail avec les pistes que je vous propose.

Le nombre de naissances en France en 2014 est de 399 284 hommes contre 381 883 femmes (source : INED).

On souhaite tester l’équiprobabilité des naissances de femmes et d'homme :



avec p0 = 0.5

Questions

* Question 1

Réaliser ce test d’équiprobabilité, et donnez la P-valeur.

* Question 2

Accepte-t-on l'hypothèse d’équiprobabilité des naissances femmes-hommes au niveau de test 5% ?

* Question 3

Supposons maintenant que la taille de l'échantillon soit différente, mais que les fréquences femmes-hommes soient les mêmes.

Calculez le nombre de femmes et d'hommes dans chacun de ces 3 cas :

* échantillon de taille n = 100
* échantillon de taille n = 1000
* échantillon de taille n = 10000
* Question 4

Pour chacun des 3 cas précédents, accepte-t-on l'hypothèse d’équiprobabilité des naissances femmes-hommes au niveau de test 5% ?

Vous répondrez aux questions grâce au langage R ou au Python.

Vérifiez-bien que vous avez les éléments suivants :

* la p-value est égale à 2.2 \* 10^-16 (une marge d'erreur de 10^-14 est admise) ;
* vous rejetez bien l'hypothèse d’équiprobabilité des naissances hommes-femmes au niveau de test 5%, la p-valeur 2.2 \* 10^-16 valeur étant (très largement) inférieure à 5%.
* toutes les valeurs sont correctes (à 1 individu près) :

|  |  |  |
| --- | --- | --- |
| taille de l'échantillon | femmes | hommes |
| 100 | 49 | 51 |
| 1000 | 489 | 511 |
| 10000 | 4889 | 5111 |

* vous rejettez l'hypothèse d'équiprobabilité au niveau de test 5% pour l'échantillon de taille 10000, et ne la rejette pas pour les échantillons de taille 1000 et 100.

# Savez-vous effectuer un test statistique ?

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Comprendre la notion d’inférence
* Etablir un lien entre des observations et un modèle probabiliste
* Tester une hypothèse sur une proportion, une moyenne ou une variance
* Tester l’adéquation à une loi par un test statistique

### Question 1

**Dans un test statistique, l’hypothèse nulle :**

* + 

se choisit de manière indifférente

* + 

désigne l’hypothèse qu’on souhaite rejeter à tout prix

* + 

est un a priori qu’on ne souhaite pas rejeter trop facilement

*C’est notre hypothèse de base !*

### Question 2

**On vous indique dans un test que la p-valeur vaut 0.03 (3%) :**

* + 

on peut rejeter l’hypothèse nulle uniquement pour un niveau de test égal à 3%

* + 

on peut rejeter l’hypothèse nulle pour un niveau de test égal à 5% mais pas 1%

* + 

on peut rejeter l’hypothèse nulle pour un niveau de test égal à 1% mais pas 5%

*La p-valeur est inférieure à 5% mais pas à 1%.*

### Question 3

**Le niveau de test correspond :**

* + 

au risque maximum de se tromper en général

* + 

au risque maximum de rejeter l’hypothèse nulle à tort

* + 

au risque maximum de rejeter l’hypothèse alternative à tort

### Question 4

**Pour tester l’adéquation à une loi normale, vous pouvez naturellement utiliser :**

* + 

le test de Kolmgorov-Smirnov

* + 

le test du Khi-deux

* + 

le test de comparaison de deux échantillons gaussiens

### Question 5

**Pour comparer deux échantillons (i.i.d) gaussiens, la méthodologie la plus classique est d'effectuer de manière séquentielle :**

* + Un test d'égalité des variances.
  + Un test d'égalité des moyennes.
  + 

Cette phrase est vraie

* + 

Cette phrase est fausse

*Si on souhaite comparer deux échantillons (i.i.d) gaussiens, il suffit en fait de comparer leurs paramètres : leur moyenne μ1 et μ2, et leur variance*σ21σ21*et*σ22σ22*.*

## Conclusion

Nous avons vu, tout au long de ce cours, les bases de la statistique inférentielle, dans le cadre de lois de probabilités paramétrées (dépendantes d'un paramètre θ), qui nous ont permis notamment :

* d'estimer le paramètre d'une loi afin de la déterminer entièrement,
* d'établir un intervalle de confiance sur le paramètre,
* et enfin de tester un paramètre.

Ces notions, pouvant apparaître parfois comme un peu abstraites, nous fournissent en fait des outils puissants d'aide à la décision. Au-delà des cas réalistes rencontrés ici, ces outils s'avèrent être extrêmement utiles (et utilisés !) dans le cadre de la modélisation statistique, notamment :

* dans les très utilisés modèles linéaires gaussiens,
* et dans les modèles ARMA applicables sur des séries temporelles.