



# STATISTIQUES :

## Tests d'hypothèse : détection d'un son



Crédit : Jurassic Park

## Objectifs du module

**Utiliser les outils statistiques standards permettant de réaliser un test d'hypothèse, et réaliser ce test.**

Un test d'hypothèse statistique est la vérification d'une hypothèse portant sur une variable aléatoire, sur la base d'observations (mesures) de ladite variable. La conclusion du test sera le rejet ou le non-rejet de l'hypothèse. Dès lors qu'on tire une conclusion d'un jeu de donnée, il est nécessaire d'associer un critère de significativité statistique à cette conclusion.

Il existe une multitude de tests d'hypothèse en statistique. On peut par exemple comparer 2 échantillons afin de déterminer s'ils sont statistiquement significativement différents. On peut aussi comparer un jeu de données à un modèle statistique censé décrire le comportement d'une variable aléatoire.

L'objectif de ce module est triple :

- Que vous vous formiez à la construction de lois de probabilités (probability distributions).
- Que vous appréhendiez via un exemple les caractéristiques systématiquement présentes dans un test d'hypothèse. Et ce de manière à ce que vous soyez en mesure de vous approprier dans le futur les tests d'hypothèse que vous pourrez avoir à réaliser.
- Si vous avez le temps, de méditer plus en profondeur sur ce qu'on peut dire d'un test d'hypothèse et ce que l'on ne peut/doit pas.

## Modalités

- Durée du projet : 3 jours.
- Ce projet sera réalisé en autonomie
- Mais l'entraide est très fortement encouragée !
- Vous produirez un jupyter notebook contenant la résolution/réponse aux questions du sujet.

## Description du projet

Pour les 3 journées de cours consacrés à ce projet, nous allons nous intéresser à la détection de signal sonore. Cela nous amènera en première itération à réaliser une étude

sur un bruit. En seconde itération, nous réaliserons une étude sur un son. Puis en troisième et dernière itération nous nous intéresserons à la détection d'un signal sonore présent au milieu du bruit. Cela constituera votre test d'hypothèse : pour réaliser une détection, nous allons devoir faire une hypothèse qui nous permettra de décider si oui ou non, un signal sonore est détecté.

Vous aborderez les notions suivantes :

- Lois de probabilités en statistique (continues et discrètes). Probability distributions.
- Éléments de probabilités
- Tests d'hypothèses (choix d'une hypothèse nulle, d'une hypothèse alternative, critère de significativité et p-value)

Si cela remonte à loin, vous aurez besoin de vous rafraîchir la mémoire sur certaines notions :

- Les bases du calcul différentiel et intégral (Qu'est ce qu'une dérivée ? Une intégrale ? etc). Vous n'aurez pas besoin de calculer des intégrales analytiquement mais il sera important de comprendre ces concepts pour avancer
- Les bases de probabilités.

**Sujet :** Depuis quelque temps, bien que vivant dans un endroit très calme la nuit, vous vous demandez si des sons ne vous dérangent pas pendant votre sommeil. Bien décidé à avoir le coeur net et plutôt que de vous demander à chaque réveil si celui-ci est dû à son quelconque ou juste à une mauvaise qualité de sommeil, vous décidez de réaliser un enregistrement sonore d'une nuit, et de créer un test qui permettra de détecter les sons qui se seront produits durant cette nuit.

Tous les enregistrements sont disponibles avec le sujet, chacun en double :

- un au format ".wav" (ça serait dommage de ne pas utiliser vos oreilles alors que vous travaillez sur des signaux sonores)
- l'autre et en format .txt.

Chaque enregistrement est **échantillonné à 48 kHz**.

Pour éviter de perdre du temps, la variable "signal sonore" contenue dans chaque fichier .txt est déjà normalisée, pour avoir des amplitudes de signaux n'excédant pas quelques unités. Vous n'aurez donc pas besoin de modifier vos données pour travailler dessus.

# Itération 1

## Etudes du bruit et d'un son

### Objectifs de l'activité

- Que vous étudiez 2 jeux de données correspondant chacun à une variable aléatoire continue.

### Compétences

- Construire les fonctions “probability density function” et “cumulative distribution function” d'une variable aléatoire continue.

### Information importantes

- Toutes les lois de probabilités porteront sur la **valeur absolue** de l'amplitude mesurée. De manière générale, par défaut toute référence à une amplitude renvoie à la valeur absolue de l'amplitude mesurée. Le cas contraire vous sera explicité. L'abréviation **va** renvoie à la valeur absolue, et **vr** à la valeur réelle.

### Consignes :

#### A. Etude du bruit

Avant de commencer à réfléchir à un quelconque test de détection de son, vous décidez d'étudier le bruit. Le bruit c'est en fait tout ce qui ne vous intéresse pas, un ensemble de sons qui s'additionnent pour donner une fluctuation très faiblement audible et qui ne vous dérange pas pendant votre sommeil (vous vivez dans un endroit calme). Ce qui vous intéresse in fine, ce sont les sons (on distingue donc bien les notions de son et de bruit). Mais pour espérer les détecter des **sons**, il vous faut d'abord caractériser **le bruit**.

**A.0** - Écouter le bruit en ouvrant le fichier “bruit.wav”

**A.1** - Importer le bruit “bruit\_seul.txt” dans un array numpy

**A.2** - Tracer quelque plots amplitude\_bruit(temps) pour quelque intervalles de votre choix. En **vr** pour cette question.

**A.3** - Calculez la “probability density function” (pdf) et la “cumulative distribution function” (cdf) de l'amplitude (en **va**) et les afficher

**A.4** - Caractériser analytiquement les 2 lois de probabilités calculées question précédentes.

**A.5** - Évaluer la probabilité de détection du bruit pour différentes plages d'amplitudes de détection.

## B. Etude d'un son

Vous avez réalisé un enregistrement pendant la nuit. En inspectant rapidement la courbe amplitude(temps), il vous semble voir quelque chose de particulier, et vous soupçonnez fortement qu'il ne s'agisse pas du bruit mais d'un son (une tentative d'effraction ?). Vous trouverez l'extrait en question dans le fichier "**son.wav**".

**B.0** - Écouter l'extrait

**B.1** - Importer le fichier "**son.txt**" dans un array numpy et faire un plot amplitude(temps) comparant une zone exclusivement bruitée et une zone de son. En **vr** pour cette question.

**B.2** - Identifier approximativement la zone de son et réaliser la pdf et la cdf de l'amplitude du son (en **va**).

**B.3** - Identifier pour quelques seuils d'amplitude (en **va**) les moments où l'amplitude enregistrée est supérieure à ce seuil. Conclure.

### Livrables

- Exercices terminés dans le jupyter notebook

# Itération 2

## Test d'hypothèse - détection d'un son

Plutôt que de faire des tests sur les occurrences d'amplitude supérieure à une valeur seuil, vous décidez de faire une étude similaire portant sur le nombre d'occurrences durant un certain temps. De cette manière, si le bruit atteint ponctuellement une amplitude importante, on peut se douter qu'il a très peu de chance de réitérer cet "exploit" beaucoup de fois dans un intervalle de temps court, contrairement au son qui lui devrait l'atteindre plus de fois sur un intervalle de même durée. Tout l'enjeu sera de quantifier cette "chance" (ie cette probabilité), pour déterminer le "niveau de confiance" du test.

### A. Un peu de probabilités

- On note  $a_l$  une valeur d'amplitude. La probabilité d'obtenir sur une unique mesure une amplitude  $A$  telle que  $A > a_l$  est notée  $P_{>a_l}$
- On note  $N_{ech}$  le nombre d'amplitudes mesurées sur un intervalle (échantillon).
- On note  $k$  un entier positif inférieur ou égal à  $N_{ech}$

**A.1** - Trouver  $P_k^{N_{ech}}$  la probabilité de tirer  $k$  fois une amplitude supérieure à  $a_l$  lors de  $N_{ech}$  mesures

**A.2** - Identifier la fonction python et ses arguments permettant d'obtenir  $P_k^{N_{ech}}$

**A.2** - En posant  $N_{ech} = 20$

**A.2.a** - Calculez plusieurs fois, chacune pour une valeur  $P_{>a_l}$  de votre choix

**A.2.b** - Affichez en une sur un plot en barres

**A.2.c** - Affichez les toutes sur un même graphe

### B. Construction et réalisation de tests d'hypothèse

Vous voulez donc construire votre test d'hypothèse. Vous vous appuyez pour ce test sur la loi binomiale étudiée partie A. Pour cela, vous avez besoin de définir une hypothèse nulle, une hypothèse alternative, ainsi qu'un critère de significativité.

**B.1** - Caractériser l'hypothèse nulle  $H_0$

**B.2** - Caractériser l'hypothèse alternative  $H_1$  (ou  $H_a$  selon les goûts)

**B.3** - Mesure/statistique :

**B.3.a** - Identifier l'élément mesurable (ie la statistique) qui permettra de réaliser le test

**B.3.b** - Rappeler la loi de probabilité de cet élément mesurable (ie cette statistique) "sous"  $H_0$

**B.3.c** - Identifier la p-value d'un test.

**B.4** - Définir le seuil de "significativité"  $\alpha_{sig}$

**B.4.a** - Pour un seuil  $\alpha_{sig}$ , identifier dans quelles situation  $H_0$  sera rejetée, ou bien non-rejetée.

**B.5** - Réaliser des tests pour différents choix de mesure/statistique. Pour chacune des statistiques choisies :

**B.5.a** - Calculer la loi de probabilité de la mesure/statistique "sous"  $H_0$

**B.5.b** - Calculer la valeur de la mesure/statistique

**B.5.c** - Calculer la "p-value" de votre test. A quoi correspond-elle ?

**B.5.d** - Prendre la décision concernant l'hypothèse posée.

**B.5.e** - Estimer le nombre de "faux positifs"

**B.6** - Regarder la vidéo : <https://www.youtube.com/watch?v=jy9b1HSqtSk> puis répondre par vrai ou faux à chacune des questions suivantes

**B.6.a** - Pour chaque détection (ie  $H_0$  rejetée) j'ai une probabilité  $1 - p_{val}$  qu'il s'agisse d'une "vraie détection"

**B.6.b** - En supposant  $H_0$  vraie, j'ai une probabilité  $\alpha_{sig}$  de conclure le rejet de  $H_0$