III – Inférence et décision

On s'intéresse dans ce TP à la population de brochets présente dans la Deûle [1].

A – Estimation paramétrique

On dispose, dans le fichier `samples.mat`, de deux jeux de données X et Y relatifs à la taille (en cm) des brochets au sein d'échantillons prelevés en 2012 et 2022, respectivement.

```
load samples.mat
hist(x)            % 2012
```

Il semble raisonnable d'après le théorème central limite de supposer que la variable X suit (du moins approximativement) une loi normale $\mathcal{N}(\mu, \sigma^2)$, dont nous allons maintenant estimer les paramètres.

1) Intervalle de confiance pour μ

Pour un échantillon $(X_1, X_2, \ldots, X_n)$ indépendant tiré d'une loi $\mathcal{N}(\mu, \sigma^2)$, nous savons que la moyenne échantillonnale

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

est une variable aléatoire distribuée selon une $\mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big)$, ce qui nous donne après normalisation

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

Remplaçant σ par la valeur S_X fournie par l'estimateur sans biais de la variance,

$$S_X^2 := \frac{1}{n-1} \sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2,$$

on obtient une variable aléatoire

$$Z := \frac{\overline{X} - \mu}{S_X/\sqrt{n}} \quad \text{approximativement} \quad \mathcal{N}(0, 1).$$

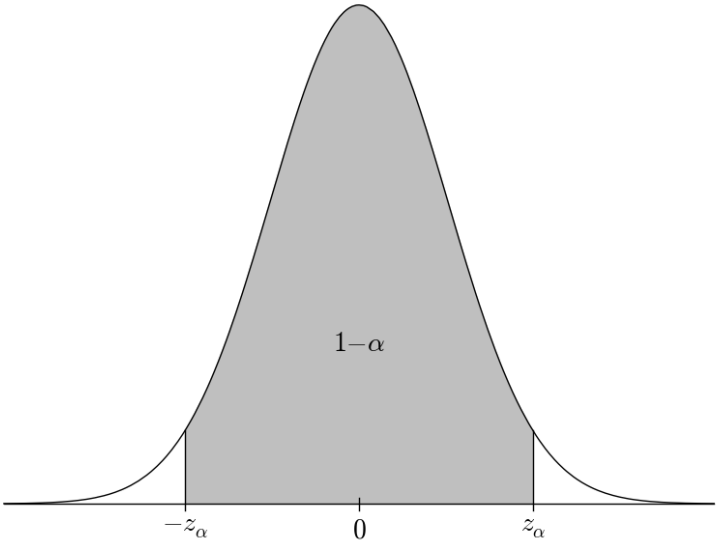
1. données fictives

Pour obtenir un intervalle de confiance au niveau $1-\alpha$ pour μ, soit z_α la valeur pour laquelle

$$\mathbb{P}[-z_\alpha \leqslant Z \leqslant z_\alpha] = 1-\alpha,$$

i.e. telle que

$$F_Z(z_\alpha) = 1 - \frac{\alpha}{2}.$$



On peut alors vérifier (n'est-ce pas ?) que

$$I_\alpha := \left[\overline{X} - z_\alpha \frac{S_X}{\sqrt{n}},\ \overline{X} + z_\alpha \frac{S_X}{\sqrt{n}}\right]$$

est un intervalle (aléatoire) contenant μ avec probabilité $1-\alpha$.

a) Commencez pas simuler le tirage de plusieurs échantillons pour bien sentir comment les réalisations de ces intervalles se comportent : la commande `sim_I(mu, sigma, alpha, n, m)` fournie par le fichier `sim_I.m` permet d'observer les intervalles de confiance au niveau $1-\alpha$ obtenus pour m échantillons de taille n tirés d'une $\mathcal{N}(\mu, \sigma^2)$. Faites varier les paramètres et consignez vos observations (observer notamment le difficile compromis à trouver entre *précision* et *confiance*).

b) Déterminer maintenant l'intervalle de confiance pour la valeur conventionnelle $\alpha = 5\,\%$ obtenu avec notre échantillon :

```
n = size(x,1)    % 100
xbar = mean(x)   % moyenne échantillonnale
s = std(x)       % écart-type échantillonnal, normalisé avec n-1

alpha = .05
z = norminv(1-alpha/2)

I = [xbar - z*s/sqrt(n), xbar + z*s/sqrt(n)]
```

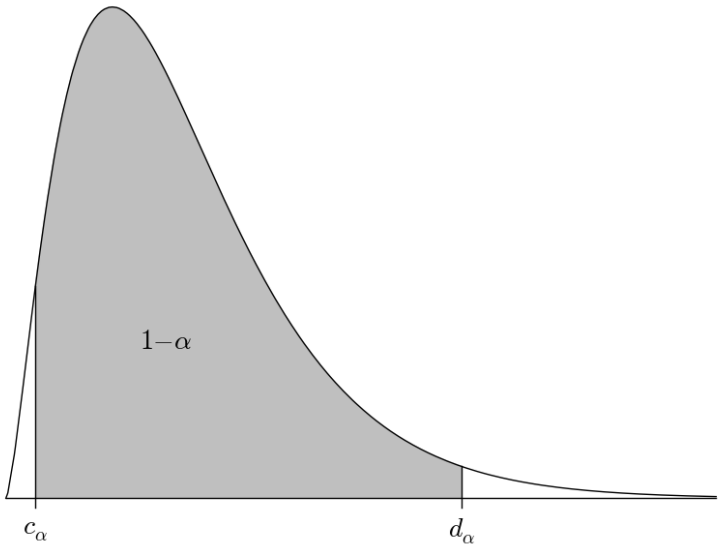
Quelle interprétation peut-on faire de cet intervalle ? Expérimenter avec différentes valeurs de α.

2) Intervalle de confiance pour σ

Pour obtenir un intervalle de confiance pour σ, on procède de façon similaire, en utilisant la loi de notre estimateur :

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Pour un intervalle de confiance de niveau α, nous devons utiliser des quantiles pour une loi du chi-carré.




```
alpha = .05

c = chi2inv(alpha/2, n-1)
d = chi2inv(1-alpha/2, n-1)
```

On a alors

$$\mathbb{P}\left[c_\alpha \leqslant \frac{(n-1)S_X^2}{\sigma^2} \leqslant d_\alpha\right] = 1-\alpha,$$

d'où on conclut que

$$J_\alpha := \left[\sqrt{\frac{(n-1)}{d_\alpha}}S_X, \sqrt{\frac{(n-1)}{c_\alpha}}S_X\right]$$

est un intervalle de confiance au niveau $1-\alpha$ pour σ.

a) Créer une commande `sim_J` adaptée de `sim_I` permettant de simuler des réalisations de cet intervalle J_α. Quels sont les comportements semblables et différents par rapport à I_α ?

b) Calculer les bornes de cet intervalle pour notre échantillon et en déduire un intervalle de confiance pour σ. Comment se comporte-t-il lorsque l'on modifie α ?

B – Tests d'hypothèses

Au cours des dernières années, diverses mesures ont été prises pour tenter de réhabiliter la faune et la flore du cours d'eau, et la MEL aimerait savoir si celles-ci ont eu un impact positif sur la taille des brochets. Quelques brochets ont donc été mesurés récemment, on trouve les résultats dans la variable `y` stockée dans `samples.mat`.

```
hist(y)  % 2022
```

Ils ont effectivement *l'air* un peu plus grands… Mais comment faire pour savoir si cette différence n'est due qu'à l'aléa du tirage de l'échantillon, ou si, au contraire, elle témoigne d'une réelle évolution de la taille moyenne μ dans la population ?

1) Égalité des espérances

Nous allons procéder à un test d'hypothèse. Il est encore vraisemblable de supposer que Y suit une loi normale, disons $Y \sim \mathcal{N}(\mu', \sigma'^2)$. Pour simplifier, supposons pour l'instant que l'écart-type n'a pas changé : $\sigma' = \sigma$ (nous verrons plus bas si cette simplification est justifiée).

La question que l'on se pose est donc : « est-ce que $\mu' > \mu$? »

Pour formuler cela sous forme de test d'hypothèse, on prend comme hypothèse nulle (conservatrice, mais permettant des prédictions) :

$$H_0 : \mu' = \mu$$

et comme hypothèse alternative :

$$H_1 : \mu' > \mu.$$

On va se demander à quel point les données observées sont incompatibles avec H_0 ; si elles le sont *trop*, on rejettera celle-ci pour croire plutôt à H_1. Si elles ne le sont *pas trop*, il n'y a pas de raison de rejeter H_0 ; les déviations au modèle donné par H_0 ne sont pas alors jugées *statistiquement significatives*.

Si H_0 est vraie, alors $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$ sont toutes i.i.d. selon la même loi $\mathcal{N}(\mu, \sigma^2)$. On a alors

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{et} \quad \overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right),$$

d'où pour la différence des moyennes

$$\overline{Y} - \overline{X} \ \sim \ \mathcal{N}\left(0, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right)$$

(pourquoi ?), donc

$$\frac{\overline{Y} - \overline{X}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\,\sigma} \ \sim \ \mathcal{N}(0, 1).$$

Comme précédemment, nous avons besoin d'une estimation de σ ; or, sous nos hypothèses, nous disposons de plus de données qu'avant pour estimer celui-ci, on remplacera donc σ par S donné par

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}.$$

Bref, *sous* H_0, la statistique

$$Z := \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\,S} \quad \text{est approximativement} \quad \mathcal{N}(0, 1).$$

On se donne comme précédemment – *avant de voir les données* – un seuil de décision, disons encore :

```
alpha = .05
```

puis on se pose la question : quelle est la probabilité [2], si H_0 est vraie, d'observer des valeurs de Z

2. souvent appelée valeur p

aussi (ou plus) extrêmes que celle que l'on a ?

Notre critère de décision est alors :

- si $p < \alpha$, on rejette H_0 en faveur de H_1 ;
- si $p \geqslant \alpha$, on conserve H_0.

Calculer la valeur p pour nos jeux de données :

```
m = size(y,1)  % 20

s = sqrt( ((n-1)*var(x) + (m-1)*var(y))/(m+n-2) )

z = (mean(y) - mean(x)) / s / sqrt(1/m + 1/n)

p = 1 - normcdf(z)
```

Quelle est votre conclusion ?

2) Égalité des variances

Reste que cette conclusion n'est valable que sous l'hypothèse, faite pour simplifier les calculs, que $\sigma' = \sigma$. Effectuons un autre test d'hypothèse pour déterminer si celle-ci est raisonnable :

$$\begin{cases} H_0 : \sigma' = \sigma \\ H_1 : \sigma' \neq \sigma \end{cases}$$

(cette fois, il n'y a aucune raison de privilégier une inégalité dans un sens plutôt que dans l'autre pour l'hypothèse alternative).

Sous H_0, le quotient

$$Q := \frac{S_X^2}{S_Y^2}$$

suit une loi connue de MATLAB sous le nom de *distribution F à $n-1$ et $m-1$ degrés de libertés*.

Implémenter ce test d'hypothèse ; la déviation à H_0 est-elle statistiquement significative ?

Notez que puisque l'on ne sait pas à l'avance si la valeur observée `q` de Q sera grande ou petite, la bonne notion de valeur p est ici

`p = 2*min( fcdf(q,n-1,m-1), 1 - fcdf(q,n-1,m-1) )` .

Vous pouvez observez où la valeur échantillonnale de `q` tombe par rapport à la densité de la variable aléatoire Q qu'on obtient sous H_0 par un `plot` de

$t \mapsto$ `fpdf(` t `, n-1, m-1 )` .