PIERRE_24

Geste de savoir

Comptez avec moi les retours !

vendredi 09 août 2024

Table des matières

Introduction

Chaque année depuis trois ans, j'aide à organiser une marche ADEPS pour le compte de l'association caritative dont je fais partie et dont je suis par ailleurs le *webmaster* à mes heures perdues. Vu que notre but est de, grâce à cet événement, renflouer les caisses de l'association (parce que malheureusement, les dons n'y suffisent pas), la question est de savoir quand est-ce que nos marcheurs reviennent de leur promenade, et s'ils viennent alors consommer dans notre buvette.

Et bien que la question parait simple, la réponse demande de faire un peu de statistiques. Bref, ce billet est l'occasion pour vous comme pour moi de réviser ces statistiques en cette période estivale

Un peu de contexte

L'ADEPS est un service public belge (et plus précisément wallon, fédéralisation oblige) chargé de la promotion du sport (ainsi que de l'encadrement des formations sportives et de la gestion des athlètes). Dans ce but, des marches sont co-organisées tous les dimanches de l'année avec des associations locales. Tous les groupes et associations peuvent ainsi proposer d'organiser, un dimanche de leur choix, des promenades dans leur commune. Moyennant le respect d'un cahier des charges, l'ADEPS sponsorise l'événement et se charge de sa promotion. La marche est gratuite, mais les ventes de boissons et de nourriture permettent aux associations de rentrer dans leurs frais, voir de dégager un bénéfice comme c'est le cas ici.

1. Posons le problème

À la question *combien l'événement nous rapporte-t-il*, il est facile de donner une réponse : l'état du stock est connu et l'argent en caisse en début et fin de journée également, quelques opérations simples suffisent.

Ce que je voudrais savoir, c'est par exemple de savoir quelle est la proportion des gens qui reviennent consommer, disons, sur le temps de midi. Il ne m'est pas possible d'avoir directement accès à cette information car on ne s'amuse pas à compter le nombre de personnes qui rentrent de promenade. Par contre, je me suis amusé à relever, manuellement et toutes les heures :

— Le nombre de marcheurs qui avaient **démarré** une promenade durant la dernière heure (car les marcheurs doivent s'inscrire avant de commencer le parcours, question d'assurance), et
— Le nombre de boissons et de nourritures vendu durant la dernière heure (pareil, on a l'info, cette fois grâce au bon vieux système du ticket boisson™ et du guichet unique).

Autrement dit, **si j'arrive à estimer quand les marcheurs reviennent**, je peux aligner ces chiffres avec ceux du nombre de consommations vendues[1] et avoir une réponse à ma question. À priori, c'est relativement facile et ça ne mérite pas vraiment un billet. On connait la distance des parcours (en plus se sont des multiples de 5 km) et on peut estimer qu'un marcheur marche à du 5 km/h en moyenne (*vide infra*). Autrement dit, on connait le temps que va prendre un marcheur pour faire le parcours, et *yapluka*.

Sauf que si je fais ça, j'ai de grosse différence avec la vente de consommations. Typiquement, j'estime un pic de retours vers 11h (tout les marcheurs du matin sur les différents parcours), tandis que les chiffres indiquent un pic de consommation entre 12 et 13h. Nos marcheurs seraient-ils rentrés manger chez eux ?

Avant de sauter sur les conclusions, il convient d'examiner les prémisses :

1. En procédant comme indiqué plus haut, je pars du principe que tous les marcheurs comptabilisés pour une heure donnée sont partis en même temps. Dans les faits, on est plus proches de marcheurs qui partent régulièrement (ne serait-ce que parce que le nombre de personne au bureau des inscriptions limite le nombre de départ simultanés). Le profil des retours devrait donc être quelque chose d'assez étalé.
2. Par ailleurs, le profil des marcheurs est très varié, et va de la famille nombreuse avec enfants et poussettes aux marcheurs chevronnés qui font ça tous les dimanches : la vitesse moyenne de ces deux groupes n'est probablement pas la même. Ce qui a pour effet d'étaler d'autant plus le profil des retours sur les longs trajets.

Bref, disons carrément des gros mots : le temps de départ d'un marcheur peut être modélisé par une distribution **uniforme**, $T_d(t_0) \sim U(t_0, t_0 + 1)$ (un marcheur peut partir n'importe quand dans l'intervale $[t_0, t_0 + 1[$), tandis que la vitesse des marcheurs peut être modélisée par une **distribution normale** : $V \sim N(\mu, \sigma^2)$ (en km/h). Et donc, la distribution de l'heure de retour est donnée par :

1. En partant par exemple du principe que chaque personne qui revient consomme alors une boisson. Même si on a quand même eux deux ou trois piliers de comptoir. Soit.

$$T_r(t_0, d) = T_d(t_0) + T_p(d) = T_d(t_0) + \frac{d}{V} \sim U(t_0, t_0 + 1) + \frac{d}{N(\mu, \sigma^2)},$$

où d est la distance du parcours (en km).

Ok, bon, T_r est une distribution comme une autre … Non ?

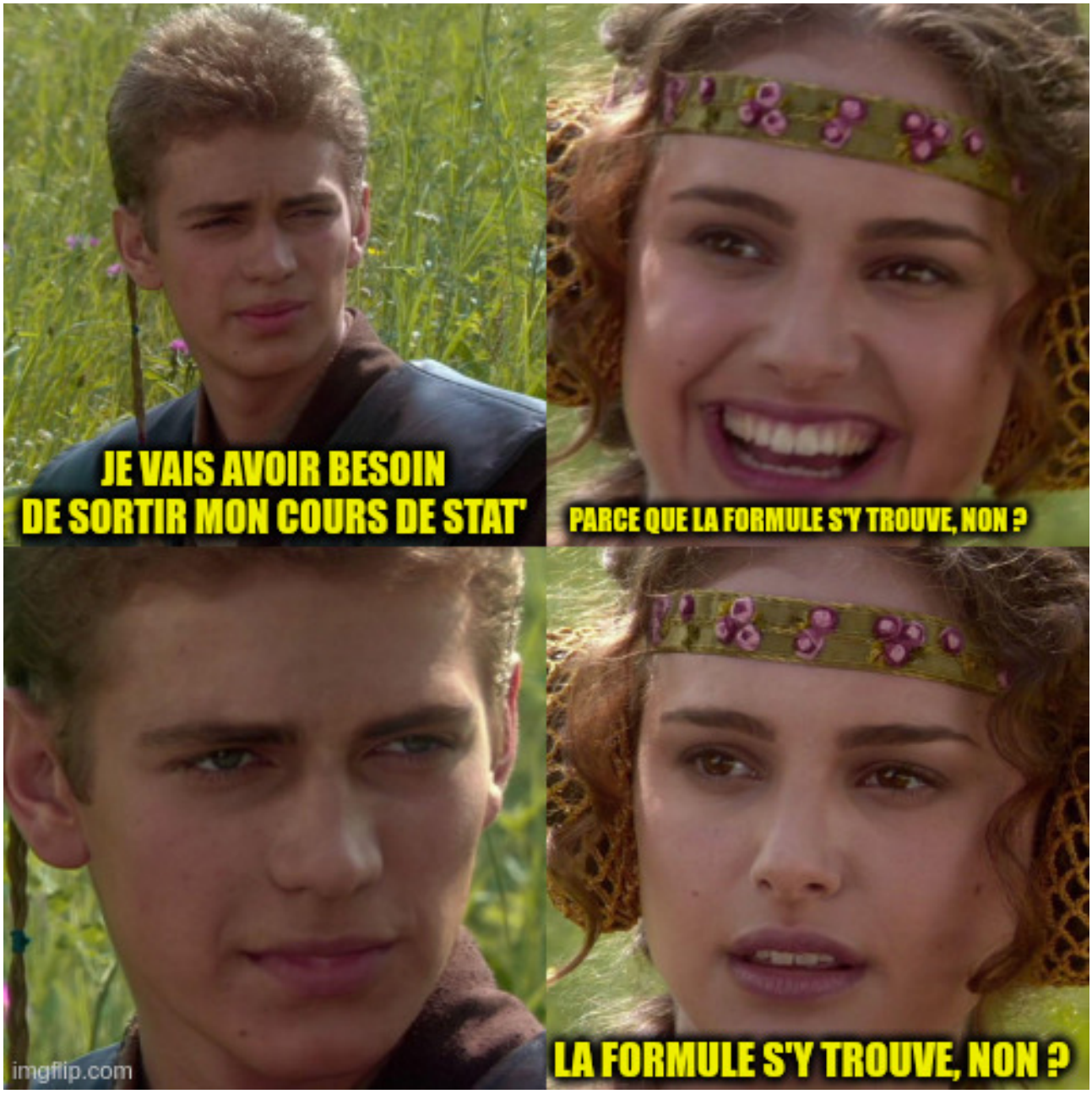

FIGURE 1.1. – Oui, un mème, j'ose. Faut vivre avec son temps.

… pas tout à fait .

2. Faisons des stat's, donc

Si vous avez peur des maths, pas de problème, rendez vous à la section suivante pour les résultats

2.1. Rappels de statistique

Soit une variable aléatoire X. La **fonction de densité de probabilité** (FDP), $f_X(x)$ donne, pour un x donné, la probabilité de sa réalisation (si X est discrète, on parle de fréquence) :

$$f_X(x) = Pr[X = x].$$

La **fonction de répartition** (FR) calcule la probabilité que X prenne une valeur inférieure ou égale à a (si X est discrète, on parle de fréquence cumulée) :

$$F_X(a) = Pr[X \leq a] = \int_{-\infty}^{a} f_X(x)\,dx.$$

On peut alors remarquer que $f_X(x) = \frac{d}{dx} F_X(x)$.

Par ailleurs, la probabilité que X soit compris entre a et b est donnée par :

$$Pr[a \leq X \leq b] = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x)\,dx.$$

Exemple

Si $X \sim \mathcal{U}\{1, 2, 3, 4, 5, 6\}$, c'est à dire si X est une variable aléatoire représentant le résultat d'un dé, alors :

- $Pr[X = 2] = \frac{1}{6}$, c'est à dire qu'il y a une chance sur 6 de faire un 2,
- $Pr[X \leq 2] = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, c'est à dire qu'il y a une chance sur 3 de faire un chiffre inférieur ou égal à 2, et
- $Pr[2 \leq X \leq 5] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$, c'est à dire qu'il y a 2 chances sur 3 de faire un chiffre compris entre 2 et 5.

Autrement dit, si je veux connaitre la probabilité qu'un marcheur parti après 9h sur un parcours de 5 km revienne entre 11 et 12h, je dois calculer $Pr[11 \leq T_r(9, 5) \leq 12]$, en utilisant la notation de la section précédente.

2.2. À la bourrin

Pour estimer la FDP puis la FR de T_r, on peut s'amuser à générer un grand nombre de valeurs en choisissant au hasard une valeur pour T_d et pour V. L'histogramme des valeurs ainsi obtenues (c'est à dire la fréquence de leur apparition) sera alors une estimation de la densité de probabilité. Mieux encore, on peut demander à un ordinateur de le faire pour nous. En utilisant ensuite l'intégration de Riemman , on peut approximer la FR.

Et la vitesse des marcheurs, alors ?

J'ai pris une estimation issue de ce papier , qui estime la vitesse moyenne d'un piéton à 1.36 m/s (σ^2=0.19 m/s) dans une zone avec large trottoir, ce rapprochant d'une situation avec faible densité de piéton, ou les marcheurs vont à leur rythme. Notez ceci dit que la valeur pourrait être encore plus élevée, comme indiqué dans d'autres articles (voir figure 8 de celui-ci ou un peu tout le papier mais en particulier la figure 5 de celui-là).
On pourrait s'amuser à discuter le fait que sur les parcours plus courts, le profil des marcheurs est plus familial, là ou les parcours plus longs sont prisés par des personnes plus aguerries, et probablement plus rapides.

Par exemple, pour un parcours de 5 km et un départ après 9h (entre 9h et 10h, donc), en tirant 1 million de valeurs au hasard à l'aide de ce code …

```python
import numpy
import matplotlib.pyplot as plt

distance = 5 # km
t0 = 9 # h
# vitesse moyenne et déviation standard issue de https://dx.doi.org/10.1016/j.sbspro.2013.11.160
vitesse_moyenne = 4.896 # km/h
vitesse_std = 0.684 # km/h

# tire au hasard des nombres suivant la loi de probablité
# voir https://numpy.org/doc/stable/reference/random/index.html#module-numpy.random
gna = numpy.random.default_rng()
N = 1_000_000
Tr = gna.uniform(t0, t0 + 1, N) + distance / gna.normal(vitesse_moyenne, vitesse_std, N)

# calcule la fonction de densité de probabilité (FDP, ça ne s'invente pas)
FDP, bin_edges = numpy.histogram(Tr, bins=200, range=(t0,t0 + 3), density=True)

# plot
figure = plt.figure(figsize=(10, 6))
ax1, ax2 = figure.subplots(1, 2)
```

```
22
23 ax1.scatter(bin_edges[:-1], FDP)
24 ax1.set_ylabel('Densité de probabilité')
25 ax1.set_xlabel('Heure de retour')
26
27 # on "integre" la FDP pour obtenir la FR
28 ax2.scatter(bin_edges[:-1], numpy.cumsum(FDP) *
       numpy.diff(bin_edges))
29 ax2.set_ylabel('Fonction de répartition')
30 ax2.set_xlabel('Heure de retour')
31
32 plt.tight_layout()
33 plt.show()
```

… On obtient quelque chose qui ressemble à :

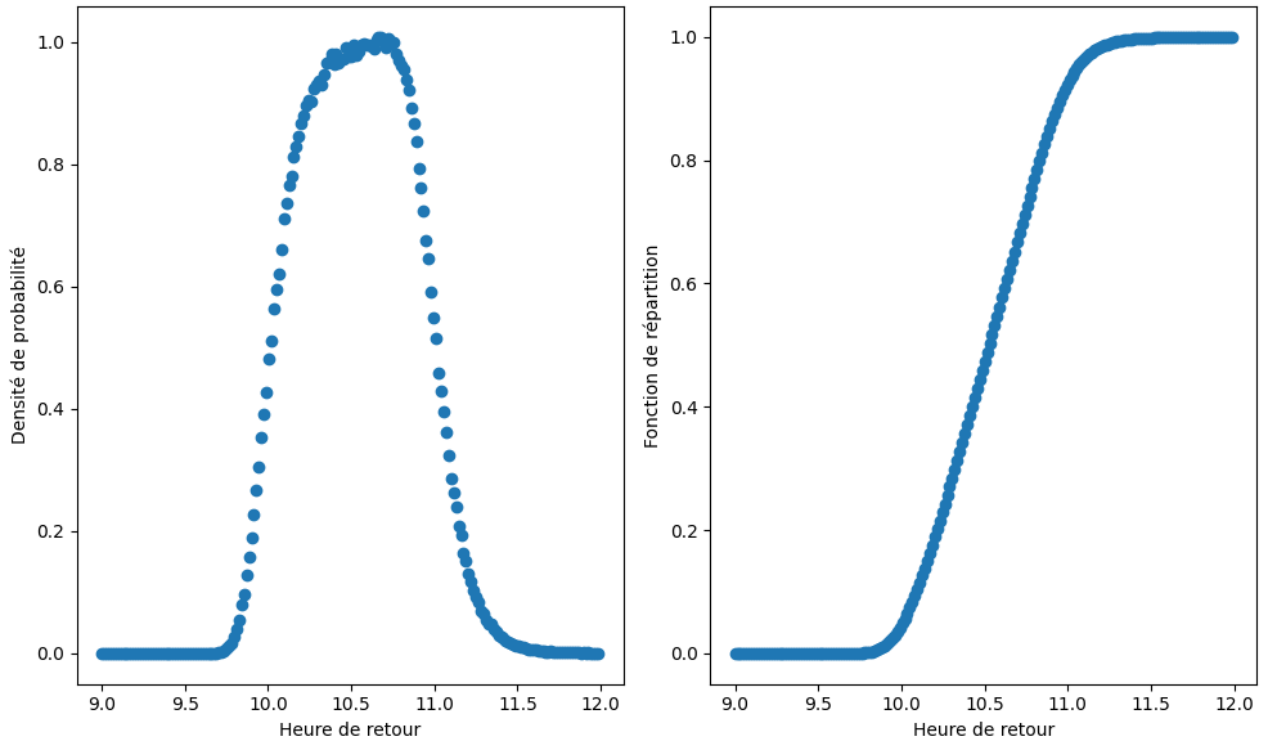

FIGURE 2.2. – Densité de probabilité (gauche) et fonction de répartition (droite) pour $T_r(9, 5)$, à partir de valeurs aléatoires.

Tandis que le graphe de gauche représente la répartition des probabilités individuelles (et a un profil un peu différent de ce qu'on a l'habitude de voir, puisqu'il s'agit d'une somme de lois normales inverses), c'est le graphe de droite qui permet de répondre à des questions intéressantes. Tout d'abord, puisqu'il s'agit d'une probabilité, les valeurs de celui-ci varient entre 0 et 1. Par ailleurs, il permet par exemple d'estimer que :

— La probabilité de rentrer avant 10h30 est d'environ 48%, puisque $Pr[T_r(9, 5) \leq 10.5] \approx 0.48$. Il s'agit simplement de la valeur correspondant à 10.5 sur le graphe de droite.

— La probabilité de rentrer entre 11h et 12h est d'environ 8%, puisque $Pr[11 \leq T_r(9,5) \leq 12] = Pr[T_r(9,5) \leq 12] - Pr[T_r(10,5) \leq 11] \approx 0.08$, avec $Pr[T_r(9,5) \leq 12] \approx 0.92$ et $Pr[T_r(9,5) \leq 12] \approx 1$ (valeurs lues, une fois encore, sur le graphe de droite).
— S'il y a des marcheurs qui sont partis peu après 9h et qui marchent vite, ils peuvent arriver avant 10h : $Pr[T_r(9,5) \leq 10] \approx 0.03$.

2.3. Plus de subtilité, donc plus de maths

Il est tout de fois possible d'obtenir des résultats plus précis en arrêtant de jouer aux dés et en sortant du papier et un crayon. Tout d'abord, on peut remarquer que pour ce qui est de la vitesse, on calcule en fait une distribution normale inverse , donc il est possible de connaitre la FDP.

Inverse d'une variable aléatoire

Soit une variable aléatoire X et soit une seconde variable aléatoire définie comme $Y = \frac{1}{X}$, on a que la FR de Y est donné par :

$$F_Y(y) = Pr[Y \leq y] = Pr\left[\frac{1}{X} \leq y\right] = Pr\left[X \geq \frac{1}{y}\right] = 1 - Pr\left[X < \frac{1}{y}\right] = 1 - F_X\left(\frac{1}{y}\right).$$

Et dès lors, en utilisant la dérivation de fonction composées ,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[1 - F_X\left(\frac{1}{y}\right)\right] = \frac{1}{y^2} f_X\left(\frac{1}{y}\right).$$

Autrement dit, la FDP pour une loi inverse normale, que je vais noter $InvN(\mu, \sigma^2)$, est :

$$f_{InvN}(y; \mu, \sigma^2) = \frac{1}{y^2} f_N(y; \mu, \sigma^2) = \frac{1}{y^2\, \sigma\, \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\frac{1}{y} - \mu}{\sigma}\right)^2\right].$$

On peut ensuite gérer le produit d'une variable aléatoire par une constante :

Multiplication d'une variable aléatoire par une constante

Soit une variable aléatoire X et soit une seconde variable aléatoire définie comme $Y = c\,X$ avec $c > 0$, on a que la FR de Y est donné par :

$$F_Y(y) = Pr[Y \leq y] = Pr[c\,X \leq y] = Pr\left[X \leq \frac{y}{c}\right] = F_X\left(\frac{y}{c}\right).$$

Et donc, en dérivant,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y}{c}\right) = \frac{1}{c} f_X\left(\frac{y}{c}\right).$$

Et donc, on peut trouver la FDP pour le temps de parcours [noté $T_p(d)$ plus haut] comme :

2. Faisons des stat's, donc

$$T_p(d) = \frac{d}{N(\mu, \sigma^2)} = d \times InvN(\mu, \sigma^2) \Rightarrow f_{T_p}(y; d) = \frac{1}{d} f_{InvN}\left(\frac{y}{d}; \mu, \sigma^2\right).$$

Et finalement, si on veut une expression pour la FDP de T_r, il va falloir additioner $T_p(d)$ et une distribution uniforme sur $[t_0, t_0 + 1[$.

Somme de deux variables aléatoires indépendantes

Soit deux variables aléatoires X et Y, indépendantes, et soit une troisième variable aléatoire définie comme $Z = X + Y$, la FDP est donnée part le produit de convolution des FDPs de X et Y :

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(z - y)\, f_Y(y)\, dy.$$

La démo est un peu plus compliquée, mais suis le principe de "travailler sur la FR puis dériver pour obtenir la FDP" qu'on a vu avant.

Autrement dit, pour évaluer la FDP de $T_r = T_d + T_p$, on doit évaluer

$$f_{T_r}(t; t_0, d) = \int_{t_0}^{t_0+1} f_{T_p}(t - y; t_0)\, f_{T_d}(y; d)\, dy,$$

où j'ai utilisé le fait que la FDP de T_d est de zéro en dehors de $[t_0, t_0 + 1[$. Malheureusement, évaluer cette intégrale est un petit peu au delà de mes compétences, d'autant vu la forme de T_p qui n'est pas une "simple" Gaussienne. Notez que c'est néanmoins possible pour la somme de distributions plus simples. Je vais donc opter pour une convolution numérique issue des méthodes de nos amis de la manipulation du signal.

En utilisant ce code implémentant la formule ci-dessus pour T_r, on peut obtenir le résultat suivant :

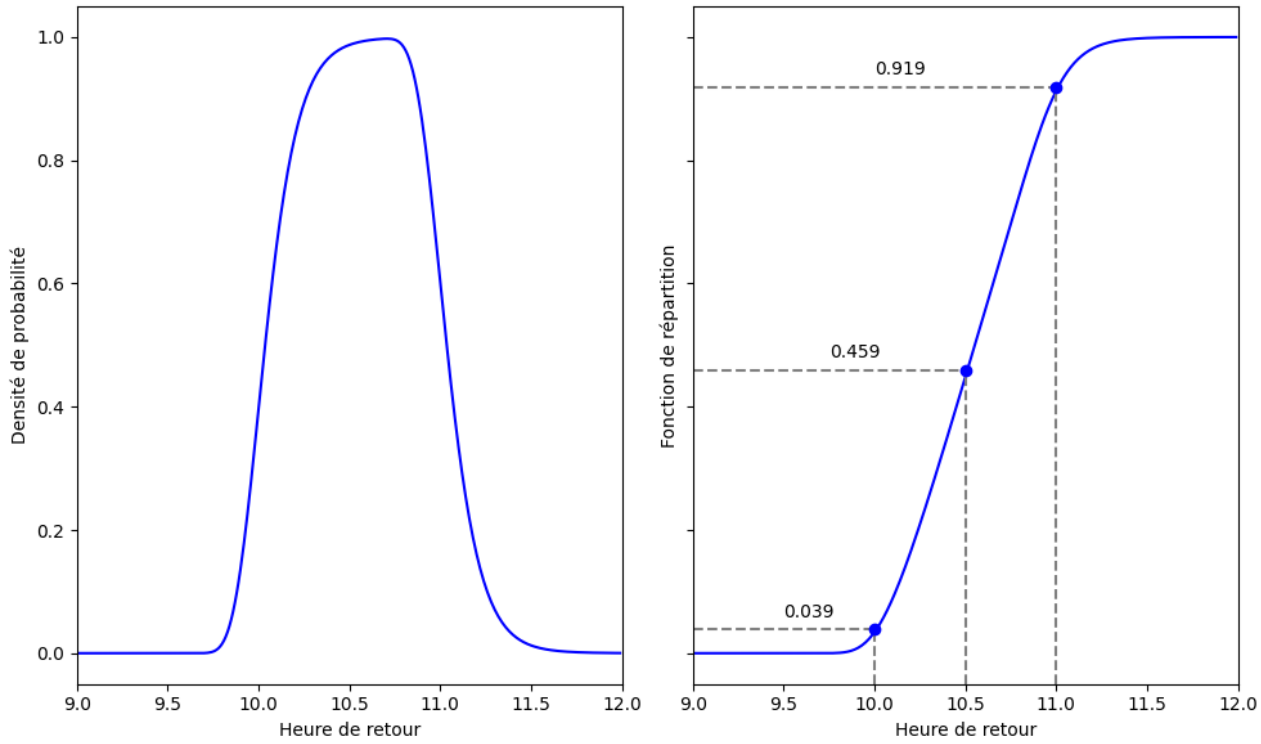

FIGURE 2.3. – Densité de probabilité (gauche) et fonction de répartition (droite) pour $T_r(9,5)$, obtenu à partir de l'approche analytico-numérique dévellopée ci-dessus. Sur le graphe de droite, on voit les probabilités "correcte" pour l'arrivée à différentes heures. Par exemple, celle de rentrer avant 10h30 est plutôt de l'ordre de 46%.

On constate que cette approche donnne des résultats équivalents à la précédente, quoi que plus rapidement et de manière un peu plus précise. Et que la forme de la FDP n'est définitivement pas commune 🍊

? Ça aurait pas une gueule de Poisson, ton truc ?

Ça y ressemble et on pourrait imaginer modéliser le nombre de marcheur revenant de marche avec, mais ce n'est pas exactement les mêmes prémisses.

3. Et donc ?

Grâce aux quelques développements de la section précédente, on sait évaluer la probabilité qu'un marcheur rentre dans un certain intervalle de temps. Voyons ce que ça donne avec les 4 parcours de la marche :

Période	**5.6 km**	**10.7 km**	**16.0 km**	**20.0 km**
$[t_0, t_0+1[$	0.9	0	0	0
$[t_0+1, t_0+2[$	**81.5**	3.5	0	0

$[t_0+2, t_0+3[$	17.6	**70.7**	5.4	0
$[t_0+3, t_0+4[$	0	25.2	**59.2**	15.4
$[t_0+4, t_0+5[$	0	0.6	32.1	**56.5**
$[t_0+5, t_0+6[$	0	0	3.1	24.1
$[t_0+6, t_0+7[$	0	0	0.2	3.5
$[t_0+7, t_0+8[$	0	0	0	0.4
$[t_0+8, t_0+9[$	0	0	0	0.1

TABLE 3.2. – Probabilité (en %), pour un marcheur parti à $t_d \in [t_0, t_0+1[$, de finir le parcours dans un intervalle de temps donné. Obtenu à partir d'ici .

Les chiffres correspondent à peu près à l'estimation qu'on aurait pu faire en utilisant simplement la distance et la vitesse moyenne (à savoir qu'un parcours de 16 km prend à peu près 3h), mais la distribution obtenue tient donc compte des départs différés et du profil de vitesse des marcheurs.[1] Dit autrement, si on compte que 20 marcheurs se lancent sur le parcours de 16 km entre 9h et 10h, on aura statistiquement, 1 marcheur (motivé !) qui reviendra avant midi, 12 qui reviendront entre midi et 13h, et les 7 derniers reviendront après 13h.

Bref, si on additionne tout ça en utilisant le nombre de départs , on obtient ceci :

1. Encore une fois, on pourrait discuter du fait que les marcheurs sur des parcours plus long vont probablement un peu plus vite … Quoiqu'ils fatiguent probablement …

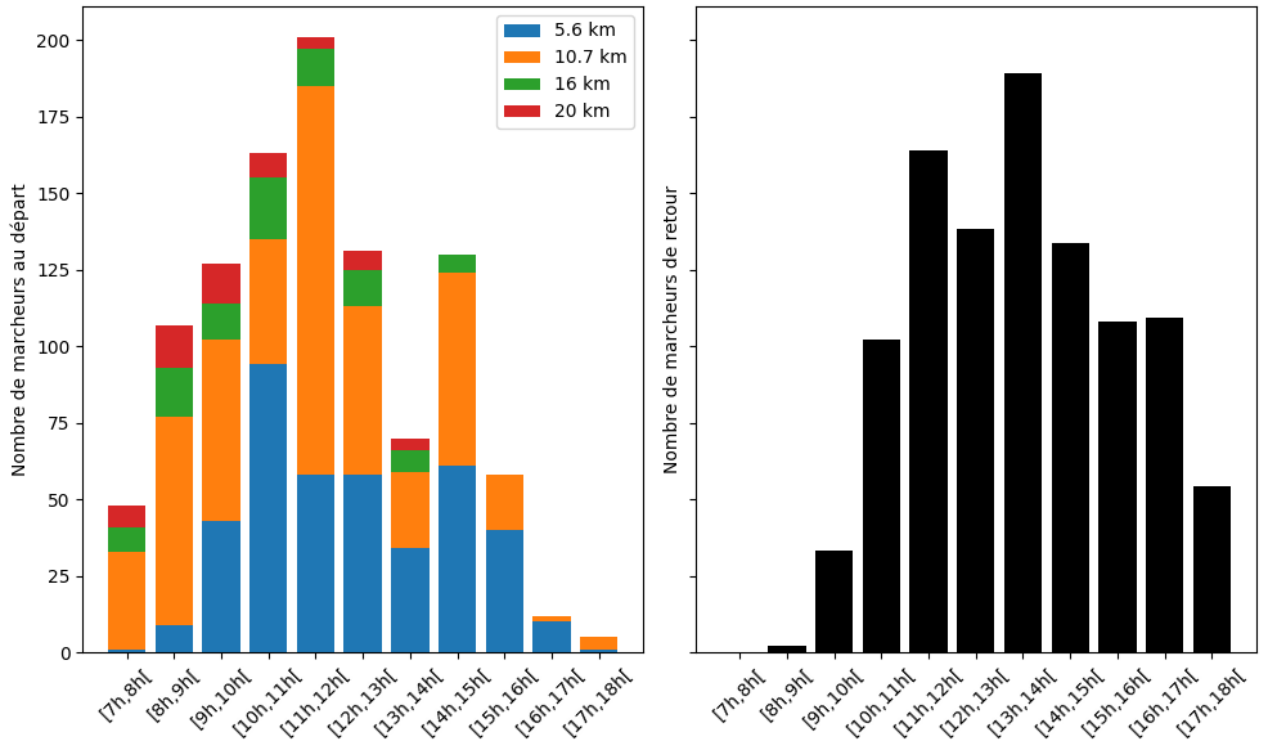

FIGURE 3.4. – Nombre de départs (gauche) et de retours (droite) pour chaque intervalle de temps, estimé à partir de la probabilité de retours et du nombre de départs.

On peut observer différentes choses sur ce graphe. En vrac, que 80% de nos marcheurs viennent en matinée (avant 12h) pour faire des longs parcours (et sont plutôt motivés, avant 9h !), mais que les retours sont plutôt étalés sur le midi et l'après-midi (11h-17h). Il y a également un petit rebond des départs en début d'après-midi (après 14h, généralement pour des parcours plus court) qui vient s'ajouter à la vague des retours de milieu/fin de l'après-midi. On observe également qu'il y a des gens qui la joue stratégique en partant un peu avant midi pour éviter le *rush*. Manque de pot, 2h plus tard, les pains saucisses avaient bien diminué[2] 🍊
Plus qu'à comparer avec les chiffres de vente (mais, au doigt mouillé, je dirais qu'environ 25% des gens consomment ensuite).

Conclusion

Eeeeeeeeeeeeeeeeeeet … Tout ça pour ça, en fait 🍊
Prenez-le comme moi qui m'amuse un peu avec les chiffres (j'aime bien les statistiques) sans trop de prétentions non plus (d'ailleurs, si j'ai fait des bêtises, n'hésitez pas à me le signaler).
Et sinon, faites du sport, c'est bon pour la santé.
L'icone du billet est issue d'une image de Silvia Natalia sur the Noun Project

2. C'est compliqué, la gestion du stock 🍊🤷

Liste des abréviations

FDP fonction de densité de probabilité. 4, 5, 7– 9

FR fonction de répartition. 4, 5, 7, 8