INSA ROUEN
À taille humaine,
à l'échelle du monde

Projet de statistiques - M8

La vérité sur la nuit du 15 avril 1912

A l'attention de
Stéphane Canu

Étudiants :
Anthonin Lizé
Simon Rohou

2010/2011

Table des matières

Introduction

Nous avons choisi ce projet puisque nous avons été tous les deux passionnés à un moment ou à un autre par l'histoire dramatique qu'a été le naufrage de l'Insubmersible, alias le *Titanic*.

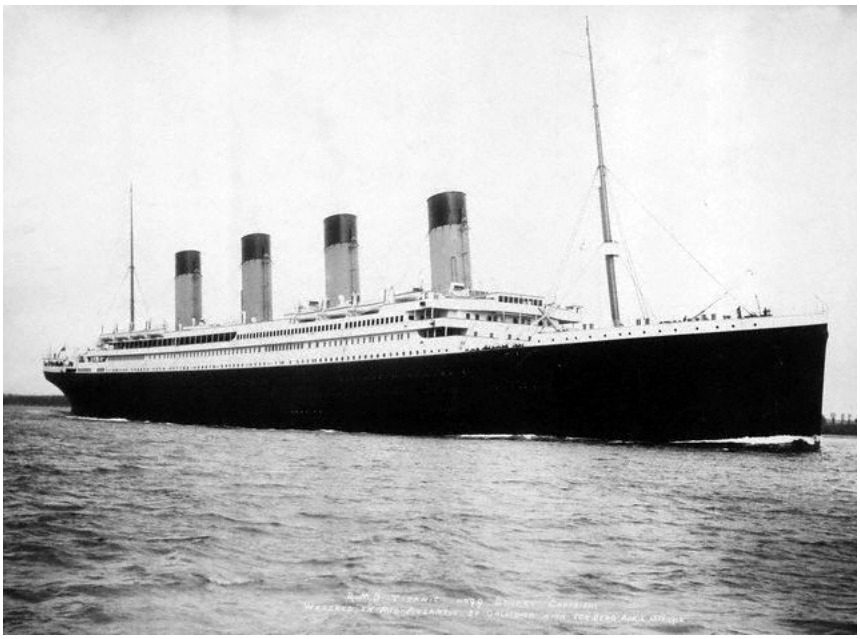
Nous pensions que grâce à ce projet nous pourrions confirmer le privilège des classes sociales à l'époque de ce drame et ainsi confirmer ce que tout le monde pense, ce que tout le monde a vu dans les films ou livres traitant de ce sujet par les statistiques. Bien sûr, ce projet n'a pas grand intérêt dans la vie de tous les jours, mais il est un moyen d'appliquer notre cours sur un sujet qui nous motive.

Du fait que cette histoire ait attiré plein d'historiens et de passionnés à travers le monde, nous avons pu récolter tout un tas de données sur l'ensemble des passagers : le nom, l'âge, la classe, le numéro de billet d'embarcation, le prix de billet, le type de ce billet, le groupe de passagers (simple passagers, servants, H&W Guarantee Group, musiciens, Cross Channel Passenger, ouvriers postaux), le port d'embarcation, l'emploi, le canot de sauvetage pris (si embarqué), le numéro du corps (si mort et retrouvé), le sexe et bien sûr la survie. Nous nous servirons uniquement de l'âge, la classe, le prix du billet et du sexe pour notre étude qui consistera à répondre aux questions suivantes :

– L'adage « *les femmes et les enfants d'abord* » a-t'il été respecté lors du naufrage ?

– Un passager aurait-il pu prédire sa probabilité de survie en fonction du prix de son billet ?

Pour mener à bien cette étude, il sera donc judicieux de ne regarder parfois que les observations concernant les passagers.



Le *Titanic* le 10 avril 1912 à Southampton

1. Description des variables

Voici un aperçu de nos données :

Name	Age	Class	Fare	Sex	Alive
ABBING, Mr Anthony	42	3	7	1	0
ABBOTT, Mrs Rhoda Mary 'Rosa'	39	3	20	0	1
ABBOTT, Mr Rossmore Edward	16	3	20	1	0
ABBOTT, Mr Eugene Joseph	14	3	20	1	0
ABBOTT, Mr Ernest Owen	21	4	0	1	0
ABELSETH, Miss Karen Marie	16	3	7	0	1
ABELSETH, Mr Olaus Jørgensen	25	3	7	1	1
ABELSON, Mr Samuel	30	2	24	1	0
ABELSON, Mrs	28	2	24	0	1
ABRAHAMSSON, Mr Abraham August Johannes	20	3	7	1	1
ABRAHIM, Mrs Mary Sophie Halaut	18	3	7	0	1
ABRAMS, Mr William	33	5	0	1	0
ÅDAHL, Mr Mauritz Nils Martin	30	3	7	1	0
ADAMS, Mr John	26	3	8	1	0
ADAMS, Mr R.	26	5	0	1	0
AHIER, Mr Percy Snowden	20	4	0	1	0
AHLIN, Mrs Johanna Persdotter	40	3	9	0	0
AKERMAN, Mr Albert	28	4	0	1	0
AKERMAN, Mr Joseph Francis	35	4	0	1	0
AKS, Mrs Leah	18	3	9	0	1
AKS, Master Frank Philip	0,833	3	9	0	1

FIG. 1.1 – Extrait des données sur les passagers

L'âge est en années, le prix en Livre Sterling (£), le sexe indique *homme* si il égale 1, *femme* si 0 et la survie est égale à 1 si la personne a survécu, 0 sinon.

À noter que, par manque d'informations, nous avons dû enlever 32 observations incomplètes, ce qui réduit l'étude à un échantillon de **2208 observations**.

Après quelques calculs sous *MatLab*, on obtient :

	Age	Class	Fare	Sex
Moyenne	29.7731		19.4819	
Médiane	29	3	7	
Ecart type	11.9113		43.1756	
Variance	141.8779		1864.1	

FIG. 1.2 – Statistiques sur les variables étudiées

Les champs grisés montrent qu'il n'y a aucun sens à calculer ces valeurs, étant donné que ce sont des variables qualitatives.

1 Focus sur la variable « alive »

La variable « *alive* » est dans notre étude la variable à expliquer. 0 signifie que le passager n'a pas survécu au drame, 1 signifie le contraire. Voici les effectifs totaux :

Morts	**Survivants**
1495	713

FIG. 1.3 – Effectifs totaux de la variable *alive*

Si on restreint l'étude aux passagers, on obtient les effectifs suivants :

Morts	**Survivants**
816	501

FIG. 1.4 – Survie parmi les passagers

2 Focus sur la variable « âge »

Pourquoi une étude autour de l'âge ? Tout simplement parce que la règle d'or présente sur un bateau lors d'un naufrage est : « *Les femmes et les enfants d'abord !* ». Cet adage a-t-il été respecté ou cet ordre était-il un prétexte pour rassasier les requins ?

Lisez jusqu'au bout et peut être que la réponse viendra. Tout d'abord un premier graphique, qui représente l'âge en fonction de la survie et une fonction de répartition empirique :

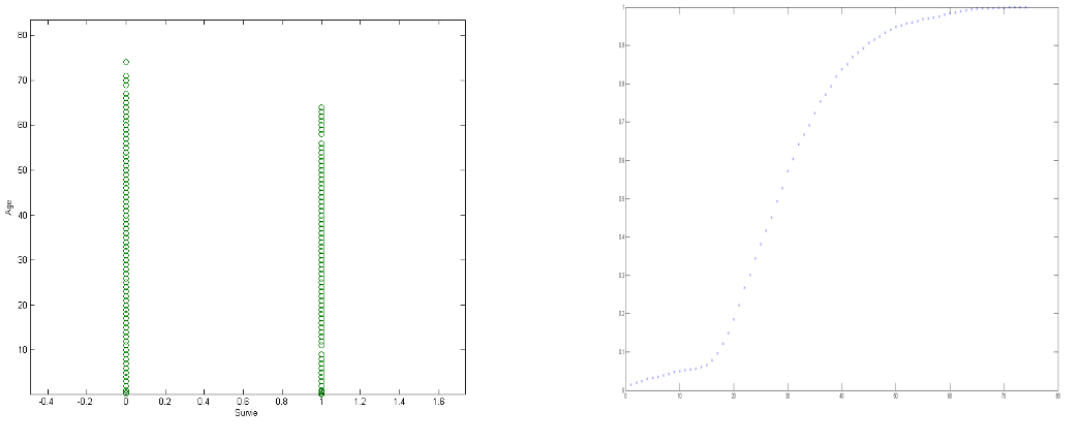


FIG. 1.5 – Age et survie

Pas très parlant me direz-vous. Un histogramme sera mieux adapté à la situation :

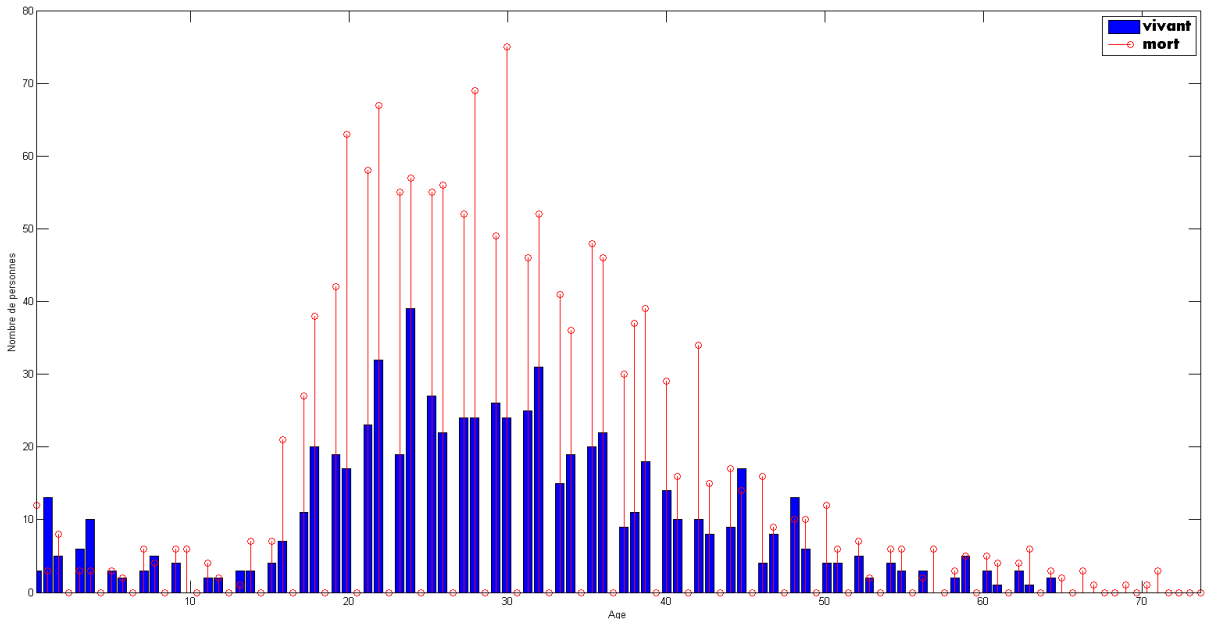


FIG. 1.6 – Histogramme des variables *âge* et *survie*

On peut voir que les nouveaux nés n'ont malheureusement pas résisté au naufrage, mais que les enfants en général s'en sont mieux sortis que les ados/adultes[1]. De plus, on remarque qu'il y a très peu de survivants chez les personnes âgées. Ceci est certainement dû à leurs conditions physiques lors du drame. Donc à première vue, la partie âge de l'adage est vérifiée.

On obtient également les boites à moustaches suivantes :

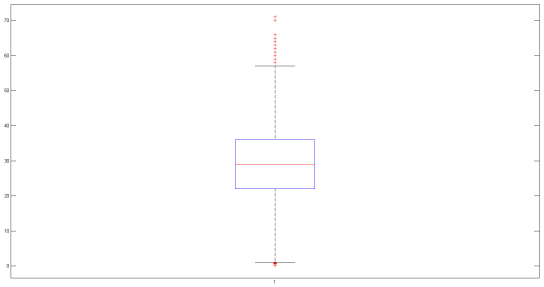


FIG. 1.7 – Boites à moustaches représentant ceux qui n'ont pas survécu

[1]Notez qu'à l'époque on n'était plus considéré comme enfant à partir d'une quinzaine d'années.

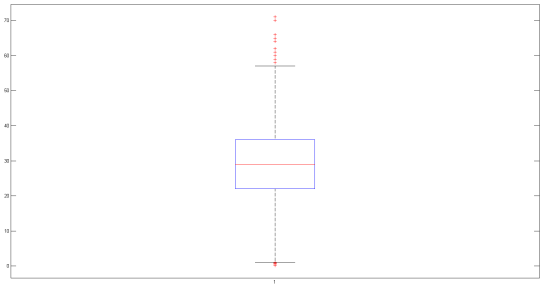


FIG. 1.8 – Boites à moustaches représentant ceux qui ont survécu

3 Focus sur la variable « sex »

Pour continuer sur l'étude de cet adage, nous allons maintenant nous concentrer sur la description de la variable « *sex* ».

Il n'y a ici aucune utilité à faire un graphique ou nuage de points à moins de trouver un intérêt à la représentation d'un carré. Nous avons donc directement fait un histogramme des valeurs, ce qui donne :

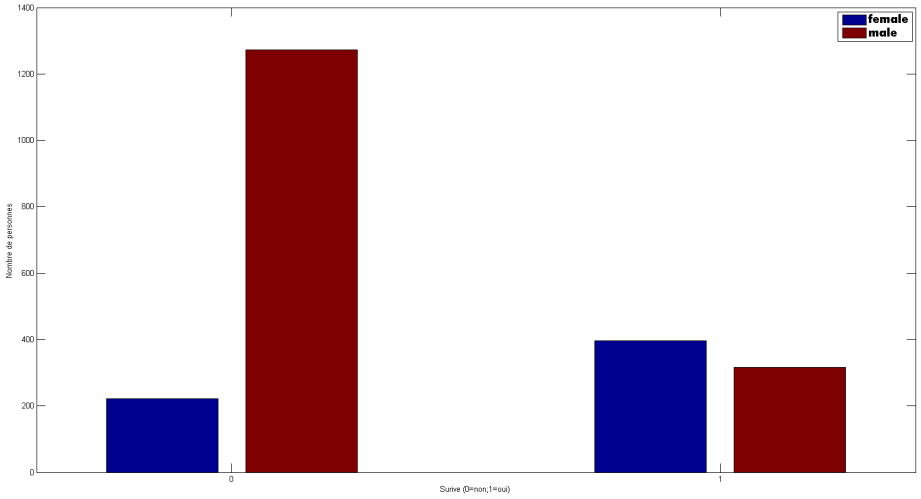


FIG. 1.9 – sexe = f(survie)

Il y a beaucoup plus de morts chez les hommes que chez les femmes. Cela dit, les hommes étaient très nombreux et par conséquent, le nombre de survivantes chez les femme est à peu près équivalent au nombre de survivants chez les hommes.

On ne peut donc pas conclure grand-chose, il faudrait étudier les rapports, qui seraient plus significatifs.

	Femmes	**Hommes**
Morts	222	1273
Vivants	397	316

FIG. 1.10 – Récapitulatif de la survie en fonction du sexe

On retient :

- $\dfrac{\text{Morts}}{\text{Vivants}} = \dfrac{222}{397} = 0,559$ pour les femmes ;
- $\dfrac{\text{Morts}}{\text{Vivants}} = \dfrac{1273}{316} = 4,02$ pour les hommes.

Le rapport pour les hommes est supérieur à celui des femmes. La partie de l'adage par rapport au sexe semble également avoir été respectée.

4 Focus sur la variable « class »

Les passagers sur le *Titanic* étaient classés par classe, en fonction de leur rang social et du type de ticket qu'ils achetaient (par conséquent, l'étude de cette variable sera très liée à celle de la variable « *fare* »).

Grâce à MatLab on obtient l'histogramme suivant :

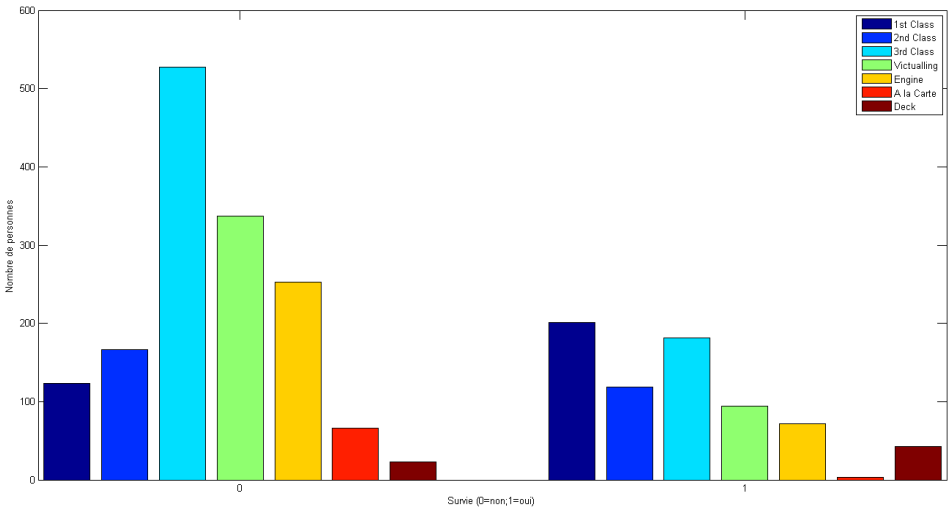


FIG. 1.11 – Répartition des survies dans les différentes classes

Ce qui correspond aux effectifs suivants :

	Première classe	**Deuxième classe**	**Troisième classe**
Morts	123	166	527
Vivants	201	119	181
Ratio Vivants/Morts	1,63	0,71	0,34

FIG. 1.12 – Récapitulatif de la survie en fonction de la classe

À première vue, les passagers de première classe ont eu plus de chance que ceux de la troisième.

5 Focus sur la variable « fare »

Ce qui différenciait les passagers à bord du paquebot était avant tout l'argent qu'ils possédaient (et par conséquent, le prix d'achat de leur billet). Si on représente le nombre de morts et de vivants en fonction du prix de leur billet, on obtient l'histogramme suivant :

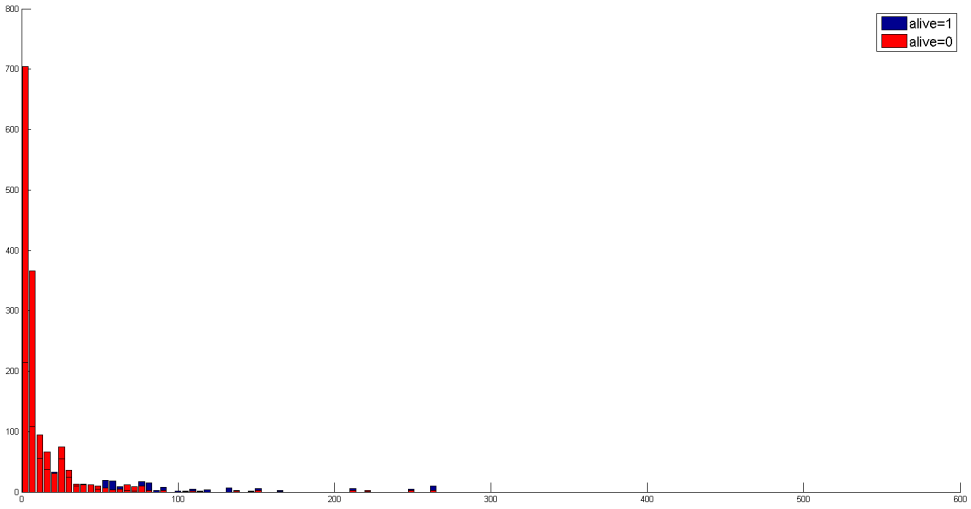


FIG. 1.13 – Survie et prix de billets

On observe que le nombre de survivants (en bleu) dépasse le nombre de morts (en rouge) à partir d'un peu moins de 100 £.

2. Régression linéaire

1 Régression multiple avec toutes les variables

Ici on cherche à savoir si la survie peut s'exprimer en fonction des variables considérées, à savoir l'âge, le prix du billet, la classe et le sexe.
Pour cela, nous avons suivi le modèle suivant :

$$alive = \alpha_0 + age \cdot \alpha_1 + class \cdot \alpha_2 + fare \cdot \alpha_3 + sex \cdot \alpha_4 \tag{2.1}$$

Sous forme matricielle (pour simplifier l'écriture), cela nous donne :

$$y = X \cdot \alpha \tag{2.2}$$

Avec :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = alive = \begin{pmatrix} alive_1 \\ \vdots \\ alive_n \end{pmatrix} \;\; ; \;\; \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_4 \end{pmatrix} \;\; ; \;\; X = \begin{pmatrix} 1 & age_1 & class_1 & fare_1 & sex_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & age_n & class_n & fare_n & sex_n \end{pmatrix} \tag{2.3}$$

$$\tag{2.4}$$

L'étude étant réduite aux observations ne concernant que les passagers, nous avons $n = 1317$.
Nous avons ensuite écrit le script suivant :

```
[n p] = size(X)
ahat = (X'*X)\(X'*y)
z = X*ahat;
erreur = y - X * ahat;
MatR2 = corrcoef(y,z);
R2 = MatR2(1,2) * MatR2(1,2)
```

Ce qui nous donne :

$$\alpha = \begin{pmatrix} 1.0919 \\ -0.0023 \\ -0.1601 \\ 0.0001 \\ -0.4803 \end{pmatrix} \;\; ; \;\; R^2 = 0.3498 \tag{2.5}$$

Ce qui est de loin insuffisant car le coefficient de corrélation R^2 est éloigné de 1. Nous allons donc faire des régressions linéaires en ne prenant en compte que certaines variables.

2 Régressions multiples spécifiques

2.1 Tableau des R^2

Voici les R^2 obtenus pour chaque variable, prise individuellement :

Variable	Modèle	R^2
age	$alive = \alpha_0 + age \cdot \alpha_1$	$9.1239.10^{-0.04}$
class	$alive = \alpha_0 + class \cdot \alpha_1$	0.0966
fare	$alive = \alpha_0 + fare \cdot \alpha_1$	0.0618
sex	$alive = \alpha_0 + sex \cdot \alpha_1$	0.2831

FIG. 2.1 – R^2 des variables étudiées

Nous remarquons que la variable « *sex* » a le plus gros R^2.

2.2 Régressions

Variable	Modèle	R^2	α
sex	$alive = \alpha_0 + sex \cdot \alpha_1$	0.2831	$\begin{pmatrix} 0.6855 \\ -0.5238 \end{pmatrix}$
sex + class	$alive = \alpha_0 + sex \cdot \alpha_1 + class \cdot \alpha_2$	0.3463	$\begin{pmatrix} 1.0055 \\ -0.4952 \\ -0.1469 \end{pmatrix}$
sex + class + fare	$alive = \alpha_0 + sex \cdot \alpha_1 + class \cdot \alpha_2 + fare \cdot \alpha_3$	0.3463	$\begin{pmatrix} 0.9930 \\ -0.4936 \\ -0.1434 \\ -0.0001 \end{pmatrix}$

FIG. 2.2 – R^2 des variables étudiées

Notre R^2 n'augmente pas après les variables « *sex* » et « *class* » (étant donné que « *fare* » est très lié à « *class* », il est logique que cela ne bouge pas).

Notre R^2 étant toujours faible, on en déduit qu'il y a des points aberrants à retirer de l'étude.

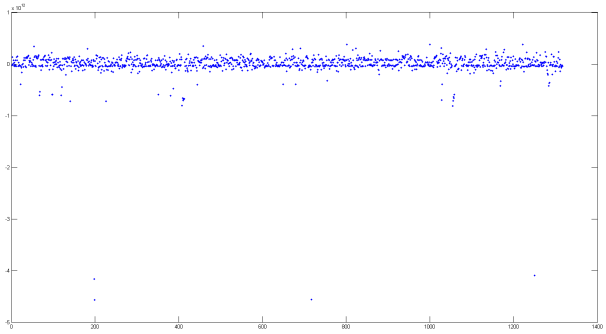
3 Régression sans points aberrants

3.1 Détection des points aberrants

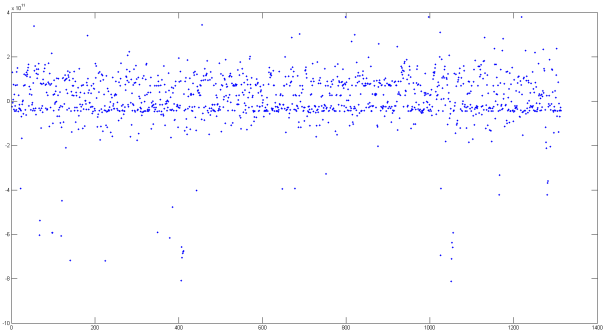
On utilise ce script pour enlever les points aberrants :

```
XC = [age class fare sex erreur contributions];
j = 1;
for i = 1:length(XC(:,1))
  if(vabs(XC(i,6)) < 4e12)
    Xn(j,:) = XC(i,:);
    yn(j,1) = DP(i,5); // represente la survie
    j=j+1;
  end
end
```

$4e^{12}$ étant la marge établie selon l'étude de ce nuage de points, qui est une représentation des contributions :



On obtient les contributions suivantes :



Et le R^2 vaut 0.1920, ce qui n'est toujours pas suffisant. En réduisant la marge à $1e^{11}$, on ne monte qu'à 0.2360.

4 Régression avec A.C.P. (Analyse en Composantes Principales)

On se propose d'effectuer une Acp afin de mieux représenter nos données.

4.1 L'As épais

Pour mener à bien notre Acp, il nous faut premièrement centrer et réduire nos données. Soit X notre matrice des données, μ son espérance et σ son écart type. Notre variable centrée réduite X_n s'écrit comme suit :

$$X_n = \frac{X - \mu}{\sigma} \tag{2.6}$$

Travaillant dans le domaine des statistiques, μ est ici la moyenne. Nous définissons maintenant la matrice S telle que :

$$S = \frac{{X_n}^T X_n}{n} \tag{2.7}$$

S est par écriture définie positive ET symétrique.
Nous obtenons les valeurs propres et vecteurs propres suivants :

$$D_n = diag(D) = \begin{pmatrix} 0.4627 \\ 0.5354 \\ 0.8056 \\ 1.1876 \\ 2.0065 \end{pmatrix} \tag{2.8}$$

$$V = \begin{pmatrix} -0.0567 & 0.2934 & 0.5878 & -0.7498 & 0.0541 \\ -0.6978 & -0.0790 & 0.3859 & 0.2864 & -0.5252 \\ -0.6336 & -0.3870 & -0.2859 & -0.2893 & 0.5322 \\ 0.2461 & -0.7252 & -0.0806 & -0.4013 & -0.4959 \\ 0.2187 & -0.4816 & 0.6460 & 0.3331 & 0.4382 \end{pmatrix} \tag{2.9}$$

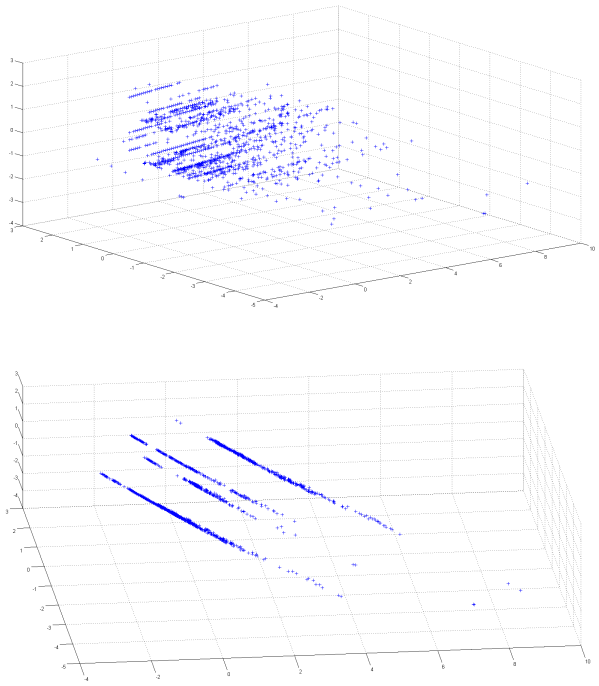
(2.10)

Voici les indices de qualité concernant les valeurs propres :

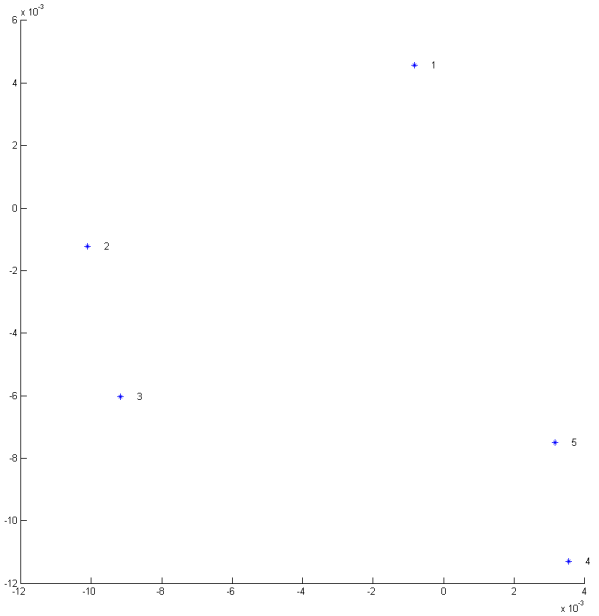
Valeurs propres	**Indice**
2.0065	0.4015
1.1876	0.6391
0.8056	0.8003
0.5354	0.9074
0.4627	1

Fig. 2.3 – Indices de qualité

On observe qu'avec les trois premières composantes principales (CP) on représente 80% des observations. Ce qui donne sous MatLab :



On distingue bien 4 groupes de données différents. La représentation des variables nous donne :



L'échelle étant en 10^{-3}, cela indique que nos variables sont très corrélées car elles sont très proches. Nous allons essayer de faire une régression linéaire sur ces variables pour voir si notre R^2 change ou non.

4.2 Et ça régresse !

Soit U la composante principale des données définie par :

$$U = X_n \cdot V \tag{2.11}$$

Où X_n est notre matrice centrée réduite et V notre matrice de vecteurs propres. Si on pose V_n comme étant les vecteurs propres centrée et réduits, on aura une meilleure approximation de nos données en faisant le calcul :

$$M = C \cdot {V_n}^T \tag{2.12}$$

On obtient :

$$R^2 = 1 \quad \text{et} \quad \alpha = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} \tag{2.13}$$

Nous avons tout d'abord crié à l'erreur en voyant cela, mais nous avons cherché et nous n'en avons pas trouvées. Nous avons aussi fait des prédictions sur la matrice, et nous tombons juste. Le 1 doit être une approximation d'un R^2 tel que :

$$0.99 \ll R^2 < 1 \tag{2.14}$$

Nos variables semblent donc être très fortement corrélées, nous allons le vérifier via plusieurs tests.

```
   -0.7369    4.7751   -0.1993   -3.0185   -0.8204
    0.1885    4.5833   -0.4660   -3.2855   -1.0202
    0.0805   -0.2259   -0.2372    1.6087   -1.2262


ahat =

   -1.0000
   -1.0000
   -1.0000
   -1.0000


R2 =

     1

------Fin du traitement------
>>
```

Fig. 2.4 – Calculs sous Matlab

3. χ^2 ces variables prédit le mieux la survie ?

1 L'adage « les femmes et les enfants d'abord »

1.1 « Mr Murdoch, j'ai un enfant ! »

On se propose de réaliser un test du χ^2 afin de savoir si, oui ou non, les enfants ont été privilégiés lors de l'accident. On pose l'hypothèse H_0 qui stipule qu'aucun traitement de faveur n'a été accordé aux enfants.

Concrètement, on part donc du principe que les variables « *âge* » et « *survie* » sont indépendantes. Comme nous l'avons présenté dans la première partie de ce dossier, la variable « *âge* » est quantitative. Le test du χ^2 s'appliquant aux variables qualitatives, il va falloir classer les différents âges par catégories.

Dans notre cas, nous allons simplement différencier les enfants des adultes. Et comme la définition *d'adulte* peut varier en fonction des pays et des époques, nous ferrons différents tests avec différentes limites d'âges.

Pour le moment, nous fixons la **limite à 18 ans**.
Une étude des données nous permet de dresser le tableau de contingence suivant :

	Enfants	**Adultes**	**Total**
Survivants	106	607	713
Morts	163	1332	1495
Total	269	1939	**2208**

FIG. 3.1 – Tableau de contingence O

Il faut à présent bâtir l'hypothèse nulle à partir de ces observations. Cela est possible grâce à l'hypothèse d'indépendance des variables précédemment formulée. On calcule les valeurs comme suit :

$$E_{i,j} = \frac{\sum_{k=1}^{I} O_{k,j} \times \sum_{l=1}^{J} O_{i,l}}{n} \quad \text{avec ici :} \quad n = 2208 \tag{3.1}$$

$$\tag{3.2}$$

On obtient ainsi le tableau hypothétique E suivant :

	Enfants	**Adultes**	**Total**
Survivants	86,86	626,13	713
Morts	182,13	1312,86	1495
Total	269	1939	**2208**

FIG. 3.2 – Tableau hypothétique E

On dresse ensuite un ultime tableau dont chaque cellule est calculée à partir des cellules des deux précédents tableaux :

$$\chi^2_{i,j} = \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{3.3}$$

	Enfants	**Adultes**	**Total**
Survivants	4,215	0,585	4,800
Morts	2,010	0,279	2,289
Total	6,226	0,864	**7,089**

FIG. 3.3 – Test du χ^2

La valeur obtenue est **7,089**. Il faut maintenant établir la distance critique en fonction du nombre de degrés de liberté du test. Ici, nous avons 2 variables de 2 valeurs chacune. Nous avons donc : $(2-1) \times (2-1) = 1$ degré de liberté. D'après les tables du χ^2, la distance critique pour un risque admis de 5% avec 1 degré de liberté est de **3,84**.

On remarque que la valeur obtenue est supérieure à la valeur critique correspondante. Nous sommes donc en mesure de **rejeter l'hypothèse d'indépendance des variables.**

1.2 Enfant, ado ou adulte ?

Ces résultats sont valables si l'on considère que les enfants concernés par l'adage sont les passagers de moins de 18 ans. Seulement au début du $XX^{ème}$ siècle, on ne considérait pas un individu de 18 ans comme un « *enfant* ». Ainsi, afin de s'assurer que cette partie de l'adage reste valable pour une limite d'âge inférieure, nous avons renouvellé l'étude en fixant d'autres limites.

Nous obtenons finalement ces résultats :

Age	**Distance**
5	18,42
6	18,99
7	17,24
8	19,6
9	19,23
10	14,94
11	14,31
12	14,93
13	17,16

Age	**Distance**
14	15,61
15	15,14
16	10,5
17	7,65
18	7,09
19	5,54
20	1,26
21	0,5
22	0,44

FIG. 3.4 – Distance en fonction de la limite d'âge

On remarque que tant que la limite fixée est inférieure ou égale à 19 ans, l'hypothèse d'indépendance reste valable. De plus, les distances obtenues grandissent quand la limite décroit, ce qui nous prouve que plus les enfants étaient jeunes, plus ils étaient amenés à obtenir une place à bord d'un canot de sauvetage.

Ainsi, nous pouvons affirmer qu'une partie de l'adage a été respectée lors du naufrage.

1.3 Discrimination sexuelle ou galanterie ?

Il convient maintenant de s'intéresser à l'autre partie de l'adage, concernant les femmes. Nous allons comparer la dépendance des deux variables « *sexe* » et « *survie* ». S'agissant de deux variables qualitatives, un autre test du χ^2 convient parfaitement pour cette étude.

Nous dressons le tableau de contingence O suivant :

	Hommes	**Femmes**	**Total**
Survivants	316	397	713
Morts	1273	222	1495
Total	1589	619	**2208**

FIG. 3.5 – Tableau de contingence O

De la même manière que pour le test précédent, nous obtenons le tableau hypothétique :

	Hommes	**Femmes**	**Total**
Survivants	513,11	199,88	713
Morts	1075,88	419,11	1495
Total	1589	619	**2208**

FIG. 3.6 – Tableau hypothétique E : les variables sont indépendantes (hyp. H_0)

Enfin, on calcule la distance :

	Hommes	**Femmes**	**Total**
Survivants	75,722	194,382	270,104
Morts	36,114	92,705	128,819
Total	111,836	287,087	**398,923**

FIG. 3.7 – Test du χ^2

Nous admettons la même marge d'erreur que pour le précédent test. Le nombre de degré de liberté étant le même (c'est à dire $(2-1)\times(2-1)=1$), la distance critique reste **3,84**.

On remarque que la distance du test est franchement supérieure à la distance critique. L'hypothèse d'indépendance des variables « *sexe* » et « *survie* » est donc clairement rejetée. Toutefois, nous ne comparerons pas cette distance avec celle du précédent test puisque le nombre de femmes à bord du navire est bien supérieur au nombre d'enfants : ces observations n'ont pas le même degré de précision.

Ainsi, l'adage « *les femmes et les enfants d'abord* » a bien été respecté lors du naufrage.

2 Le prix du billet pour la survie du passager

2.1 « Les vrais hommes créent leur propre chance »

On pourrait effectivement se demander si, à l'instar du film *Titanic*, certains pots de vins n'auraient pas été fixés durant la nuit du 14 au 15 avril en échange de réservations de places à bord de l'un des canots de sauvetage. Nos données ne nous permettent pas toutefois de lire l'Histoire avec une telle précision. Nous nous contenterons donc de justifier le lien entre richesse et survie.

Des observations de prix de billets aussi précises que celles que nous avons ne nous sont pas d'une grande utilité pour répondre à la question : « *Un passager aurait-il pu prédire sa probabilité de survie en fonction du prix de son billet ?* »
Nous avons donc choisi de classer ces prix par catégories. Ces prix étant relativement étendus (de 3£ à 512£ le billet), nous avons jugé que 16 catégories seraient suffisantes pour bien représenter ces différents prix.

On dresse donc un nouveau tableau de contingence O :

Prix	**3 à 8**	**9 à 14**	**15 à 20**	**21 à 27**	**28 à 32**	...	**137 à 221**	**222 à 512**	**Total**
Survie	101	68	45	64	37	...	18	37	517
Mort	343	145	46	85	50	...	11	17	803
Total	444	213	91	149	87	...	29	54	1320

FIG. 3.8 – Tableau de contingence O

De la même manière que pour le test précédent, nous obtenons le tableau hypothétique :

Prix	**3 à 8**	**9 à 14**	**15 à 20**	**21 à 27**	**28 à 32**	...	**137 à 221**	**222 à 512**	**Total**
Survie	173,9	83,42	35,64	58,36	34,07	...	11,36	21,15	517
Mort	270,1	129,57	55,36	90,64	52,92	...	17,64	32,85	803
Total	444	213	91	149	87	...	29	54	1320

FIG. 3.9 – Tableau hypothétique E

Enfin, on calcule la distance :

Prix	**3 à 8**	**9 à 14**	**15 à 20**	**21 à 27**	**28 à 32**	...	**137 à 221**	**222 à 512**	**Total**
Survie	30,56	2,85	2,46	0,54	0,25	...	3,88	11,88	99,68
Mort	19,68	1,84	1,58	0,35	0,16	...	2,50	7,65	64,18
Total	50,24	4,69	4,04	0,90	0,41	...	6,38	19,53	**163,56**

FIG. 3.10 – Test du χ^2

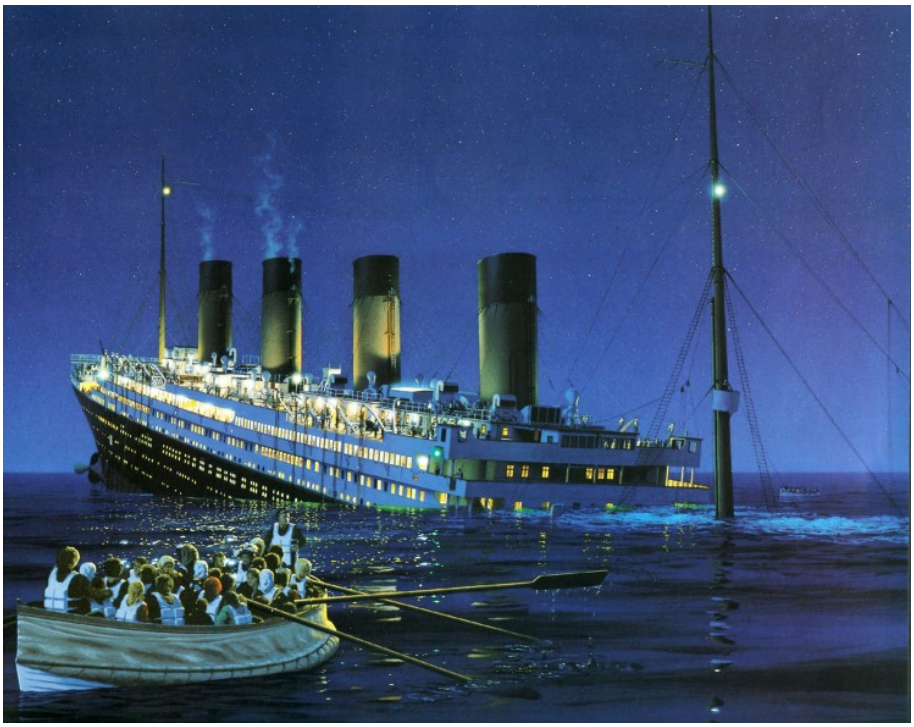
Nous admettons la même marge d'erreur que pour le précédent test. Pour un nombre de degré de liberté de $(2-1)\times(16-1)=15$, la distance critique est de **25**.

Nous remarquons donc que là encore, l'hypothèse d'indépendance des variables est clairement rejetée. La valeur élevée de la distance obtenue nous prouve que plus un passager avait payé son billet cher, plus il avait de chance de survivre lors du naufrage. Il vallait mieux être riche pendant la nuit du 15 avril 1912...

Conclusion

Ce projet nous aura permis de lire scientifiquement des données historiques liées à un évènement tragique. Nous sommes donc maintenant en mesure de dire que l'adage dont il a été question dans ce dossier a été respecté. De même, nous avons démontré que la survie d'un passager était très liée à sa richesse. Cette dernière conclusion nous prouve que des privilèges étaient encore valables sur ce paquebot titanesque.

D'une manière plus globale, ce projet de statistiques nous a permis de nous concentrer sur un sujet qui nous tenait beaucoup à cœur : retracer scientifiquement l'histoire d'un tel évènement historique n'est pas banal pour nous. Il faut bien dire que c'est peut-être la première et dernière fois que nous utilisons les mathématiques pour lire l'Histoire. Cette application de la M8 sur un sujet intéressant nous permet de conclure un semestre de principes du traitement de l'information sur une nouvelle approche concrète de la matière.



Vue d'artiste du naufrage du *Titanic* le 15 avril 1912

Annexes

1 Fiche technique des données du projet

Nom : Données précises sur les passagers ayant embarqué à bord du Titanic.
Taille : 2240 observations - 13 variables

Description des variables :

Variable	Sujet	Commentaire
1	Nom complet	
2	Age	(en années)
3	Catégorie	(classe / fonction)
4	Numéro du billet	
5	Prix du billet	(en livres)
6	Type de billet	
7	Groupe	
8	Ville d'embarquement	
9	Profession	
10	Bateau de sauvetage occupé	(0 si inexistant)
11	Numéro du corps repêché	(0 si inexistant)
12	Survie	(1 = oui, 0 = non)
13	Numéro de cabine	

2 A propos des données

Les données proviennent du site *Encyclopedia Titanica* et ont été finement récupérées à l'aide d'un code Php afin d'en sortir un fichier *.dat*.

http ://www.encyclopedia-titanica.org/titanic-passengers-and-crew/

Ces données sont inspirées des listes de passagers qui ont été publiées dans les journaux après le naufrage. Elles ont de plus été complétées par ce qu'a publié Walter Lord dans son livre *A Night to Remember*. Au moment de leur publication, ces listes étaient les plus précises qu'il soit ; mais en fait, ces données contenaient de nombreuses erreurs et omissions.

Encyclopedia Titanica a depuis tenté de rassembler les meilleures informations possibles à partir de diverses listes disponibles. Les données actuelles contiennent donc d'innombrables additions et corrections. Les listes ont été mises à jour avec les contributions des descendants des passagers et des historiens, en référence aux documents originaux.