École de Statistique, Biostatistique et Sciences Actuarielles

LSTAT2200 – Échantillonnage et Sondages

Projet

Auteur:
Warnauts Aymeric-87031800

Professeur:
Marie-Paul Kestemont

Louvain-la-Neuve
2021/2022

Introduction

Dans le but de collaborer avec les zirobourdons, nous aiderons la communauté à sélectionner deux échantillons de graines qu'ils nous acheterons et planteront intégralement. Le premier échantillon fourni nous permettra de collecter certaines informations après la plantation de cette série de graines sélectionnées, notamment sur le profit total effectué après la revente mais également sur le prix individuels des fleurs de notre échantillon après leur floraison. De ce fait, nous obtiendrons les profits marginaux de chaque type de graines (jaunes, bleues, rouges) et nous pourrons dès lors visualiser la variance propre à chaque groupe dans les retours sur investissment. Afin que les test effectués puissent fournir des données exploitables pour améliorer l'offre, il est impératif d'analyser finement les retours d'expérience. Après avoir récolté toutes les informations necessaires, nous nous efforcerons de constituer un deuxième échantillon optimal en terme de rentabilité, tout en nous efforcant de respecter scrupuleusement le budget de **15 000** euros de nos clients. Au terme de notre analyse, nous utiliserons nos résultats pour répondre à une série de questions que nos clients se posent afin de produire leurs prochaines fleurs indépendemment de nous.

	Poids	Couleur	Intensité	Régularité
	Min. :0.2000	Bleu :47893	Min. :1.000	Min. :1.000
	1st Qu.:0.4000	Jaune:32783	1st Qu.:3.000	1st Qu.:1.000
	Median :0.7000	Rouge:40670	Median :4.000	Median :2.000
	Mean :0.7959	NA	Mean :3.903	Mean :1.783
	3rd Qu.:1.0400	NA	3rd Qu.:5.000	3rd Qu.:2.000
	Max. :5.0100	NA	Max. :5.000	Max. :3.000

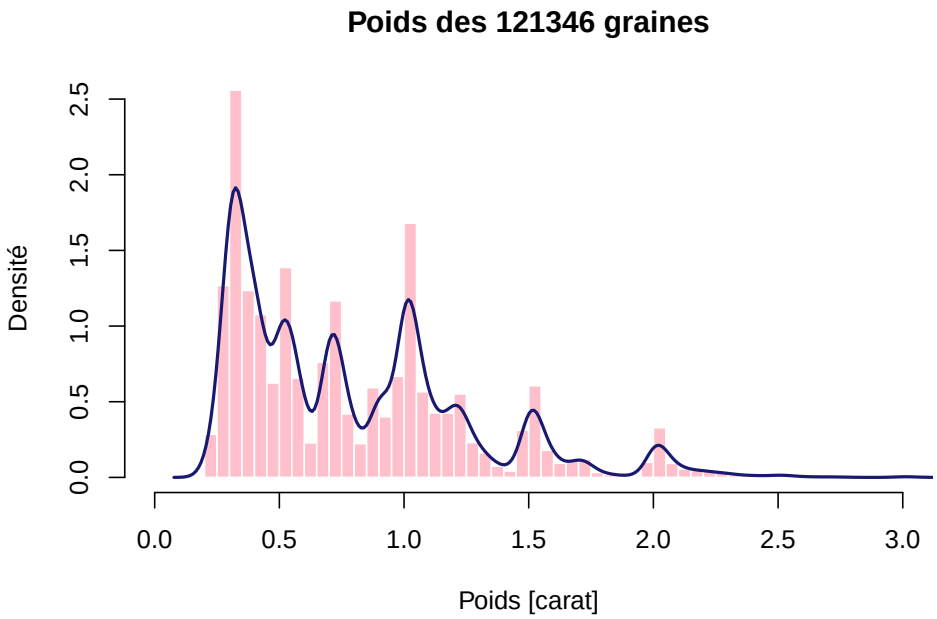
Dès lors, au vu de la nature du cas étudié, différenciant le coût de chaque type de graines plantées et étudiant la relation entre les caractéristiques de chacune d'entre elles avec la rentabilité attendue, il parrait judicieux de privilégier un échantillonnage représentatif de notre problématique. Le prix des fleurs n'étant pas connu, des questions préallable se posent; en effet, la rentabilité suite à la floraison de chaque type de graines est-elle la même? Si oui, il parraitra logique de privilégier le type de graines au coût le plus faible, étant donné l'équité du coût de plantation, dans notre cas, les graines bleues, pour maximiser les profits de notre client. De plus les préférences sont-elles liées à l'intensité de la couleur, la régularité de la forme et au poids de la graine au sein même de chaque type de fleurs? Dans notre cas, nous choisirons l'option qui maximise le profit, car une graine de mauvaise qualité au sein d'un niveau de la variable `Couleur` ne coûte pas moins cher qu'une graine dite de bonne qualité, aux caractéristiques optimisées.

A l'inverse, les préférences des consommateurs seraient-elles uniquement liées à la qualité de la fleur et donc aux autres variables telles que l'intensité et la régularité? Ces variables ont-elles à elles seulent un impact sur la vente des fleurs? Les consommateurs préfèrent-ils une fleur mixte mélangée où au contraire une fleur pure avec une haute intensité?

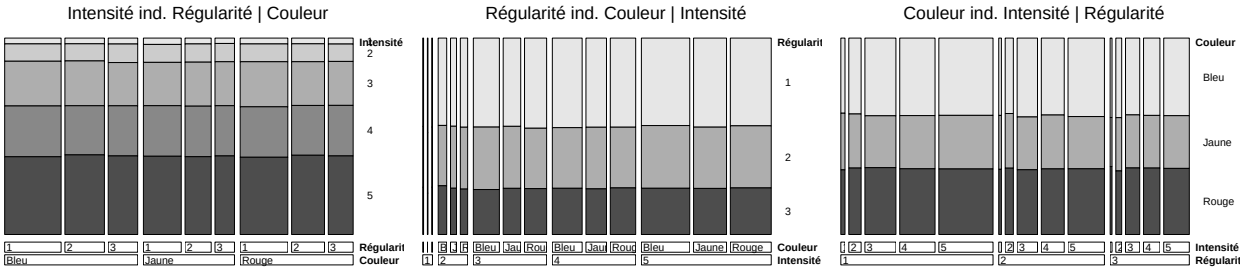
Nous tenterons de repondre à toutes ces questions, dans un premier temps a priori, sans connaissance du prix de revente de chaque fleur représentant clairement la propension à consommer de la population. A ce stade, il est très important de préciser que les premières graines récoltées seront plantées et donc non disponibles pour le deuxième échantillon formé, ce qui implique d'appréhender un plan d'échantillonnage global sans remise.

Il est très important de bien comprendre notre problématique et de répondre correctement à la première série de question que nos clients se posent. En effet, la corrélation entre notre variable Y `prix de vente`, inconnue à priori, et nos variables X fournies est une inconnue primordiale dans la première partie de ce projet. Ce n'est que grâce à cette information que nous pourrons apréhender un redressement convenable pour améliorer la précision des estimateurs et pour proposer notre échantillon final. Le principe général pour arriver à cette fin sera d'attribuer un poids w_i à chaque observation et de **caler** ces estimations sur des totatux connus, ces calages seront l'ajustement des poids.

Nous en discuterons dans la section suivante portant sur l'échantillonnage en lui même mais il s'avère déjà évident que la stratification selon la couleur semble judicieuse. En effet, si la rentabilité de nos graines ne dépendante que d'une ou plusieurs variables indépendamment de la variable `Couleur` on ne produira que des graines au coût le plus faible, c'est à dire les jaunes.



Nous pouvons a priori nous intéresser aux différentes formes d'associations que l'on peut retrouver dans notre dataset, en effet, lorsque l'on travaille avec des données catégorielles, ce qui est le cas pour 3 de nos 4 variables, il est intéressant de faire appel à la théorie de l'analyse de données discrètes. Nous allons donc analyser nos 3 variables `Couleur`, `Intensité` et `Régularité` à travers un tableau de contingence $I \times J \times K$ avec I, J, K respectivement les niveaux de nos trois variables pour conclure aux hypothèses d'indépendance et d'association entre elles. Ces hypothèses nous permettrons par la suite de construire un modèle log linéaire pour estimer les termes d'intéractions entre les variables si l'hypothèse de nullité de ceux-ci est rejettée.



Ce Mosaicplot représente parfaitement nos relations d'associations entre nos variables, car tout nos spinplot sont plats. Il est graphiquement évident que ces 3 variables et leurs interactions sont caractérisés par une association d'indépendance totale, elles sont indépendantes les unes des autres. Dès lors, au vu de cette analyse, on avance que *Couleur* II *Intensité* II *Régularité* et ces variables peuvent donc être analysées séparemment car il n'existe aucun lien entre elles.

Echantillonages

Premier Echantillon

L'échantillonnage en population finie, pour une entreprise, consiste à créer une sélection représentative de la population totale auprès de laquelle elle validera son offre. Ce groupe test permet d'appréhender le comportement global de la population et d'adapter sa proposition en fonction des retours de l'échantillon. L'échantillonnage en deux étapes proposé ici permet de comprendre les forces et les faiblesses des graines à commercialiser avant de proposer un échantillon final suite à un redressement. En définissant un échantillon

plus restreint dans un premier temps, le nombre de données collectées est évidemment réduit, faisant ainsi économiser des ressources pour planifier notre échantillon final.

Dans cette optique de sélection, nous nous aiderons d'un plan de sondage, p sur U définissant une loi de probabilité sur les parties de U et d'un algorithme d'échantillonage pour sélectionner un échantillon selon le plan de sondage choisi. Ici, à la différence des lois de probabilités classiques, l'aléatoire ne porte pas sur la variable mais sur le sous-ensemble d'individus observés. Nous devrons définir a priori les probabilités d'inclusion π_k, c'est à dire la probabilité que l'unité k soit retenue dans l'échantillon, pour utiliser un plan de sondage qui respecte ces probabilités. En effet, la connaissance des probabilités π_k permet une estimation sans biais d'un total sous le plan de sondage aléatoire à probabilités inégales, appelé estimateur de **Horvitz-Thompson**:

$$\hat{\tau}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

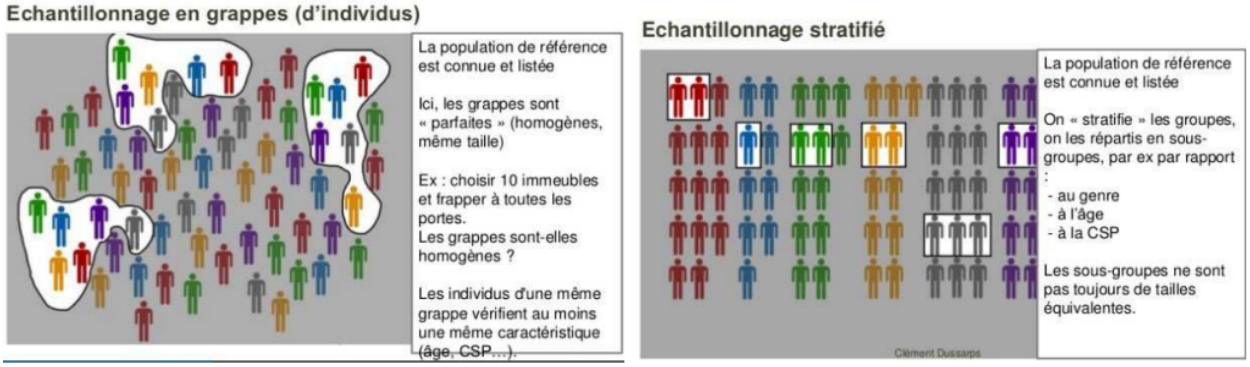
Où un individu k de l'échantillon représente $d_k = \frac{1}{\pi_k}$ individus de la population. De ce fait si certaines probabilités d'inclusion sont nulles, le π-estimateur est biaisé, ce qui peut arriver si on décide de laisser une partie de la population de coté. Cet estimateur demande la connaissance des probabilités d'inclusion d'ordre 1. Nous pourrons utiliser ces estimateurs des totaux de nos variables pour estimer des fonctions des totaux tels qu'un ratio $\frac{t_y}{t_x}$. Si certaines probabilités d'inclusion sont nulles, le π-estimateur est biaisé. On peut observer ce genre de biais lors d'un défaut de couverure ou de **cut-off sampling**. Grâce à l'estimation de la variance de cette estimateur, nous pourrons, via une approximation normale de $\hat{\tau}_{y\pi}$, obtenir des intervalle de confiance pour cet estimateur.

D'après la formule de **Yates-Grundy**, la variance est nulle si les probabilités d'inclusion sont proportionnelles à la variable d'intérêt. En pratique il est parfois plus simple de définir ces probabilités d'inclusion proportionnellement à une mesure de taille. En effet, si les individus peuvent être de tailles très différentes, on utilise les probabilités d'inclusion pour lisser les rapports $\frac{y_k}{\pi_k}$. Dès lors, après une stratification en suivant la variable `Couleur` de notre data set, nous pourrions calculer les probabilités d'inclusion de chaque type de graines. Si n désigne la taille d'échantillon souhaitée, les probabilités d'inclusion proportionnelles à une variable sont données par:

$$\pi_k = n \frac{x_k}{\sum x_l}$$

où la variable x_k doit être connue avant le tirage pour chaque individu k de U.

Une première approche de notre population aurait pu, dans une logique de coûts, nous pousser vers un **sondage en grappes et à plusieurs degrés**. En effet, en divisant notre population en M grappes, mini-populations U_g, $g = 1..., M$, en fonction de la variable `Couleur`, nous aurions pu tirer aléatoirement des unités collectives. Tout les individus des différentes grappes selectionnées, quelles quelles soient, n'auraient pas pu être sélectionnés dans leur intégralité. De ce fait un échantillon de chacune des grappes tirées aurait été prélevé. Malheureusement la composante aléatoire dans la sélection même des grappes nous posera certainement problème pour répondre de manière complète aux questions des Zirobourdons. De plus, nous devons garder en tête que la variable d'intérêt reste à ce stade inconnue et le calcul des estimateurs produits par ce type de sondage est complexe. De plus, si les grappes formées sont plus homogènes que ne le voudrait le hasard, le sondage en grappe est moins précis qu'un sondage élémentaire. Afin d'optimiser au maximum la rentabilité du second échantillon proposé, et disposant d'une vaste base de sondage, nous privilégierons d'autres méthodes d'échantillonnages plus complexes dans leurs analyses mais également plus précises.



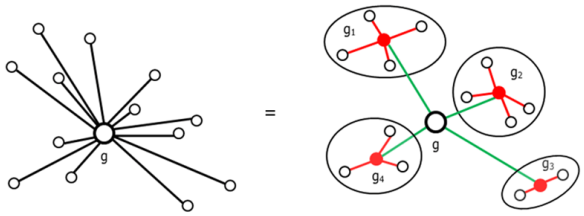
La stratégie adoptée sera la suivante, nous planifierons une **stratification à allocation proportionnelle** (STP) pour notre premier échantillon et une **stratification à allocation optimale** (STO) **en terme de coût** pour maximiser le profit en proposant notre échantillon final. Cette stratégie nous permettera de simplifier les estimateurs en prétendant des poids w_i égaux, néanmoins nous devrons rester objectif car les variances ne seront pas réduites de manière optimale. Ce type de sondage nous permettra de nous assurer que toutes les strates, c'est à dire les différents niveaux de la variable `Couleur` seront représentées. De cette façon, nous pourrons affiner nos questions et apporter des réponses plus précises quant au choix du type de fleur à sélectionner. Nous rappelons que les coûts différents pour chaque couleur de graines justifient notre choix de stratification.

Afin de mesurer la précision de notre échantillonnage, nous utiliserons notamment les mesures suivantes:

- Le biais de notre estimateur
- La variance de notre estimateur
- L' Erreur Quadratique Moyenne

Pour qu'on puisse considérer qu'un échantillon est représentatif de la population étudiée, il doit être de taille suffisante par rapport à la population, et posséder les mêmes caractéristiques que celle-ci. C'est dans cette optique que nous préleverons notre premier échantillon afin de pouvoir observer la réaction globale du prix fixé pour les différentes graines.

A ce stade, il est primordial de rappeler les concepts d'inertie qui régissent dans bon nombre d'études de classification mais également de stratification comme nous le verrons par la suite. En effet, le théorème d'**Huygens** explique parfaitement la décomposition de la variance totale liée à une variable après stratification de celle-ci.



$$\sigma^2 = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu)^2$$

On peut alors définir le rapport de corrélation $\eta^2 = \frac{\sigma^2_{inter}}{\sigma^2}$, si $\eta^2 = 0$ alors $\mu_h = \mu, \forall h$. A l'inverse, si $\eta^2 = 1$ alors $\sigma^2_{intra} = 0$ et donc il n'y a pas de variance dans les strates formées, elles sont parfaitement homogènes, c'est l'objectif que nous viserons. Dans le type de sondage stratifié, la probabilité d'inclusion est

$$\pi_i = \frac{n_h}{N_h} = f_h$$

Stratification à allocation proportionnelle

Dans le cas d'un sondage aléatoire stratifié à tailles proportionnelles, on définit un taux de sondage f qui sera le même pour toutes les strates: $n_h = f.N_h$ et la probabilité d'inclusion sera alors:

$$\pi_i = f_h = f \implies \frac{n_h}{n} = \frac{N_h}{N}$$

En effet, tout les individus dans U ont la même probabilité d'inclusion. Dès lors pour estimer μ, le résultat moyen en statostique dans la population, on fera appel à l'estimateur:

$$\hat{\mu}_{STP} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} \equiv \widehat{\mu}_{échantillon}$$

L'estimateur de la moyenne de la population est donc **non-biaisé** dans le cas d'une stratification à tailles proportionnelles. Et la variance de cet estimateur est donné par:

$$Var(\hat{\mu}_{STP}) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - f_h) \frac{\sigma^2_{h;corr}}{n_h} = \frac{(1-f)}{n} \sigma^2_{intra;corr}$$

Rappelons que le facteur $(1 - f)$ nous donne le gain de variance dû au tirage sans remise, appelé correction de population finie. De plus, si le taux de sondage est faible, la variance ne dépend que de la taille d'échantillon n. Dans notre cas, où la taille de la population étudiée N est grande, on peut montrer que $s^2_{intra;corr} = s^2_{intra} + \frac{1}{N} \sum_{h=1}^{H} s^2_{h;corr}$ et on peut determiner l'effet de sondage:

$$D(STP|PESR) \approx \frac{\sigma^2_{intra}}{\sigma^2} = 1 - \eta^2 \leq 1$$

On se rend alors compte que la stratification avec une variable très liée à la variable d'intérêt est primordialle. C'est suite à cette démonstration que nous nous plongerons dans des études préalables trouvée ici pour valider nos intuitions de stratifications en fonction de la variable `Couleur`. Après des recherches empiriques et universitaires sur le thème de l'inluence de la couleur et de l'intensité de celle-ci sur les préférences d'un consommateurs, une étude sur la neuropsychologie du consommateur developpant l'influence de la couleur en marketing où l'on cite notament **Huygens** nous apporte beaucoup d'éclaircissements sur notre problématique. Nous nous pencherons brièvement sur les conclusions liées aux couleurs de nos graines, en supposant que celles-ci nous donnerons des fleurs aux mêmes teintes. L'étude avance la théorie trichromatique humaine qui développe qu'en ajustant la quantité de bleu, de jaune et de rouge ou de vert, on peut reproduire n'importe quelle longueur d'onde échantillonnée en s'appuyant sur des concepts tels que le "color after-effects" ou effet consécutifs de couleur. Ce qui suit est d'autant plus intéressant car, d'après des ecpériences effectuées, dans le cas d'une lumière bleue et d'une lumière rouge de 7 candélas au mètre carré, la lumière bleue donnera l'impression d'être plus lumineuse, cette diffférence de perception de l'intensité entre les couleurs devra être retenue dans nos interprétations. L'étude avance notament que dans les années 80, près de 2 millions de Français présentaient une déficience prononcée pour le rouge et le vert, conséquence d'un daltonisme prononcé. Dès lors, afin de transmettre une information fondée sur la couleur, et perceptible par ces personnes, il convient d'éviter des dégradés uniques de rouge et de jouer sur l'intensité d'une couleur plûtot que sur la seule teinte. Ce genre de pathologie réaffirme paradoxalement l'universalité des anatomies, psychologies en terme de visualisation des préférences et des couleurs. Un point essentiel sur lequel cette étude porte est la multimodalité et la polysensorialité de nos sensations, le rouge est chaud, le vert est frais. Ce n'est pas pour rien que des grandes marques de l'industrie terrestre comme **Coca-cola**, **Netlfix**, **Nintendo** se sont appropriées le rouge. De plus, la symbolique de la couleur des fleurs offertes est encore très encrée sur terre;

le rouge symbolise l'amour, la passion, le bleu la pureté et finalement le jaune associé notament à l'infidélité, la trahison. De plus, la stratification devrait être choisie de façon à ce que la dispersion à l'intérieur des strates soit minimisée. Les nombreuses variables catégiorielles de notre dataset nous permettent de stratifier très finement notre population.

Pour des populations élevées, la taille de l'échantillon N_H a sélectionner pour atteindre une représentativité raisonnable se calcule avec la formule de **Cochran** ici:

$$n_{cochran} = t^2 p \frac{(1-p)}{m^2}$$

Pour notre taille de population de **121346** pour un niveau de confiance de **95%** et une marge d'erreur de **5%** il nous faudrait choisir une taille d'échantillon de longueur **383** et de longueur **663** pour un niveau de confiance de **99%** avec la même marge d'erreur. On observe qu'au dela d'une population de **100000** observations à échantillonner, les échantillons représentatifs à différents seuils de marge d'erreur ne varient plus. On s'aperçoit qu'a $\pm$ **3%** la taille de ceux-ci se stabilise entre **1000** et **1100** ce qui rentrerait dans notre budget pour les côuts fixés si nous devions sélectionner 2 échantillons de cette taille. Rappelons que notre but sera d'estimer le plus precisiement possible la corrélation entre les les différentes variables et le prix de chaque fleurs après notre premier échantillon. De plus, nous devrons estimer au mieux la variance du prix de chaque fleurs afin de sélectionner notre échantillon final à taille optimale en terme de coût.

Ici, malgré notre faible taux de sondage f_h de **0.009** nous nous attendons de part la nature du sondage effectué sur le prix de vente des fleurs à un taux de réponses extrèmement élevé, ce qui nous permet d'obtenir une fiabilité dans nos résultats. Il serait également interressant de déjà calculer le coût total de notre échantillonnage en strates de longueur total **1050**. Pour cela, rien de plus simple, il nous suffit de recenser le nombres de graines de chaques catégories de la variable `Couleur`de notre subset et de les pondérer par leurs coûts totaux respectifs.

$$D\acute{e}pense_{1^{er}\acute{e}chantill} = 6x_{bleue} + 7x_{jaune} + 8x_{rouge} = 7281$$

$$Budget_{2^{ème}\acute{e}chantill} = 15000 - 7281 = 7719$$

A noter qu'on se base sur des apréhensions statistiques plus que sur des théories économiques. En effet, on prefère privilégier une analyse de la variance et une planification recouvrant l'entiéreté des combinaisons catégorielles possibles plûtot que sur une logique de rareté où les fleurs les plus rares (seulement **220** graines **jaune**, **intensité=1**, **régularité=3** disponibles) rapporteraient le plus. Nous effectuerons un redressement adéquat en fonction des observations effectuées après la première récolte. La rareté des graines jaunes se fait d'ailleurs ressentir dans chaque strates à `Intensité` et `régularité` identiques.

Une fois notre première plantation récupérée et donc les prix de vente, la variable d'intérêt, connus pour l'échantillon proposé, il nous faut estimer la variance propre à chaque strate $\sigma^2_{h;corr}$ afin de pouvoir optimiser selon nos coûts. Nous avons choisi un échantillon représentatif de la population et après avoir effectué les hypothèses préalable, nous nous permettrons de traiter l'échantillon obtenu comme indépendant de la population afin d'en tirer des conclusions statistiques tirées des fondement de l'analyse de données. Nous globaliserons par la suite ces conclusions à notre population avant de choisir un échantillon final. Il nous faudra également construire un modèle adéquat, tester la significativité des coefficients d'intéraction de nos variables. Il parrait évident qu'une analyse de la distribution de notre variable cible `Prix` nouvellement connue est primordiale pour choisir le modèle adéquat.

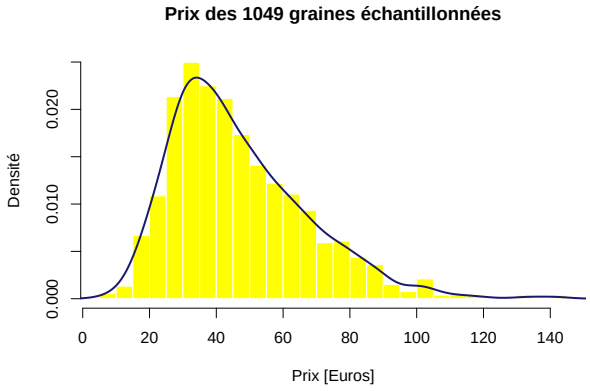
L'inspection visuelle n'étant généralement pas fiable, il est possible d'utiliser un test de significativité, comparant la distribution de l'échantillon à une distribution Normale, appellés test de **kolmogorov-Smirnov** et de test de **Shapiro-Wilk**. L'hypothèse nulle de ces tests est H_0 : ***La distribution de l'échantillon est Normale***. La méthode de **Shapiro-Wilk** fournit une meilleure puissance pour ce test et est basé sur la corrélation entre les données et les scores normaux correspondants. Nous pourrons également tester cette hypothèse de normalité au sein même des niveaux d'une variable catégorielle, ce que nous ferons pour étudier

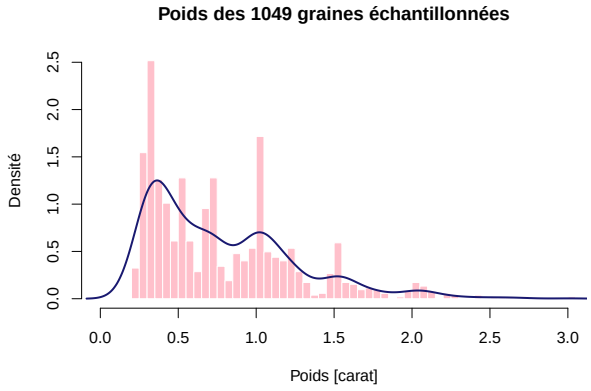
la distribution de la variable `Prix` pour les fleurs de différentes couleurs. Et notre test nous donne une p-valeur de $3.0420541 \times 10^{-20}$, ce qui nous permet de rejeter l'hypothèse nulle de normalité.

Nous pouvons dès à présent nous pencher sur le profit effectué après notre premier échantillonnage. En effet, en sommant les variables `Prix` et `Coût` nous pourrons faire la différence des deux métriques des totaux. Nous obtenons pour notre premier échantillon un prix total de **49143.34** et un coût total de **7281**, ce qui nous ramène à un profit de **41862.34** euros.

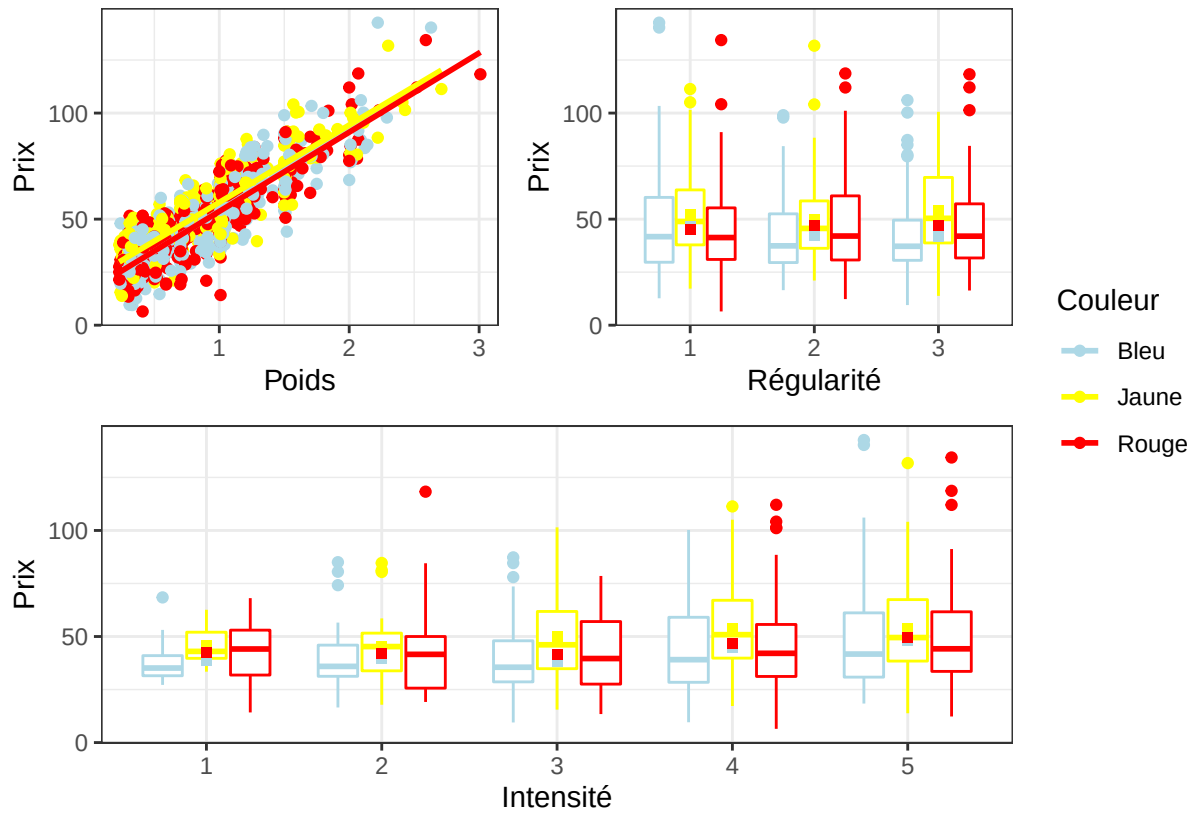
variable	statistic	p
Prix	0.9395302	0

Couleur	variable	statistic	p
Bleu	Prix	0.9127844	0e+00
Jaune	Prix	0.9584027	3e-07
Rouge	Prix	0.9402269	0e+00





L'interpretation de nos Boxplots nous permet déjà de poser des hypothèses à priori, avant même l'interprétation des tests d'hypothèses qui suivent. En effet, nous rappelerons que les bords inférieurs et supérieurs du carré central représentent le premier et le troisième quartile de la variable représentée sur l'axe vertical. le trait horizontal représente la médiane. Enfin, les moustaches s'étendent de chaque côté du carr, jusqu'aux valeurs minimales et maximales avec une exception, les outliers (valeurs extrêmes) sont représentés sous forme de points. Nous avons également choisis de représenter les moyennes des différents niveaux par couleurs sous forme de carrés afin de voir si ces moyennes se confondent.



Pour répondre aux questions des Zirobourdons, il nous faudra determiner s'il existe un lien entre les différentes variables et la variable `Prix` et si nécessaire déterminer le sens de cette interaction ou l'ordre de significativité entre les différents niveaux de nos variables catégorielles. Ces conclusions nous permetterons de choisir au mieux notre deuxième échantillon.

Tout d'abord, pour étudier la corrélation entre une variable continue (`Prix`) et une variable Catégorielle (`Couleur`, `Intensité` et `Régularité`) nous aurons recours à l'analyse de la variance (ANOVA) à un facteur qui permet de comparer les moyennes d'échantillon. L'Objectif de ce test est de conclure l'influence d'une variable catégorielle sur la loi d'une variable continue à expliquer. Ce que nous tenterons d'expliquer sera l'égalité des moyennes dans les différents niveaux des variables catégorielles; autrement dit, $H_O : \mu_1 = ... = \mu_k$.

Enfin, pour étudier la corrélation entre deux variables continues, c'est à dire dans le cas entre le `Prix` et le `Poids`, il existe un test pour analyser la dépendance entre elles, le test de corrélation de **Pearson**. L'hypothèse nulle testée est ici tout simplement l'indépendance entre les variables. En effet, le coefficient de Peaeson permet de mesurer le niveau de corrélation entre les deux variables. Il renvoie une valeur entre **-1** et **1**. Si sa valeur se rapproche de ces limites, elles sont respectivement négativement ou positivement corrélées. Au contraire un coefficient proche de **0** avance une décorrélation.

Le test le plus utilisé pour tester la relation entre une variable quantitative et une variable qualitative à deux modalitées est le test de Student, mais dans notre cas, comme avancé précedemment nous devrons utiliser une ANOVA. Cette méthode étant un test paramétrique, elle requière certaines conditions sur la distribution de probabilité des données, comme **(a)** la normalité de la population, **(b)** l'homoscédasticité des variances conditionnelles et **(c)** des sous-échantillons des modalités de la variable catégorielle indépendants. l'idée qui sous-tend le test ANOVA est le suivant; si la variation moyenne entre les groupes est suffisamment importante par rapport à la variation moyenne au sein des groupes, on peut conclure qu'au moins la moyenne d'un des groupes n'est pas égale aux autres.

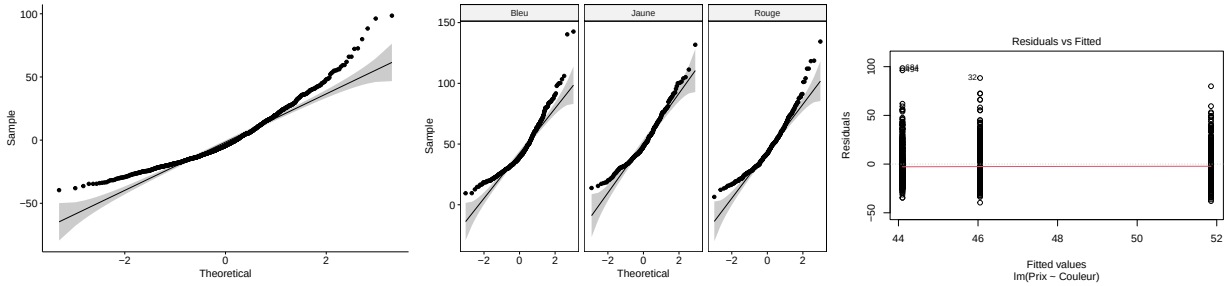
Questions de nos clients

Existe-t-il un lien entre la couleur de la graine et le prix de la fleur ? Si oui, pour quelle(s) couleur(s) pouvez-vous conclure qu'il existe une différence de prix ?

Premiérement il est intéressant de s'attarder sur les éventuels outliers liés à l'analyse croisées de ces deux variables. La presence de celles ci nous force à utiliser un test robuste de l'ANOVA dans notre cas. Nous remarquerons que les deux observations extrêmes appartiennent toutes deux au niveau `Bleue` de la variable `Couleur`.

	Couleur	Obs	Prix	Poids	Intensité	Régularité	Coût	is.outlier	is.extreme
4	Bleu	55089	140.32	2.63	5	1	6	TRUE	TRUE
5	Bleu	76838	142.62	2.22	5	1	6	TRUE	TRUE

Nous preciserons que nous sommes en présence de 18 observations considérées comme outliers. Nous continuons notre anlayse préalable en rappelant que l'hypothèse de normalité pour tout les groupes de la variable `Couleur` avaient déjà été rejetée plus haut via le test de **Shapiro-Wilk**.



C'est suite à cette incertitude sur la normalitée de nos données que nous privilégierons un test de **Kruskal-Wallis**, qui est l'alternative non-paramétrique. Néanmoins, au vu de notre plot et de notre test de **Levene** qui avance une p-valeur de 0.599141, nous pouvons supposer l'homogénéité des variances dans les différents niveaux de la variable `Couleur`.

Nous testons donc si il y a une différence significative entre les prix moyens des fleurs dans les 3 couleurs différentes de notre échantillon:

.y.	n	statistic	df	p	method
Prix	1049	32.76705	2	1e-07	Kruskal-Wallis

.y.	n	effsize	method	magnitude
Prix	1049	0.029414	eta2[H]	small

Il y a donc des différences statistiquement significatives du prix moyens entre les niveaux (p-valeur = 0) au vu du test de **Kruskal-Wallis** mais nous remarquons aussi une petite taille de l'effet qui indique au'un pourcentage de 2.94 % de la variance du Prix est expliquée par la Couleur de la fleur.

Pour répondre de manière compléte à cette question nous nous baserons sur nos résultats pour tester des comparaisons multiples par paires entre nos 3 niveaux de Couleur. Le test de **Dunn** nous permettra d'identifier les groupes différents:

.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
Prix	Bleu	Jaune	414	283	5.652953	0.0000000	0.0000000	****
Prix	Bleu	Rouge	414	352	1.629499	0.1032073	0.3096220	ns
Prix	Jaune	Rouge	283	352	-3.981345	0.0000685	0.0002056	***

Nous pouvons donc conclure cette question en affirmant plusieurs choses; dans un premier temps nous avons démontré statistiquement qu'il existe une différence significative de prix entre les fleurs **Jaunes** et les 2

autres couleurs de fleurs. Ceci vient confirmer bon nombres d'analyses effectuées à priori. De plus, nous pouvons retenir une information primoridale quant auu fait que la différence de prix entre les fleurs **Bleues** et **Rouges** n'est pas significative. En prenant du recul sur la discrimination du coût de nos graines, nous pouvons déjà conclure qu'au regard des intéractions simples entre le Prix et la Couleur, nous prefererons produire des fleurs bleues, moins cheres à la production, aux rouges dans le but de maximiser notre profit.

Existe-t-il un lien entre l'intensité de la couleur de la graine et le prix de la fleur ? Si oui, quelle est la nature de ce lien et quels sont les niveaux d'intensité pour lesquels la différence de prix est significative ?

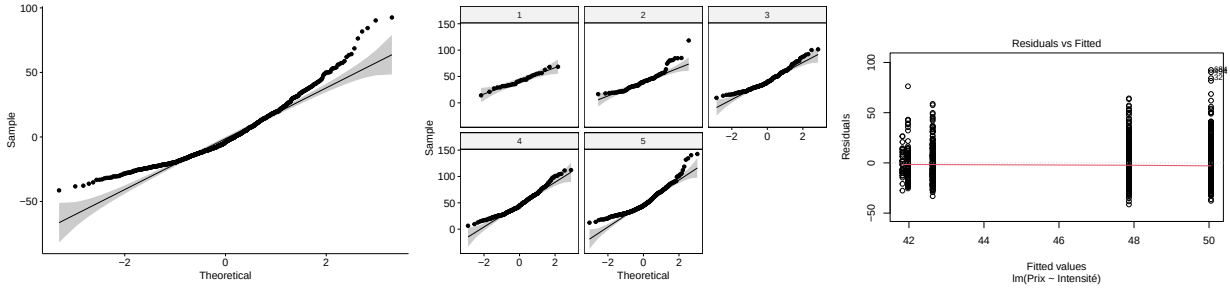
Nous devons à nouveau tester les hypothèses préalables à un test ANOVA robuste ou non afin de déterminer si nous devrons utiliser des tests alternatifs comme vus précédemment. Nous remarquons très rapidement la présence d'une valeur extrême apparternant au niveau **3** de notre variable catégorielle `Régularité`.

	Intensité	Obs	Prix	Poids	Couleur	Régularité	Coût	is.outlier	is.extreme
2	2	26922	118.28	3.01	Rouge	3	8	TRUE	TRUE

Nous préciserons à nouveau que nous pouvons compter 22 observations considérées comme outliers. Nous pouvons compléter nos interprétations graphiques quant à l'hypothèse de normalité pour les niveaux de la variable `Intensité` en effectuant le test de **Shapiro-Wilk**:

Intensité	variable	statistic	p
1	Prix	0.9795363	0.7721654
2	Prix	0.8931058	0.0000014
3	Prix	0.9517643	0.0000004
4	Prix	0.9594044	0.0000007
5	Prix	0.9218831	0.0000000

Ici, même si les résidus du niveau **1** de notre variable `Intensité` peuvent être considérés comme normaux, nous ne pouvons pas en dire autant pour les **4** autres niveaux de cette variable.



Nous nous dirigerons donc vers une alternative non-paramétrique de notre test d'égalité des prix moyens dans les différents niveaux de notre variable catégorielle. Nous choisirons la même stratégie et privilégirons un test de **Kruskal-Wallis**:

Nous testons donc si il y a une différence significative entre les prix moyens des fleurs dans les **5** niveaux d'intensité de notre échantillon:

.y.	n	statistic	df	p	method
Prix	1049	25.61131	4	3.79e-05	Kruskal-Wallis

.y.	n	effsize	method	magnitude
Prix	1049	0.0207005	eta2[H]	small

nous démontrons donc, une fois de plus au'avec une p-valeur de 3.79×10^{-5} nous devons rejeter l'hypothèse nulle pour privilégier l'hypothèse alternative tel que au moins une médiane de la population d'un groupe est

différente de la médiane de la population d'au moins un autre groupe. On rappelle que le modèle sous cette hypothèse se définit avec θ la médiane globale et τ_j, l'effet de traitement j, de la sorte:

$$X_{i,j} = \theta + \tau_j + \epsilon_{i,j} \text{où} i = 1, ..., n_j; j = i, ..., k$$

Pour répondre de manière compléte à cette question nous nous baserons sur nos résultats pour tester des comparaisons multiples par paires entre nos 5 niveaux d' Intensité. Le test de Dunn nous permettra d'identifier les groupes différents:

.y.	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
Prix	1	2	33	93	-0.4341377	0.6641885	1.0000000	ns
Prix	1	3	33	236	-0.2362995	0.8132003	1.0000000	ns
Prix	1	4	33	269	1.1317254	0.2577499	1.0000000	ns
Prix	1	5	33	418	1.6784348	0.0932622	0.9326224	ns
Prix	2	3	93	236	0.3597835	0.7190091	1.0000000	ns
Prix	2	4	93	269	2.4665706	0.0136414	0.1364138	ns
Prix	2	5	93	418	3.4143224	0.0006394	0.0063941	**
Prix	3	4	236	269	2.8328357	0.0046137	0.0461371	*
Prix	3	5	236	418	4.2667367	0.0000198	0.0001984	***
Prix	4	5	269	418	1.2121682	0.2254480	1.0000000	ns

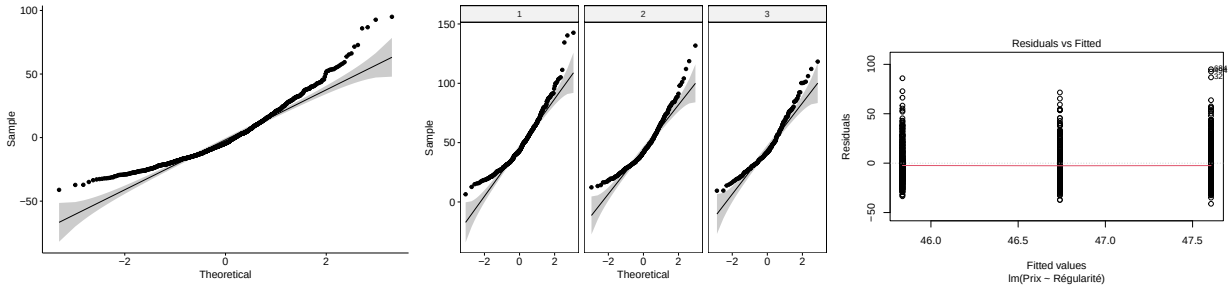
Nous pouvons donc conclure cette question en affirmant qu'il existe bel et bien un lien entre l'intensité de la couleur de la graine et le prix de la fleur même si celui-ci est faible car seulement 2.07% de la variance du prix est expliquée par l'intensité de la couleur de la graine. En effet, nous remarquons qu'il existe une différence significative pour les graines de couleur d'intensité 4 ou 5 par rapport aux niveaux de la variable `Intensité`. Néanmoins, l'interprétation des résultats est moins évidente que précédemment; on s'apercoit par exemple qu'après ajustement, notre p-valeur pour la différence entre le niveau **2** et **4** est non significatif alors qu'il l'était avant ajustement au niveau 5% (p-valeur de **0,014**), de plus la distribution assurée précédemment normale au sein du groupe **1** de la variable ne nous permet pas de distinguer une différence significative entre les niveaux **4** et **5** et ce niveau minimal de **1**. Néanmoins nous relevons un p-valeur non-ajustée proche du seuil de rejet à **5** % pour la différence entre le niveau **1** et le niveau **5**.

Existe-t-il un lien entre la régularité de la graine et le prix de la fleur ? Si oui, quelle est la nature de ce lien et quels sont les niveaux de régularité pour lesquels la différence de prix est significative ?

Pour répondre à cettet question nous procédons comme précédemment et nous remarquons premièrment que notre échantillon ne comporte pas de valeurs extrèmes lorsqu'on regroupe nos données par régularité.

Régularité	Obs	Prix	Poids	Couleur	Intensité	Coût	is.outlier	is.extreme

Nous dénotons tout de même 22 observations pouvant être considérées comme outliers. Une fois de plus notre analyse graphique nous avance une non-normalité des distributions, mais nous verifierons statistiquement.



Le test de **Shapiro-Wilk** nous donne les sorties suivantes:

Régularité	variable	statistic	p
1	Prix	0.9376346	0
2	Prix	0.9344220	0
3	Prix	0.9435732	0

Nous rejettons donc encore une fois l'hypothèse nulle de normalité dans tout les niveaux de notre variable `Régularité`, et nous nous dirigons une fois de plus vers le tet de **Kruskal-Wallis** pour tester la différence entre les prix moyens dans nos niveaux **1**, **2** et **3**.

.y.	n	statistic	df	p	method
Prix	1049	1.442824	2	0.486	Kruskal-Wallis

.y.	n	effsize	method	magnitude
Prix	1049	-0.0005327	eta2[H]	small

Cette fois ci, l'interprétation de cette varibale se fait assez aisément car le test effectué nous démontre clairement qu'avec une p-valeur de **0,486** nous ne pouvons pas rejetter l'hypothèse nulle qui avance $H_O : \tau_1 = ... = \tau_j$ et nous ne pouvons pas conclure que l'effet de traitement est significativement différent dans les **3** niveaux de la variable `Régularité`, il n'y a pas de dominnance stochastique entre les sous-échantillons.

Le poids de la graine a-t-il une influence sur le prix de la fleur ?

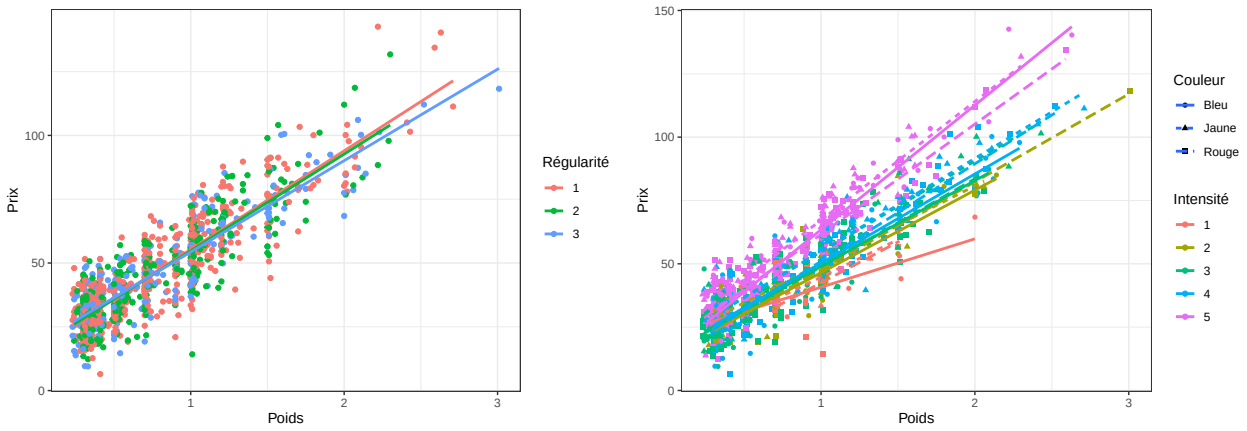
Nous commencerons ici par rappeler que notre échantillon formé, bien qu'il n'ai pas été stratifié sur une transformation catégorielle de la variable `Poids`, représente assez bien la distribution de probabilité de cette variable dans la population totale; en effet, on peut aisément arriver à cette conclusion en comparant les deux représentations graphiques de la densité de cette variable. Graphiquement, il est déjà évident que le poids et le prix sont positivement corrélés. Ce qui est encore plus frappant, c'est la manière dont les droites de régression des modèles linéaires estimés de chacune des différentes couleurs se confondent, avec la même pente. Dans le but de visualiser au mieux nos interactions mais aussi pour vérifier nos hypothèses préalables, nous décidons de représenter également notre régression en groupant nos données en fonction des 2 autres variables catégorielles `Régularité` et `Intensité` par la suite.

Statistiquement, nous calculerons la corrélation des rangs, caractérisée par le coefficient de **Spearman**. En effet, cette méthode nous permet de faire abstraction de l'hypothèse forte d'association linéaire sur les données avancée par le coefficient de **Pearson**. Plûtot que de se baser sur les valeurs des variables, cette corrélation va se baser sur leurs rangs, sur leurs positions parmi les différentes valeurs prises par les variables. On obtient un coefficient de **Spearman**, moins sensible aux valeurs extrêmes et donc plus robuste, de 0.84 et si on utilise le coefficient linéaire de **Pearson** on obtient 0.88. Ceci établi une forte corrélation positive de nos deux variables.

Lorsqu'on est en présence d'une association linéaire marquée entre deux variables, il peut être intéressant d'en évaluer les coefficients pour établir des conclusions plus précises sur les pentes et oridonnées à l'origine.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.23920	0.5836552	29.53661	0
Echant1$Poids	37.66879	0.6358377	59.24278	0

Dès lors, nous pouvons maintenant quantifier ce rapport entre la variable réponse et la variable `Poids`. Nous savons maintenant que le coeffcient qui lie ces deux variables est significatif (p-valeur de **0**) et vaut **37,66** ce qui signifie que lorsqu'on augmente le poids de notre graine d'un carat, nous augmentons notre profit estimé de **37,66** euros. Dans l'optique de maximiser notre profit, nous analyserons si la pente de cette droite de régression varie en fonction des niveaux des différentes variables catégorielles. Ceci nous amenera à discuter des termes d'interactions qu'il peut exister entre nos variables.

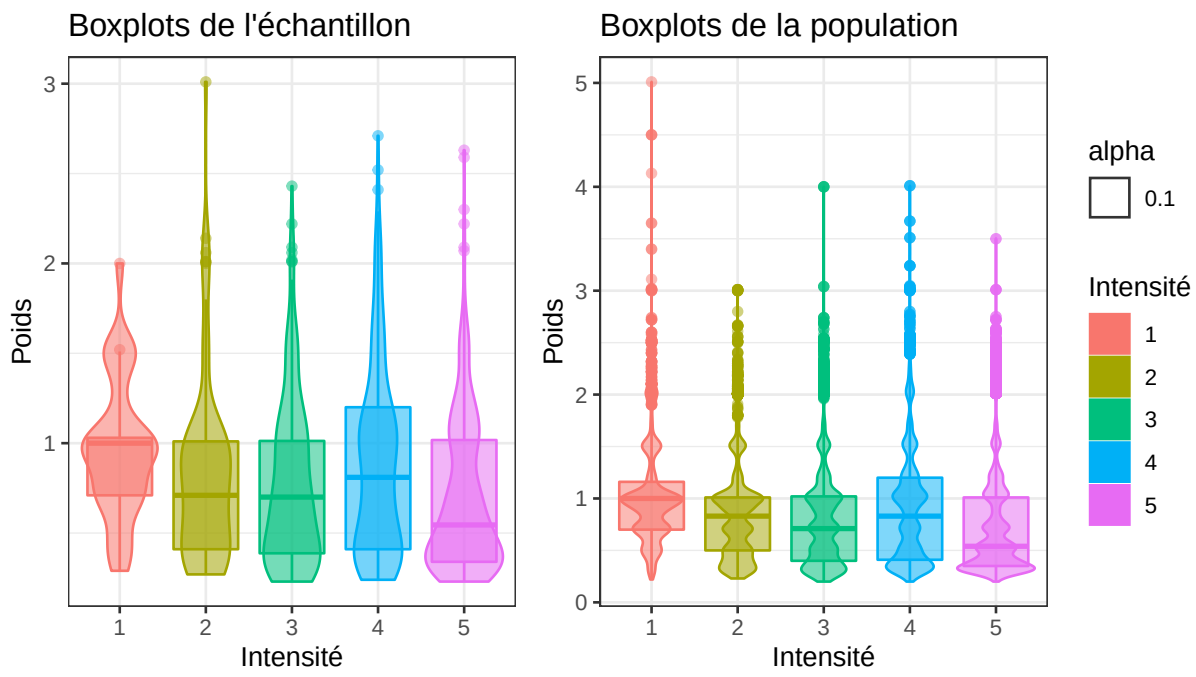


Nous afinons encore plus nos analyse car même si la régularité de notre graine ne semble pas influencer la relation existante entre le poids et le prix de la fleur liée à celle-ci nous pouvons clairement distinguer une classification ordinale croissante dans les pentes de nos régressions groupées par intensité de la couleur de la graine. Ceci va dans le sens de nos analyses précédentes car nous avions déjà conclu de l'indifférence existante dans l'impact des différents niveaux de régularité de la graine sur le prix de vente des fleurs mais au contraire nous avions démontré une différence significative au sein des niveaux **4** et **5** de l'intensité de la couleur de la graine. Au regard de la variable poids, nous pouvons prédire qu'une augmentation marginale du poids des graines de couleur d'intensité **5** provoque une augmentation proportionnelle plus élevée du prix que les autres niveaux.

```
## Prix ~ Poids * Intensité
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.551051	3.657646	5.3452545	0.0000001
Poids	23.109567	3.530570	6.5455625	0.0000000
Intensité2	-4.769129	3.968317	-1.2018015	0.2297143
Intensité3	-4.313940	3.786643	-1.1392517	0.2548608
Intensité4	-4.768573	3.774037	-1.2635205	0.2066856
Intensité5	-2.354556	3.727895	-0.6316046	0.5277842
Poids:Intensité2	10.047785	3.878116	2.5908935	0.0097068
Poids:Intensité3	12.409259	3.695522	3.3579181	0.0008138
Poids:Intensité4	13.954754	3.643360	3.8301883	0.0001357
Poids:Intensité5	23.553006	3.636664	6.4765422	0.0000000

Nous pouvons finalement quantifier les coefficients d'interactions entre les variables qui, graphiquement, nous semblaient montrer des divergences dans leurs niveaux en terme de pentes. On remarque très clairement qu'en augmentant l'intensité de la couleur de la graine, la variation du poids de celle ci entraine un effet plus important sur le prix de vente.



Ces boxplots nous démontre néanmoins que si nous n'avions pas eu le dataset représentant la totalité de la population étudiée au complet, nous n'aurions pas pu nous assurer de sélectionner les bonnes graines. En effet, nous ne pouvons pas observer graphiquement de corrélation entre le poids des graines et leur intensité de couleur. Nous pouvons même affirmer sur bases des analyses faites sur l'échantillon de 1049 graines que la médiane du poids des graines dans le niveau **5** de la variable `Intensité`est la plus faible des **5** niveaux de la variable, ce que nous confirme notre boxplots de la population totale. Nous ajouterons que notre valeur maximale pour le poids est d'environ 3 carats pour notre échantillon, alors que certaines graines obtiennent un poids plus élévé dans notre population. Nous allons essayer d'identifier ces outliers dans notre population.

Dans le but de maximiser le profit (différence entre le prix de vente de la fleur et les coûts d'achat et de plantation de la graine) qu'ils pourront espérer de cette activité d'horticulteurs, quelles graines faudrait-il choisir en priorité ?

Dans le développement de cette question, il nous faut tout d'abord rappeler les réponses que nous avons apportées précédemment. En effet, nous savons de source sûre que l'espérance de prix de vente augmente proportionnelement au poids. De plus l'espérance du prix conditionnellement à la couleur et à l'intensité de la couleur de la graine nous apporte des informations supplémentaires quant à la relation entre le poids et le prix. Lorsqu'on analyse les pentes respectives des droites de régression des différentes couleurs au sein même des niveaux de la variable `Intensité`, on s'aperçoit rapidement que là où le modèle linéaire du poids des fleurs bleues en fonction de leur prix nous donne une pente plus faible pour une intensité de **1** nous observons deux droites représentant cette relation pour les fleurs bleues et jaunes qui se confondent pour une intensité de couleur élevée de **5**. Nous nous souvenons également du coût de chaque graine de couleur bleue, jaune et rouge qui s'élevait respectivement à **6**, **7** et **8** euros respectivement, prix de plantation compris.

Afin de maximiser le profit, il faudra s'intéresser aux graines fournissant des fleurs au prix de vente le plus élevé car nous nous forcerons à maximiser la différence entre le prix et le coût pour constituer notre deuxième échantillon. C'est ainsi que nous devrons nous intéresser aux valeurs extrêmes de la variable `Poids`, car nous avons conclu qu'une relation linéaire existait entre cette variable et le prix de vente de nos fleurs.Pour se faire nous sélectionnerons en priorité les outliers de cette variable dans notre data set. de plus nous choisirons en priorité les graines de haute intensité de couleur, nous priviligierons les graines du niveau **5** de cette variable, car il est caractérisé par des pentes de droites de régresison nettement plus élevée que pour les autres nivceaux. Les droites de régression des différentes couleurs étant proches, mais pas systématiquement confondues au

sein d'un même niveau de la variable `Intensité`, nous pourrions stratifier sur cette variable afin de répartir nos coûts et nos profits, sans biaiser nos résultats.

Si les Zirobourdons décidaient d'acheter tout le stock de graines (sans contrainte de budget dans ce cas) et de vendre toutes les fleurs obtenues, quel serait le profit réalisé lors de cette opération ?

Pour répondre à cette question, nous devons estimer les prix de toutes nos combinaisons possibles grâce aux coefficients des différentes variables que nous obtenons en construisant le modèle linéaire pour chaque combinaison de ces même variable. Nous obtenons cette nouvelle colonne du prix de ventes pour notre dataset total très facilement, après avoir estimé le modèle linéaire regroupant les associations entre toutes nos variables. Nous utilisons alors simplement la focntion `predict.lm()` dans notre logiciel pour obtenir les prix de vente estimés par régression de toutes nos graines de nbtre population.

Il ne nous reste plus qu'à sommer sur cette colonne pour avoir le total du prix de revient de nos graines qui s'élève à **5702588** euros . Nous pouvons également connaître le prix de vente moyen estimé en divisant ce total par le nombre d'observations et nous obtenons **47** euros.

Afin de calculer la différence entre notre prix total estimé et nos coûts, il nous faut connaitre notre coût total pour notre population entière. Pour cela, rien de plus simple, nous créons une nouvelle variable `coût` qui fait correspondre ses niveaux `6`, `7`, `8` aux niveaux de la variable `Couleur` respectivement `bleue`, `jaune`, `rouge`. De cette façon, nous obtenons en plus du prix total estimé que précédemment le coût total qui lui n'est pas un estimateur, nous l'obtenons dans l'énoncé pour chaque couleur de graines, et celui-ci s'élève à **842199**. Nous calculons donc la différence et le profit qui s'élève donc à **4860389** pour notre population totale.

Redressement

Lorsque une partie de l'information explicative de la variable d'intérêt n'ait pas été disponible au moment de la sélection de l'échantillon, il faut l'utiliser au moment de l'estimation en redressant l'estimateur d'**Horvitz-Thompson**. On va modifier les pondérations associées à l'échantillon S en passant des poids de sondages d_k aux poids redressés w_k. On a pris connaissance d'une variable auxilière X, connue $\forall i \in U$ et **corrélée** avec Y. On l'utilisera pour améliorer la précision des estimateurs par différentes méthodes.

L'estimateur par **calage** consiste à supposer que l'on dispose d'un vecteur de variables auxiliaires X_k dont les totaux sont connus τ_k sur la population. On desirera trouver des nouveaux poids proches des poids d_k qui vérifient les équations $\sum_{k\in S} w_k x_k = \tau_k$. Dès lors, on cherche à réduire la variance de l'estimation; la variance sera nulle pour les variables auxiliaires et faible pour la variable d'intérêt bien expliquée par les variables auxiliaires. On utilisera le Lagrangien:

$$L = \sum_{k\in s} d_k G(\frac{w_k}{d_k}) - \lambda^T(\sum_{k\in s} w_k x_k - \tau_x)$$

où λ est un vecteur du multiplicateur de Lagrange, G est une fonction de distance entre le poids initial et final que l'on veut faible. Finalement on obtient que $w_k = d_k F[\lambda^T x_k]$ avec F la fonction inverse de G'. On résoud ensuite, avec le résultat obtenu, les équations non-linéaire de calage pour détermminer λ par l'algorithme de Newton jusqu'à obtention de convergence. Ceci nous fournit l'**estimateur par regression**:

$$\hat{\tau}_{y,reg} = \hat{\beta}_\pi^T \tau_x + \hat{\tau}_{e\pi} = \sum_{k\in s} w_k y_k$$

De plus, l'estimateur reg genéralisé est approcimativement sans **biais** et sa variance est approximativement donnée par les résidus de la régression de la variable y_k sur les variables auxiliaires x_k.

L'estimateur par le **ratio** suppose connu le total τ_x d'une seule variable auxiliaire tel que:

$$\hat{\tau}_{yR} = \hat{\tau}_{y\pi}\frac{\tau_x}{\hat{\tau}_{x\pi}} = \sum_{k\in S} w_k y_k \Rightarrow \hat{\tau}_{y,RC} = \tau_x \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h \bar{x}_h}$$

où $w_k = d_k \frac{\tau_x}{\hat{\tau}_{x\pi}}$. C'est un cas particulier de l'estimateur par régression généralisée, obtenu avec $X_k = x_k$ et $\sigma_k^2 = \sigma^2 x_k$. De plus, avec $E_k = y_k - Rx_k$:

$$V_p[\hat{\tau}_{y,RC}] \approx V_p[\hat{\tau}_{E\pi}] = \sum_{h=1}^{H} N_h^2 (1 - f_h)\frac{\sigma_{Eh}^2}{n_h}]$$

De même on peut définir l'estimateur par **ratio séparé** tel que:

$$\hat{\tau}_{y,RS} = \sum_{h=1}^{H} t_{xh}\frac{\hat{\tau}_{yh}}{\hat{\tau}_{xh}} = \sum_{h=1}^{H} \tau_{xh}\frac{\bar{y}_h}{\bar{x}_h}$$

On separera également la variance estimée en évaluant le terme erreur séparément pour chaque strates $e_k = y_k - R_h x_k$ pour $k \in U_h$ et $R_h = \frac{\tau_{yh}}{\tau_{xh}}$. La variance est donc réduite si les variables y_k et x_k sont approximativement proportionnelles dans les strates.

Un approche primoriale est de choisir une/des variables auxiliaires les plus explicatives pour construire notre calage. Ces variables devront également appartenir au plan de sondage. En principe, au plus on augmente les variables de calage, au plus les résidus seront faibles et donc la varaince de l'estimateur diminue. Néanmoins, les variables les plus explicatives nous permettent d'obtenir une forte diminution de cette variance.

On peut tester l'efficacité de l'estimateur par le ratio dans le cas d'un sondage aléatoire simple. Néanmoins toute l'information nécessaire à la prédiction de la variable réponse, ici le coût, est disponible dans notre dataset et il ne noous est donc pas utile de procéder à un redressement dans notre cas. En effet, la seule inconnue à ce stade est la variable prix ainsi que les différentes valeures prises par celle-ci pour toutes les combinaisons de nos variables.

Second echantillon

Stratification à allocation Optimale en terme de coûts

Dans un second temps, nous pourrions considérer, de par la nature de nos analyses et de notre population, un **sondage aléatoire stratifié**. Nous tenterons d'augmenter la précision de nos estimateurs grâce à des partitions en strates U_h de la population U. De ce fait, on obtiendra H échantillons S_h, dont l'union donnera notre échantillon total S. En suivant cette théorie, il sera dans notre intérêt de stratifier avec une variable très liée à la variable d'intérêt. Stratifier correspond souvent à un objectif de réduction des coûts d'enquète ou d'optimisation de sa gestion. Dès lors, il semble intéressant de stratifier suivant des **tailles optimales en terme de coûts**. En effet, en suivant un budget C_0, un nombre de tirage n non fixé a priori, et un vecteur $C = (C_1, ..., C_H)$ les coûts unitaires d'une observation dans une strate. On rappelle que le coût global de C fixé est donné par $C_0 + \sum_{h=1}^{H} C_h n_h$. On peut trouver les tailles optimales avec le Lagrangien:

$$n_h = argmin_{n_h}[\sum_{h=1}^{H}(\frac{N_h}{N})^2(1 - f_h)\frac{\sigma_{h;corr}^2}{n_h}]$$

$$s.c.$$

$$\sum_{h=1}^{H} n_h C_h \leq C_0$$

Les tailles optimales seront alors données par:

$$n_h = (\frac{C_0 \frac{N_H}{N} \sigma_{h;corr} / \sqrt{C_h}}{\sum_{l=1}^{H} \frac{N_l}{N} \sigma_{l;corr} \cdot \sqrt{C_h}})$$

Où f_h définit le taux de sondage lié à la strate h, N_h la taille de la strate h et N la taille de la population.

On retrouve dans cette expression l'allocation optimale de **Neyman** qui, par rapport à l'allocation proportionnelle, prendra en compte la dispersion à l'intérieur des strates. De cette manière, nous sur-représenterons quelque peu les strates plus hétérogènes et sous-représenterons quelque peu les strates ayant une faible variance interne. La tailles n_h de sous-échantillons selon l'allocation optimale de **Neyman** est donnée par:

$$n_h = (\frac{\frac{N_h}{N} \sigma_{h;corr}}{\sum_{l=1}^{H} \frac{N_l}{N} \sigma_{l;corr}}) n$$

L'allocation de **Neyman** donne, pour une stratification donnée et une variable d'intérêt donnée, l'allocation d'échantillon pour laquelle la variance du π-estimateur est minimisée. Cette allocation suppose la connaissance des dispersion dans les strates.

Mais, malgré notre optimisation en terme de coûts, notre problématique nous fait surtout référence à l'optimisation du profit pour notre producteur. De ce fait, nous ne pouvons pas seulement considérer la minimisation du côut de production des fleurs, qui est connu à priori. En effet, la maximisation du profit passe ici avant tout par la maximisation du prix de vente de nos fleurs. En effet, là où nos coûts varient très peu, dans un interval de [**6;8**], nous devons considérer une variation du prix de vente de nos fleurs très variable dans notre premier échantillon [**6.47; 142.62**]. ce qu'il nous faudra donc pour notre second échantillon c'est sélectionner des graines avec une éspérance de prix de vente élevé, quelle qu'en soit la couleur. Si nous disposons d'assez d'information pour prétendre que la couleur influence significativement le prix de vente nous aurions pu discriminer en fonction de cette variable. Malheureusement les données récoltées nous démontrent seulement une faible diffférence au sein de cette variable.

Comme nous l'avons évoqué plus tôt, nous privilégierons les graines de hautes intensités et au poids élevé. Néanmoins, nous déciderons de supprimer toutes les observations inférieures à un certain seuil de ces deux variables afin de garder une ceraine représentativité des niveaux de d'intensité car nous observons également une variabilité significative au sein même de ceux-ci. Nous choisirons donc d'éliminer les graines d'intensité **1** et de sélectionner des graines au poids supérieur à **2** pour constituer notre population de post-stratification.

Nous pourrions donc choisir, pour capturer un maximum de variance de notre échantillon de stratifier de façon optimale en terme de coûts en fonction de la variable `Couleur`. Pour ce faire nous aurons besoin des écarts-types estimés dans chaque strate correspondant aux différentes couleurs. Nous obtenons facilement ces estimations avec le logiciel Rstudio.

19.93184	20.2226	20.2903

Nous nous rendons compte que les écarts types sont extrêmement proches au sein de cette variable ce qui ne nous permet pas de stratifier de façon optimale en fonction de cette variable.

Outliers selection

Nous nous pencherons alors sur une méthode plus triviale, mais maximisant notre espérance de profit. En effet, nous nous rendons compte que nous devrons sélectionner les outliers de notre variable `Poids` de notre dataset afin de selectionner les graines les plus lourdes, à l'espérance de prix de vente la plus élevée. Nous prendrons soin de ne pas dépasser notre budget restant, qui s'élève à **7719** euros.

Il est temps de rappeler à ce stade que nous sommes dans une situation où l'on effectue deux tirages aléatoires sans remises, ce qui signifie donc que nous devons recoder une colonne binaire afin de ne pas sélectionner les mêmes lignes. Nous aurions pu choisir une alternative plus simple car nous possédons la variable `Obs`

mais celle ci ne s'actualise pas lorsqu'on supprime les lignes correspondantes aux graines sélectionnées lors du premier tirage.

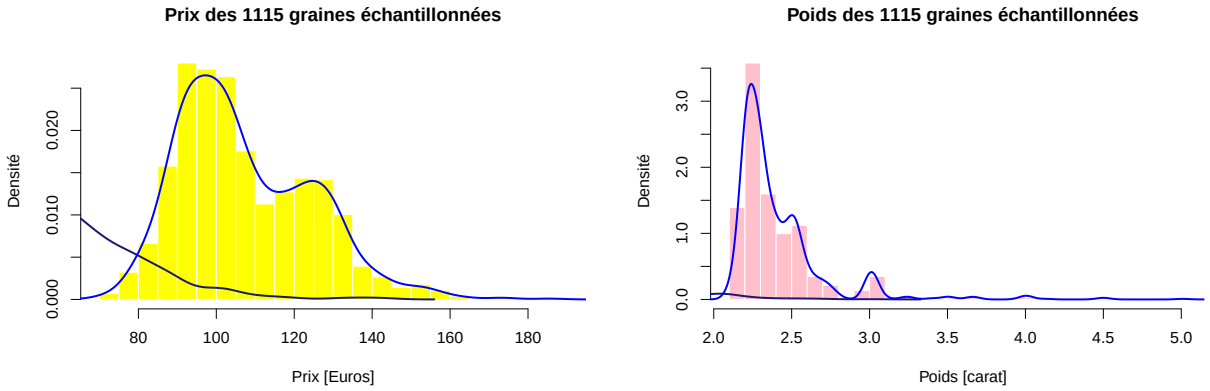
Après avoir recodé les observations de graines appartenant à notre premier échantillon, nous sélectionnons sur base des outliers du poids de notre population des graines pour un montant de **7333** euros et il nous reste donc un total de **386** euros à répartir. Nous constituons déjà la première partie de l'échantillon avec ces graines. La distribution du poids étant fortement dispersée, nous entrepenons donc de rééchantillonner en recodant les graines sélectionnées pour effectuer à nouveau une sélection sur les outliers.

Nous avons donc enlevé encore 1059 lignes dans notre population, correspondant au première graines sélectionnée pour notre second échantillon. Nous procédons denouveau à un repérage des outliers de la variable `Poids` dans cette nouvelle population. Nous nous intéresserons ici aux valeurs considérées comme extrêmes pour affiner et compléter notre sélection.

Nous isolons donc **63** graines supplémentaires qui amène notre deuxième échantillon à un total de **1051 + 63 = 1114** graines et nous répétons cette opération jusqu'à l'itération qui atteindra où dépassera notre budget maximum.

Une fois notre algorithme fini et afin de proposer notre échantillon final, nous entrepenons de fusionner les outlier-based échantillons sélectionnés à chaque itération précédemment. Nous sommons les coûts des graines de cet échantillon afin d'obtenir le total de celui-ci et de prendre connaissance du niveau atteint par cette métrique. De cette façon, nous pourrons considérer l'ajustement du ce total après la dernière itération de l'algorithme. Nous obtenons donc à la fin de notre outliers selection un total de **7727** pour notre variable `Coût` de notre second échantillon, ce qui s'avère être **8** euros de plus que le budget maximum fixé pour celui-ci. Nous entreprenons donc de retirer une graine rouge, nous remettrons dans la population non-sélectionnée celle au poids le plus faible et aux caractéristiques les plus basses. Nous choisirons une graine de Poids valant **2.19** carats et à faible intensité (**2**). Dès lors, nous rentrons totalement dans nos coûts!

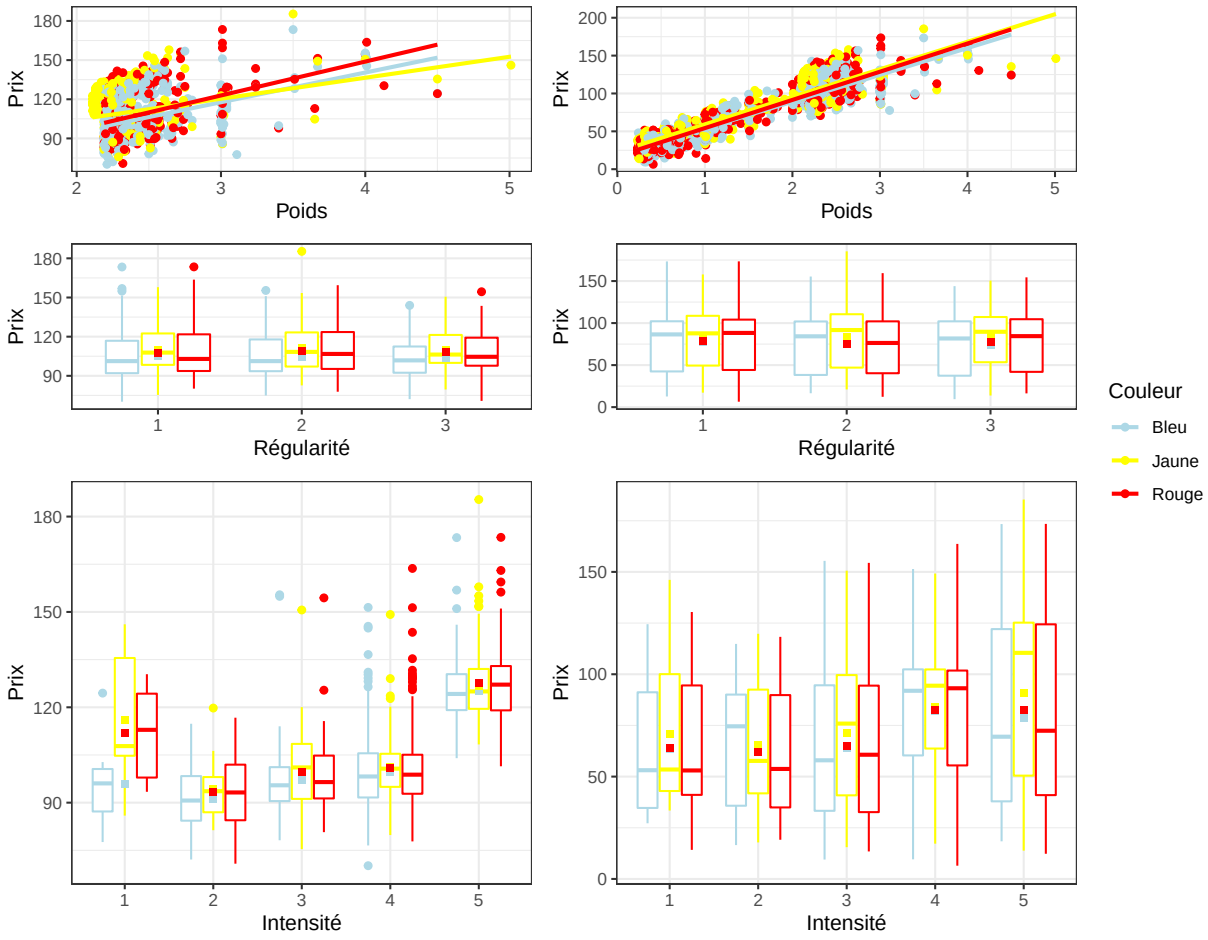
Suite à la récolte de notre deuxième échantillon, nous pouvons tirer des analyses complémentaires mais également vérifier nos conclusions précédentes, tant en terme de rentabilité qui constitue l'éssence même de ce travail d'échantillonnage qu'en terme de conclusions statistiques et d'estimations sur notre population. Premièrement, nous entreprenons à nouveau de représenter nos distributions d'échantillon pour valider notre sélection et prendre connaissance de la répartition des observations sélectionnées.



On remarque très rapidement en observant ces deux barplots que notre approche nous apportent les résultats attendus. En effet, nous pouvons observer les valeurs extrêmes sélectionnées et leurs corrélations avec la variable `Prix` en observant l'allure de leurs densités. Nous décidons de laisser apparaître en bleu foncé les courbes de densités déjà tracées précédemment afin de clarifier ces conclusions graphiques.

Nous prenons dès lors connaissance de l'efficacité de notre méthode en calculant quelques statistiques de base telles que la moyenne du prix de notre échantillon qui s'élève à **107.37** euros, ou encore les valeurs minimale et maximale de celui ci respectivement **70.17** et **185.36**. Dès lors nous avons augmenté notre profit effectué qui s'élève maintenant à **111999.5** euros soit une augmentation de **267.54** %.

Enfin, nous nous permettons de représenter les relations existantes entre nos variables pour notre second échantillon récolté (**gauche**) et pour nos deux échantillons combinés (**droite**). Nous pouvons dès lors conclure que nos analyses et nos réponses aux questions restent valides après regroupement de nos deux échantillons. La combinaison de ceux-ci nous offre une vue d'ensemble parfaite sur les relations existantes entre les variables caractérisant les observations de notre population. De plus la relation entre la variable `poids` et la variable `Prix` reste linéaire après régression par general Linear model, et ce même dans les zones extrèmes de notre échantillon combiné.



Après réunification des échantillons, il est maintenant graphiquement évident qu'une différence significative est observée dans la réaction de la variable `Prix` par rapport à l'intensité, et ce pour les graines d'intensité **5**. De plus nous observons une variabilité conséquente entre les boxplots des **3** couleurs au sein même des graines d'intensité **1**. Ceci peut justifier nos ambiguités dans notre analyse par comparaison effectuée à la lumière du test de Dunn.

Conclusion

Nous résumerons ici les conclusions prises lors des différentes étapes de nos échantillonnages. Après avoir choisi un premier échantillon sur base d'une méthode d'échantilonnage bien connue **la stratification à allocation proportionnelle**. Ce sont pour des raisons de représentativité de la population et de contrainte de budget que nous avons décider de surstratifier en fonction de 3 de nos variables afin de représenter au mieux nos variations inter et intra classes. Par la suite, après avoir balayé bon nombre de méthodes intérressantes, nous privilégierons une technique plus triviale mais construite de manière itérative, que l'on appellera **outliers**

selection. De cette manière, nous maximisons d'une part notre profit, objectif même de ce travail, et nous l'analysons en parallèle à notre premier échantillon pour obtenir une représentation plus complète de notre population.

Bibliographie

- Bernard Roullet (2008), L'INFLUENCE DE LA COULEUR EN MARKETING : VERS UNE NEUROPSYCHOLOGIE DU CONSOMMATEUR
- Guillaume Chauvet (2015), Méthodes de sondage Echantillonnage et Redressement, École Nationale de la Statistique et de l'Analyse de l'Information
- Kestemont Marie-Paule, LSTAT2200: Cours d'échantillonnage et sondage
- Laurent Rouvière, Introduction aux sondages, Université de Rennes-Service Universitaire d'Enseignement à Distance
- Site Survey Monkey: calcul de tailles d'échantillon optimal
- Tillé, Y. (2001). Théorie des sondages : échantillonnage et estimation en populations finies (Cours et exercices avec solutions), Dunod, Paris.

Appendices

Appendice 1 : Code utilisé pour ce rapport

```
knitr::opts_chunk$set(echo = FALSE)
library(kableExtra)
library(sampling)
library(stratbr)
library(ggplot2)
library(dplyr)
library(survey)
library(rstatix)
library(tidyverse)
library(ggpubr)
library(vcd)
library(tinytex)
library(pander)
options(tinytex.verbose = TRUE)
data <- read.csv("./graines.csv", stringsAsFactors = TRUE)
data$Couleur <- factor(data$Couleur)
data <- data[-1]
colnames(data) <- c("Obs","Poids","Couleur","Intensité", "Régularité")
knitr::include_graphics("./logo.jpg")
knitr::include_graphics("./roses.jpg")
summary(data[-1]) |>  kbl() |> kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
hist(data$Poids, nclass = 100, col = "pink", border = "white",
     main = paste("Poids des", nrow(data), "graines"), xlab = "Poids [carat]", ylab = "Densité",
     proba = TRUE, xlim = c(0,3), ylim = c(0, 5/2))
lines(density(data$Poids, na.rm = TRUE), lwd = 2, col = "midnightblue" )
data$Intensité <- as.factor(data$Intensité)
data$Régularité <- as.factor(data$Régularité)
tab <- xtabs(~ Régularité + Intensité + Couleur, data = data)
vcd::doubledecker(Intensité ~ Couleur + Régularité, data = tab,
                  main = "Intensité ind. Régularité | Couleur")
vcd::doubledecker(Régularité ~ Intensité + Couleur, data = tab,
                  main = "Régularité ind. Couleur | Intensité")
vcd::doubledecker(Couleur ~ Régularité + Intensité, data = tab,
                  main = "Couleur ind. Intensité | Régularité")

knitr::include_graphics("./sondage en grappes.png")
knitr::include_graphics("./stratification.png")
knitr::include_graphics("./Decomposition.png")
set.seed(2022)
n <- 1050
N <- dim(data)[1]
datastrat <- data.frame(xtabs(~Couleur + Intensité + Régularité, data = data))
datastrat <- datastrat[order(datastrat$Couleur, datastrat$Intensité, datastrat$Régularité),]
nh <- data.frame(round(n / N * datastrat$Freq))
nh$Select<- nh$round.n.N...datastrat.Freq.
nh <- nh[-1]
data <- data[order(data$Couleur, data$Intensité, data$Régularité),]
6 * sum(nh[1:15,]) + 7 * sum(nh[16:30,]) + 8 * sum(nh[31:45,])
nH <- as.vector(nh$Select)
Sample <- sampling::strata(data, stratanames = c("Couleur", "Intensité", "Régularité"),
```

```
                          size = nH, method = "srswor")
Echant <- getdata(data, Sample)
Echantillon_1 <- Echant[-(1:8)]
write.csv(Echantillon_1,"Echantillon_1.csv")
Echant1 <- read.csv("./E1_Aymeric_Warnauts.csv", stringsAsFactors = TRUE)
Echant1 <- Echant1[-1]
colnames(Echant1) <- c("Obs","Prix","Poids","Couleur","Intensité", "Régularité")
Echant1$Intensité <- factor(Echant1$Intensité)
Echant1$Régularité <- factor(Echant1$Régularité)
cout <- function(x){
  if(x == "Bleu") y <- 6
if(x == "Jaune") y <- 7
if(x == "Rouge") y <- 8
return(y)
}

data$Coût <- sapply(data$Couleur, cout)
Echant1$Coût <- sapply(Echant1$Couleur, cout)
Echant1 %>% shapiro_test(Prix) %>% kbl() %>% kable_minimal() %>%
  kable_styling(latex_options = c("hold_position"))
Echant1 %>% group_by(Couleur) %>% shapiro_test(Prix)  %>% kbl()  %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
hist(Echant1$Prix, nclass = 40, col = "yellow", border = "white",
     main = paste("Prix des", nrow(Echant), "graines échantillonnées"),
     xlab = "Prix [Euros]", ylab = "Densité", proba = TRUE)
lines(density(Echant1$Prix, na.rm = TRUE), lwd = 2, col = "midnightblue" )

hist(Echant$Poids, nclass = 40, col = "pink", border = "white",
     main = paste("Poids des", nrow(Echant), "graines échantillonnées"),
     xlab = "Poids [carat]", ylab = "Densité",
     proba = TRUE, xlim = c(0,3), ylim = c(0, 5/2))
lines(density(Echant$Poids, na.rm = TRUE), lwd = 2, col = "midnightblue" )
library(patchwork)
theme_set(theme_bw())
p1 <- ggplot(Echant1, aes(x=Poids, y=Prix, color = Couleur)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)+
  scale_colour_manual(values = c("lightblue", "yellow", "Red"))
p3 <- ggplot(Echant1, aes(x=Intensité, y=Prix, color = Couleur))+
  geom_boxplot() + scale_colour_manual(values = c("lightblue", "yellow", "Red"))+
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
p2 <- ggplot(Echant1, aes(x=Régularité, y=Prix, color = Couleur))+
  geom_boxplot() + scale_colour_manual(values = c("lightblue", "yellow", "Red"))+
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
(p1 + p2) / p3 + plot_layout(guides = "collect")
outliers1 <- data.frame(Echant1 %>% group_by(Couleur) %>% identify_outliers(Prix))
outliers1[which(outliers1$is.extreme == TRUE), ] %>% kbl() %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
model1 <- lm(Prix~Couleur, data = Echant1)
ggqqplot(residuals(model1))
ggqqplot(Echant1, "Prix", facet.by = "Couleur")
plot(model1,1)
Echant1 %>% kruskal_test(Prix~Couleur) %>% kbl() %>% kable_minimal() %>% kable_styling(latex_options = c
Echant1 %>% kruskal_effsize(Prix~Couleur) %>% kbl() %>% kable_minimal() %>% kable_styling(latex_options
Echant1 %>% dunn_test(Prix~Couleur, p.adjust.method = "bonferroni") %>%
```

```r
  kbl() %>% kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
outliers2 <- data.frame(Echant1 %>% group_by(Intensité)
                        %>% identify_outliers(Prix))
outliers2[which(outliers2$is.extreme == TRUE), ] %>% kbl()%>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
Echant1 %>% group_by(Intensité) %>% shapiro_test(Prix)  %>% kbl()  %>% kable_minimal() %>% kable_stylin
model2 <- lm(Prix~Intensité, data = Echant1)
ggqqplot(residuals(model2))
ggqqplot(Echant1, "Prix", facet.by = "Intensité")
plot(model2,1)
Echant1 %>% kruskal_test(Prix~Intensité) %>% kbl() %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
Echant1 %>% kruskal_effsize(Prix~Intensité) %>% kbl() %>% kable_minimal() %>%
  kable_styling(latex_options = c("hold_position"))
Echant1 %>% dunn_test(Prix~Intensité, p.adjust.method = "bonferroni") %>%
  kbl() %>% kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
outliers3 <- data.frame(Echant1 %>% group_by(Régularité) %>%
                          identify_outliers(Prix))
outliers3[which(outliers3$is.extreme == TRUE), ] %>% kbl()  %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
model3 <- lm(Prix~Régularité, data = Echant1)
ggqqplot(residuals(model3))
ggqqplot(Echant1, "Prix", facet.by = "Régularité")
plot(model3,1)
Echant1 %>% group_by(Régularité) %>% shapiro_test(Prix)  %>% kbl() %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
Echant1 %>% kruskal_test(Prix~Régularité) %>% kbl() %>% kable_minimal() %>%
  kable_styling(latex_options = c("hold_position"))
Echant1 %>% kruskal_effsize(Prix~Régularité) %>% kbl() %>% kable_minimal() %>%
  kable_styling(latex_options = c("hold_position"))
summary(lm(Echant1$Prix ~ Echant1$Poids))$coefficients %>% kbl() %>%
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
ggplot(Echant1, aes(x=Poids, y=Prix, color = Régularité)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)
ggplot(Echant1, aes(x=Poids, y=Prix, color = Intensité, shape = Couleur, linetype = Couleur)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)
formula(lm(Prix~Poids*Intensité, data = Echant1))
summary(lm(Prix~Poids*Intensité, data = Echant1))$coefficients |> kbl() |>
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
p1 <- ggplot(Echant1, aes(x = Intensité, y = Poids, color = Intensité, fill = Intensité)) +
  geom_boxplot(alpha = 0.5) + labs(title = "Boxplots de l'échantillon") + geom_violin(mapping = aes(alp
p2 <- ggplot(data, aes(x = Intensité, y = Poids, color = Intensité, fill = Intensité)) +
  geom_boxplot(alpha = 0.5) + labs(title = "Boxplots de la population") + geom_violin(mapping = aes(alp
  theme(legend.position = "NONE")
p1 + p2 + plot_layout(guides = "collect")
sum(Echant1$Prix)
sup <- as.vector(Echant1$Obs)
data <- data[order(data$Obs),]
data1 <- data[-sup,]
outliersf <- data.frame(data1 %>% group_by(Intensité) %>% identify_outliers(Poids))
outliersf[which(outliersf$is.extreme == TRUE), ]
reg <- lm(Prix~Poids*Intensité*Couleur*Régularité, data = Echant1)
coefficients(reg) |> kbl(col.names = "Coef", vline = "|") |>
  kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
```

```r
prix.data <- predict.lm(reg, data)
sum(prix.data)
datastrat2 <- data1[which(as.numeric(data1$Intensité) > 1 & data1$Poids > 2), ]
data1[which(data1$Intensité == 5  & data1$Poids >= 2), ]
Nh <- as.numeric(table(datastrat2$Couleur))
Sh <-  sapply(unique(Echant1$Couleur), function(x) {
sd(Echant1$Prix[Echant1$Couleur == x])
})
t(Sh) |> kbl() |> kable_minimal() %>% kable_styling(latex_options = c("hold_position"))
data$select <- 0
data <- data[order(data$Obs),]
for (i in Echant1$Obs) {
  data[i,]$select <- 1
}
data$select <- as.factor(data$select)
Echant20 <- data.frame(data[which(data$select !=1),] |> identify_outliers(Poids))
Echant20
sum(Echant20[which(Echant20$Intensité !=1 & Echant20$Poids > 2.18),]$Coût)
7719 - 7333
Echant2a <- Echant20[which(Echant20$Intensité !=1 & Echant20$Poids > 2.18),]
for (i in Echant2a$Obs) {
  data[i,]$select <- 1
}
data$select <- as.factor(data$select)
Echant20 <- data.frame(data[which(data$select !=1),] |> identify_outliers(Poids))
sum(Echant20[which(Echant20$is.extreme),]$Coût)
Echant2b <- Echant20[which(Echant20$is.extreme),]
for (i in Echant2b$Obs) {
  data[i,]$select <- 1
}
data$select <- as.factor(data$select)
Echant20 <- data.frame(data[which(data$select !=1),] |> identify_outliers(Poids))
sum(Echant20[which(Echant20$is.outlier & Echant20$Intensité == 5 &
                     Echant20$Poids > 2.1 & Echant20$Couleur == "Jaune"),]$Coût)
Echant2c <- Echant20[which(Echant20$is.outlier &
                             Echant20$Intensité == 5 &
                             Echant20$Poids > 2.1 & Echant20$Couleur == "Jaune"),]
nrow(data[which(data$select == 1),])
for (i in Echant2c$Obs) {
  data[i,]$select <- 1
}
data$select <- as.factor(data$select)
Echant2 <- rbind(Echant2a,Echant2b,Echant2c)
Echant2[-802,]
Echant2[which(Echant2$Couleur == "Rouge" &
                Echant2$Poids < 2.2 &Echant2$Intensité == 2), ]$select <- 1
Echant2 <- Echant2[which(Echant2$select !=1),]
sum(Echant2$Coût)
Echantillon_2 <- as.vector(Echant2$Obs)
write.csv(Echantillon_2,"./Echantillon_2.csv", row.names = FALSE)
Echantfinal <- read.csv("./E2_Aymeric_Warnauts.csv", stringsAsFactors = TRUE)
Echantfinal <- Echantfinal[-1]
colnames(Echantfinal) <- c("Obs","Prix","Poids","Couleur","Intensité", "Régularité")
Echantfinal$Intensité <- factor(Echantfinal$Intensité)
```

```r
Echantfinal$Régularité <- factor(Echantfinal$Régularité)
Echantfinal$Coût <- sapply(Echantfinal$Couleur, cout)
hist(Echantfinal$Prix, nclass = 40, col = "yellow", border = "white",
     main = paste("Prix des", nrow(Echantfinal), "graines échantillonnées"),
     xlab = "Prix [Euros]", ylab = "Densité", proba = TRUE)
lines(density(Echant1$Prix, na.rm = TRUE), lwd = 2, col = "midnightblue" )
lines(density(Echantfinal$Prix, na.rm = TRUE), lwd = 2, col = "blue" )

hist(Echantfinal$Poids, nclass = 40, col = "pink", border = "white",
     main = paste("Poids des", nrow(Echantfinal), "graines échantillonnées"),
     xlab = "Poids [carat]", ylab = "Densité", proba = TRUE)
lines(density(Echant1$Poids, na.rm = TRUE), lwd = 2, col = "midnightblue" )
lines(density(Echantfinal$Poids, na.rm = TRUE), lwd = 2, col = "blue" )
p1 <- ggplot(Echantfinal, aes(x=Poids, y=Prix, color = Couleur)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  scale_colour_manual(values = c("lightblue", "yellow", "Red"))
p3 <- ggplot(Echantfinal, aes(x=Intensité, y=Prix, color = Couleur)) +
  geom_boxplot() + scale_colour_manual(values = c("lightblue", "yellow", "Red")) +
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
p2 <- ggplot(Echantfinal, aes(x=Régularité, y=Prix, color = Couleur)) + geom_boxplot() +
  scale_colour_manual(values = c("lightblue", "yellow", "Red")) +
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
Echantanalyse <- rbind.data.frame(Echant1, Echantfinal)
p4 <- ggplot(Echantanalyse, aes(x=Poids, y=Prix, color = Couleur)) + geom_point() +
  geom_smooth(method = "glm", se = FALSE) + scale_colour_manual(values = c("lightblue", "yellow", "Red")
p5 <- ggplot(Echantanalyse, aes(x=Intensité, y=Prix, color = Couleur)) + geom_boxplot() +
  scale_colour_manual(values = c("lightblue", "yellow", "Red")) +
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
p6 <- ggplot(Echantanalyse, aes(x=Régularité, y=Prix, color = Couleur)) + geom_boxplot() +
  scale_colour_manual(values = c("lightblue", "yellow", "Red")) +
  stat_summary(fun = mean, geom = "point", shape = 15) + theme(legend.position = "none")
(p1 + p4 + p2 + p6) / (p3 + p5) + plot_layout(guides = "collect")
cbind(datastrat, nh) %>%  kbl() %>% kable_minimal(full_width = FALSE, position = "center") %>%
  row_spec(1:15, color = "blue") %>%  row_spec(16:30, color = "orange") %>%
  row_spec(31:45, color = "red") %>% kable_styling(latex_options = c("hold_position"))
```

Appendice 2: Premier échantillon sélectionné

	Couleur	Intensité	Régularité	Freq	Select
1	Bleu	1	1	637	6
16	Bleu	1	2	439	4
31	Bleu	1	3	340	3
4	Bleu	2	1	1905	16
19	Bleu	2	2	1316	11
34	Bleu	2	3	1063	9
7	Bleu	3	1	4906	42
22	Bleu	3	2	3441	30
37	Bleu	3	3	2479	21
10	Bleu	4	1	5573	48
25	Bleu	4	2	3774	33
40	Bleu	4	3	2874	25
13	Bleu	5	1	8528	74
28	Bleu	5	2	6114	53
43	Bleu	5	3	4504	39
2	Jaune	1	1	480	4
17	Jaune	1	2	306	3
32	Jaune	1	3	210	2
5	Jaune	2	1	1357	12
20	Jaune	2	2	954	8
35	Jaune	2	3	712	6
8	Jaune	3	1	3276	28
23	Jaune	3	2	2302	20
38	Jaune	3	3	1713	15
11	Jaune	4	1	3825	33
26	Jaune	4	2	2643	23
41	Jaune	4	3	1957	17
14	Jaune	5	1	5917	51
29	Jaune	5	2	4070	35
44	Jaune	5	3	3061	26
3	Rouge	1	1	548	5
18	Rouge	1	2	370	3
33	Rouge	1	3	291	3
6	Rouge	2	1	1667	14
21	Rouge	2	2	1157	10
36	Rouge	2	3	849	7
9	Rouge	3	1	4219	37
24	Rouge	3	2	2825	24
39	Rouge	3	3	2147	19
12	Rouge	4	1	4724	41
27	Rouge	4	2	3218	28
42	Rouge	4	3	2469	21
15	Rouge	5	1	7230	63
30	Rouge	5	2	5119	44
45	Rouge	5	3	3837	33