Rapport du Projet

Introduction

La réussite de toute entreprise nécessite une étude de faisabilité et aussi de rentablilité. C'est dans ce cadre que nous avons été contacté par les Ziroboudons du pays de Rêverose, pour la réalisation d'une étude de faibilité de leur projet d'horticulture. Notre mission ici est d'étudier les différentes caractéristiques des graines ainsi d'en dégager une estimation du prix de vente des fleurs issues desdites graines. Pour cela, nous disposons d'une base de données de toutes les graines et leurs caractéristiques mais pas d'information sur les prix de vente des fleurs. Ainsi, d'abord, nous étudierons la base de données afin de dégager des tendances de celle-ci. Ensuite nous posséderons à un premeir échantillonnage pour évaluer les caractéristiques du prix des fleurs. Et enfin, en nous basant sur les informations de ce premier échantillon, nous sélectionnerons un second échantillon qui nous permettra de faire de réelles estimation sur le prix des fleurs et aussi répondre au préoccupations diverses des Ziroboudons.

Présentation et description des données

Présentation générale des données

La base de sondage à notre disposition est un stock de 121 346 graines, avec leurs caratéristiques : Id, Poids, Couleur, Intensité, Régularité.

- Id: le numéro d'identification de la graine, de 1 à 121 346
- Poids: le poids de la graine, exprimé en carats (un carat vaut 0,2 grammes)
- Couleur: la couleur de la graine (bleu, rouge ou jaune)
- Intensité: l'intensité de la couleur de la graine. L'intensité est quantifiée par un nombre entier entre 1 et 5. Une note de 1 correspond à l'intensité minimale alors qu'une note de 5 correspond à l'intensité maximale (une couleur de faible intensité est une couleur mixte mélangée - une couleur de haute intensité est une couleur pure sans mélange)
- Régularité: la régularité de la forme de la graine. La régularité est quantifiée par un nombre entier entre 1 et 3. Les graines les moins régulières obtiennent une note de 1. La note de 3 est attribuée aux graines les plus régulières

Table 1: Extrait de la base de sondage

Id	Poids	Couleur	Intensité	Régularité
1	0.51	Rouge	4	2
2	0.34	Bleu	5	2
3	0.96	Rouge	3	1
4	0.31	Bleu	4	1
5	1.14	Rouge	3	2
6	0.39	Bleu	5	2
7	0.90	Rouge	3	3
8	0.70	Rouge	3	1
9	1.20	Jaune	2	2
10	0.46	Bleu	4	2

Description des données

Du tableau des statistiques descriptives des données, nous pouvons lire que le poids des graines varient de 0.2 à 5.01 pour une moyenne de 0.796 et un poids médian de 0.7.

Table 2: Statistique descriptive de la base

Poids	Couleur	Intensité	Régularité
Min. :0.2000	Bleu :47893	1: 3621	1:54792
1st Qu.:0.4000	Jaune:32783	2:10980	2:38048
Median :0.7000	Rouge:40670	3:27308	3:28506
Mean :0.7959		4:31057	
3rd Qu.:1.0400		5:48380	
Max. :5.0100			

Une analyse de la répartition des poids en fonction de la couleur des graines, de l'intensité des couleurs et de la regularité des gtaines pourra nous permettre de bien mesurer la répartition des graines de notre stock.

D'abord, voyons la répartition des graines en fonction des couleurs, de intensité des couleurs et aussi de la regularité des graines.

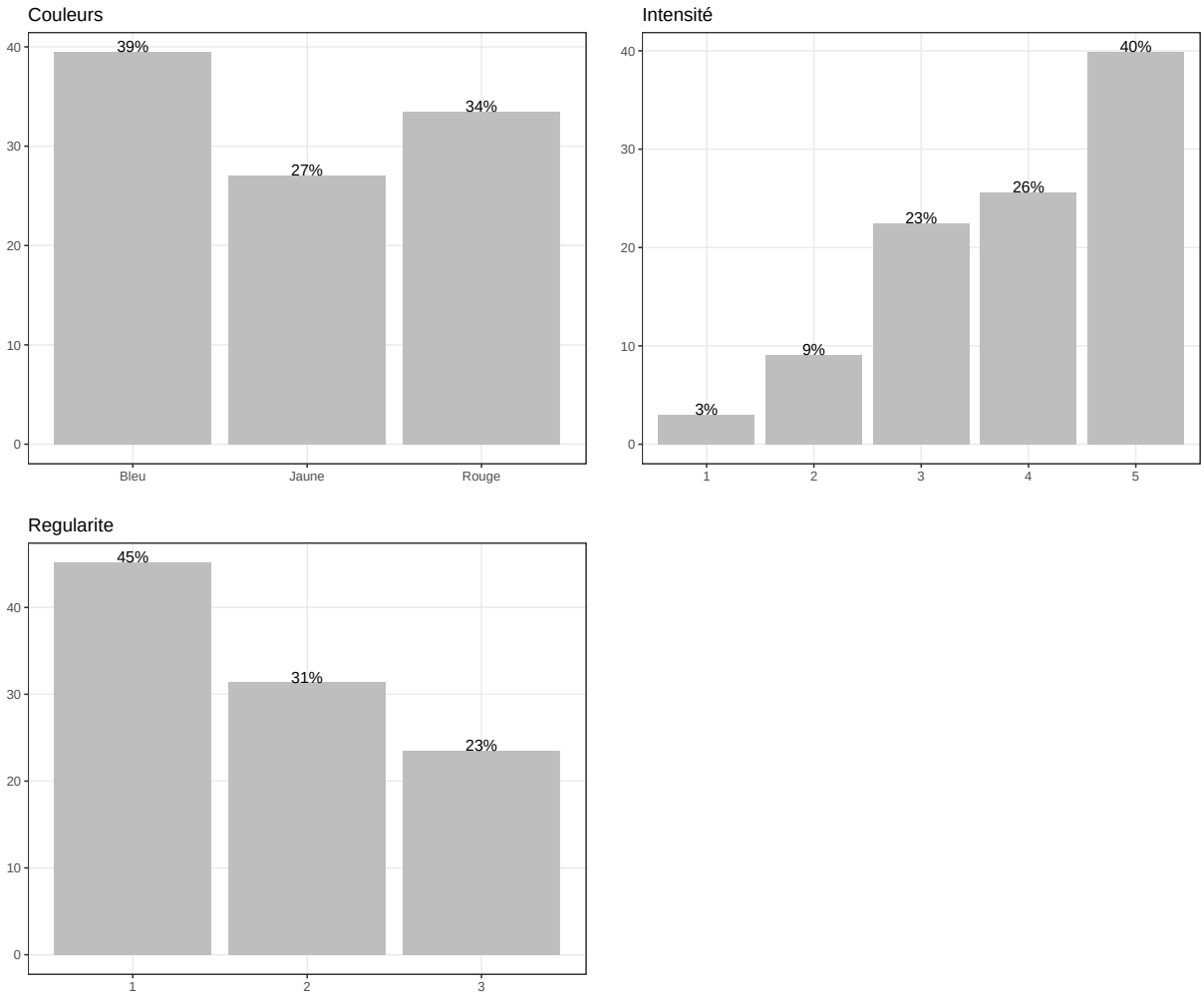


Figure 1: Répartition des graines en fonction de leurs caractéristiques

L'analyse de la répartition des graines montre que :

- en fonction des couleurs, 39% des graines sont blues, contre 27% et 34% pour les jaunes et les rouges respectivement.
- les graines d'une couleur pure sans mélange (intensité 5) représentent 40% de stock tandis que celles de couleur mixte mélangée (intensité minimale : 1) sont seulement 3% du stock.
- quant à la regularité des graines, 45% sont de forme moins régulière pour 23% qui sont de forme de forme plus régulières. Les autres (31%) sont de forme intermédiaire.

Au vue de cette repartition de notre stock, il sera bien de savoir le lien qui pourrait exister entre la couleur des graines, leurs intensités, la forme de ces graines et aussi la répartition du poids des graines au sein de ces caractéristiques. Cette étude est représentée dans les graphiques suivantes :

Des graphiques de la figure 2, il convient de retenir :

- la distribution de la couleur, de l'intensité des couleurs et de la forme de la graine est identique au sein des différents groupes. Ceci s'observe sur les graphiques en barre par l'égalité des proportions au sein des groupes (des graphiques en barre qui sont plats). En conséquence, il y a une *independance mutuelle* entre les caractéristiques : couleur de la graine, intensité de la couleur de la graine et la forme de la graine.
- quant à la distribution du poids au sein des caractéristiques *couleur de la graine* et *regularité de la graine*, en observant les courbe des densité du poids entre les différentes couleurs et entre les forme de la graine, on a tendance à dire que ces courbes sont identiques. Ce fait est confirmé par les caractéristiques de boites à moustache. En effet, lesdites boites à moustache présentent les mêmes quartiles. Ce qui nous permet de conclure que la distribution du poids des graines au sein des caractéristiques *couleur* et *regularité de la graine* est homogène.
- Pour ce qui est de l'intensité des couleurs, le graphique de la densité des poids en fonction de cette dernière montre que la distribution du poids n'est pas homogène au sein des modalités de cette caractéristique. Une analyse de la variance pourra nous conforter dans cette position.

Table 3: Analyse de la variance du poids au sein de la caractéristiques Intensité

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Intensité	4	955.7272	238.9318097	1109.117	0
Residuals	121341	26139.9236	0.2154253		

Le tableau 3 ci-dessus, confirme bien que la distribution du poids n'est pas homogène au sein de la caractéristique *Intensité de la couleur des graines* (p-value $< 5\%$)

Methodologie d'échantillonnage et estimation des paramètres

Afin de répondre aux préoccupations des Ziroboudons, nous avons besoin des informations sur le prix de vente des fleurs produites par les graines. Pour cela, nous devons constituer un premier echantillon. Ce premier echantillon nous permettra de récolter des informations principalement sur le prix de vente des fleurs. Alors quelles seront les caractéristiques de ce premier échantillon ?

Pour la sélection du premier echantillon, partant des informations dont nous disposons : *contraintes budgétaires*, *catégorisation des graines* en trois groupes selon la couleur des graines, la *différence entre les prix d'achat des graines* et aussi le *coût de plantation d'une graines*, nous stratifierons la base de sondage en fonction de la couleur des graines. Cette stratification nous permettra de recolter les informations nécessaires et de parfaire le second echantillon qui constituera la base même de notre analyse pour répondre aux questions des Ziroboudons. Aussi, nous décidons d'allouer une montant maximal de *3000 euros* à cette étude préliminaire.

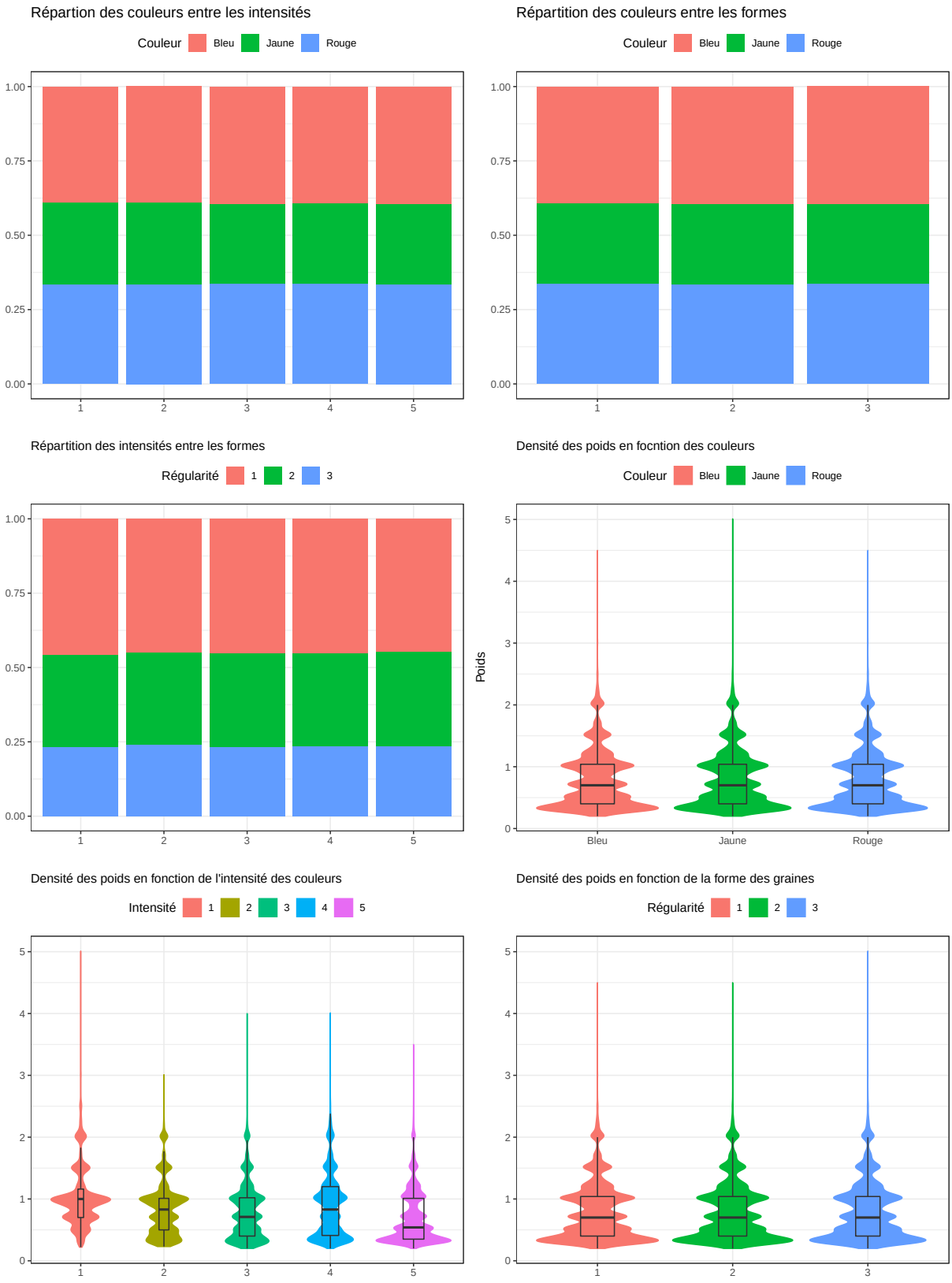


Figure 2: Analyse croisée de couleur, intensité, regularité et poids

Sélection et étude du premier echantillon

Sélection de l'échantillon_1

Notre base de sondage est stratifiée en trois strates : U_j, U_b, U_r pour respectivement la strate des graines Jaunes, des Bleues et des Rouges. Au sein de chacune des strates, nous allons prélever de façon aléatoire et sans remise échantillon de taille $n = 400$ soit donc un taux de sondage de 0.003. Nos $n = 400$ graines seront prélevées proportionnellement au sein de chaque strate. Ainsi, nous prélevons un nombre n_h avec $h \in \{j, b, r\}$ de graine au sein des strates avec un taux de sondage $f_h = 0.0033$. Nous avons donc :

$$n_j = 158 \qquad n_b = 108 \qquad n_r = 134$$

Pour connaitre le prix de vente des fleurs produites par ces graines, nous devons d'abord:

- acheter ces graines (les prix étant de 6 euros pour une graine jaune, contre respectivement 5 et 7 euros pour une graine bleue et une graine rouge) et ensuite
- les planter (le prix de plantation des graines 1 euros chacune des graines).

Ainsi, ce premier échantillon que nous constituons nous coûte : 2850 euros

Auprès de ce premier échantillon de 400 graines dont 158 jaunes, 108 bleues et 134 rouges, nous allons nous intéresser à la distribution du prix des fleues produites par lesdites graines.

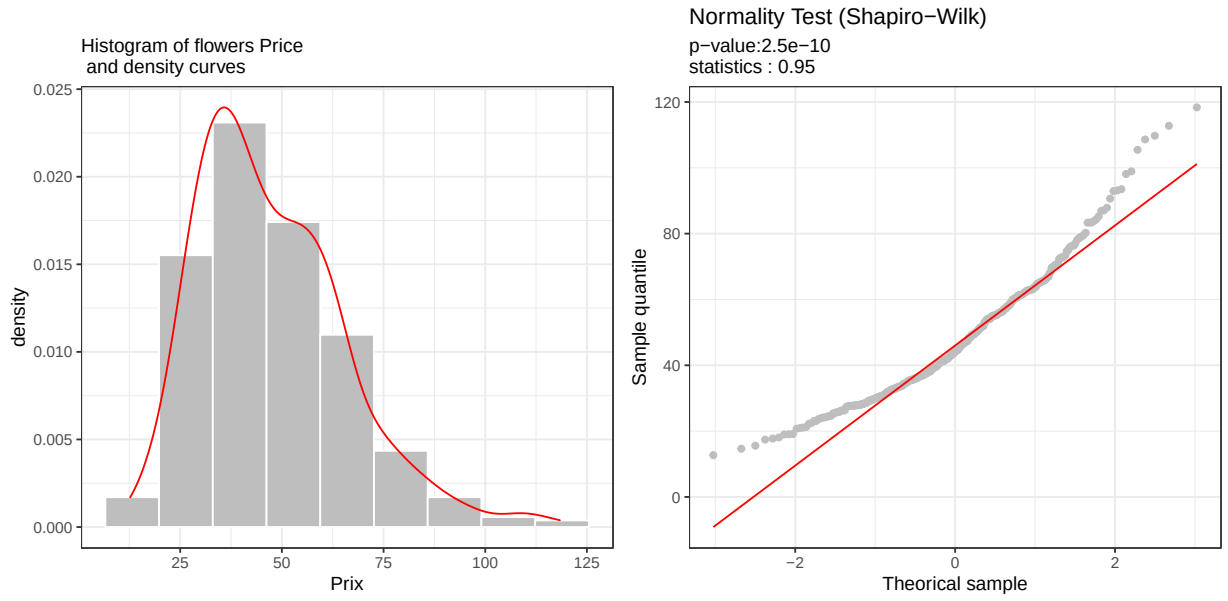
Etude de l'échantillon_1

- **Repartition du Prix**

D'abord analysons le tableau d'analyse descriptive du prix des fleurs et l'histogramme de la distribution de ce prix au sein des 400 graines sélectionnées.

Table 4: Statistiques descriptives du prix des fleurs dans Echantillon_1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.69	33.7	44.31	47.43	58.31	118.35



Au travers de l'histogramme du prix des fleurs, nous remarquons que les prix s'étant de 12 *euros* à 120 *euros* environ avec une classe modale [35;45] et un prix moyen de 47.43 *euros*. Une idée serait de voir si cette distribution pourrait correspondre à une distribution normale. Visuellement, à partir de la courbe de densité, on voit que non, néanmons, confirmons cette assertion à travers le qqplot. Ce dernier montre bien que les quantiles du Prix ne s'alignent pas à celles d'une normale. Ceci est aussi confirmé par les résultats du test de normalité basé sur le test de `Shapiro-Wilk`.

- **Distribution du Prix des fleurs en fonction des caractéristiques des graines**

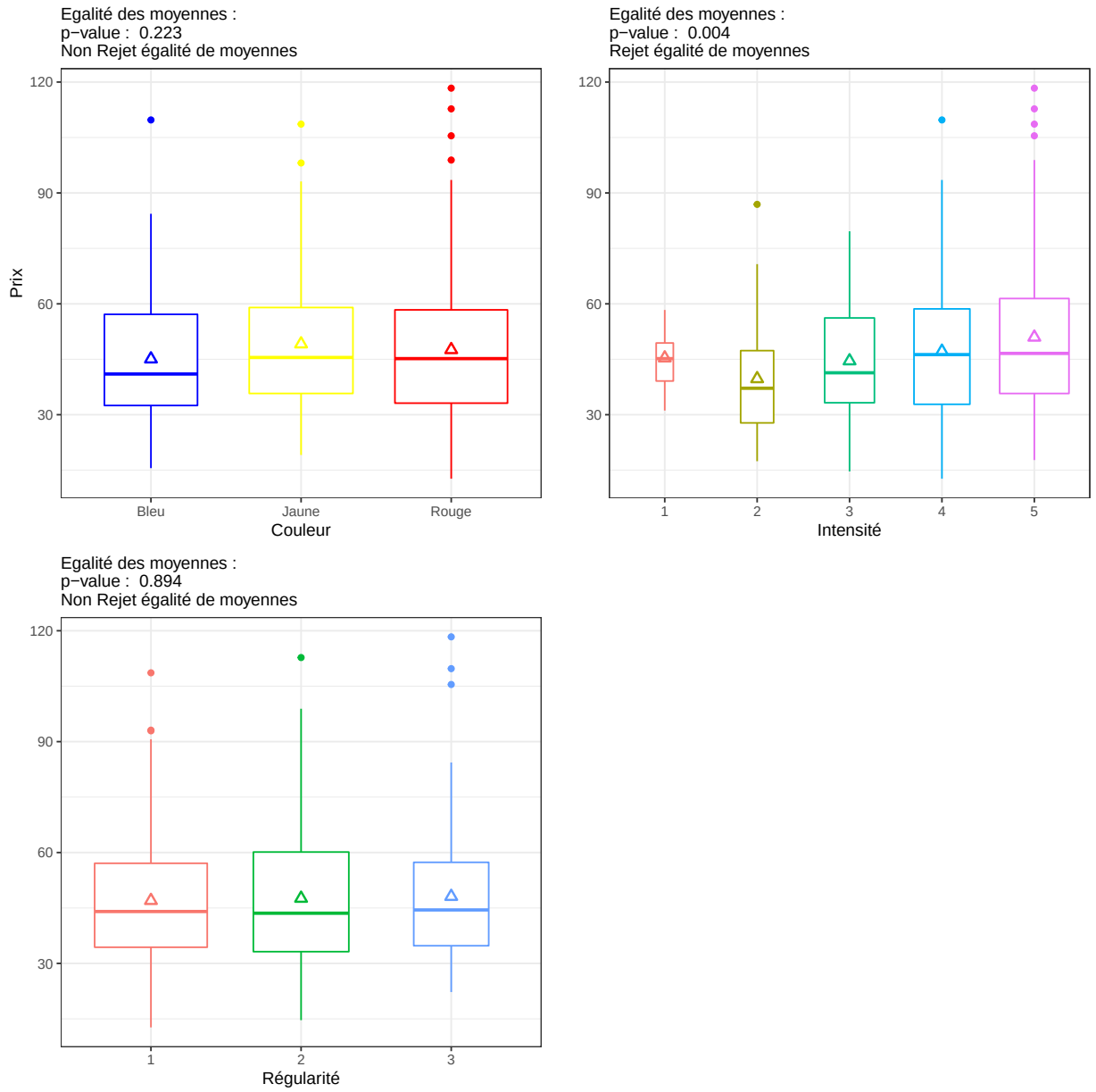


Figure 3: Distribution du Prix en fonction des caractéristiques des graines

L'étude du graphique de la distribution du Prix en fonction des caractéristiques des graines montre qu'au sein de notre Echantillon_1, le prix moyen des fleurs est d'environ `45 euros`. Ce prix reste invariant que la graine soit d'une couleur ou d'une autre ou d'une régularité à une autre. Cependant, le prix de la fleur restent dépendant de l'intensité de la graines.

Observons à présent le lien entre le prix de la fleur et le poids de la graine (Figure 4).

De ce graphique, on retient que le prix de la fleur et le poids de la graine liée à ladite fleurs sont liés par une relation linéaire avec un coefficient de corrélation linéaire évalué à 0.858. Quant à la distribution du prix et dupoids en fonction de l'intensité de la couleur des graines, on remarque que la liaison linéaire est toujours présentes.

En somme, avec ce premier échantillon, on retient que la variable d'intérêt qu'est le prix de la fleur est liée fortement au poids et qu'elle varie en fonction de l'intensité de la couleur de la graine ayant produit cette fleur. En effet, plus l'intensité de la couleur de la graine est élévée, plus la fleur produites par cette graine est vendu à un prix élévé.

Ainsi, pour la constitution de notre second échantillon, on privilégera une stratification optimale en fonction de l'intensité de la couleur de la graine. En effet, ce que nous recherchons, c'est d'avoir des strates assez homogènes du point de vue de l'intensité des graines de ladite strate et donc avec une variance du prix de la fleur assez faible.

Pour cette stratification, nous avons besoin de la variance au sein de chacune des strates de notre variable d'intérêt: le prix de la fleur. Ne connaissant pas cette variance au sein de la population des graines, nous nous contenterons de son *estimation*[1] à base de notre premier echantillon : *Enchantillon_1*. Ceci est résumé dans le tableau suivant.

Table 5: Ecart-type estimé du prix au sein des categories de l'intensité des graines

Intensité	1	2	3	4	5
Prix	8.82273	15.43024	15.82787	17.45957	20.04197

Rappelons aussi que pour notre premier échantillon, nous avons dépensé 2850 *euros* et donc nous disposons de maximum 12150 *euros* pour la réalisation de ce second echantillon; sachant en effet que les graines coûtent (achat et plantation) : $c_j = 7$ *euros*; $c_b = 6$ *euros*; $c_r = 8$ *euros*.

[1] Notons aussi que puisque le prix de la fleur et le poids des graines présentent une forte correlation, on pourrait aussi utiliser les variances de cette variable en lieu et place de celle du prix de la fleur.

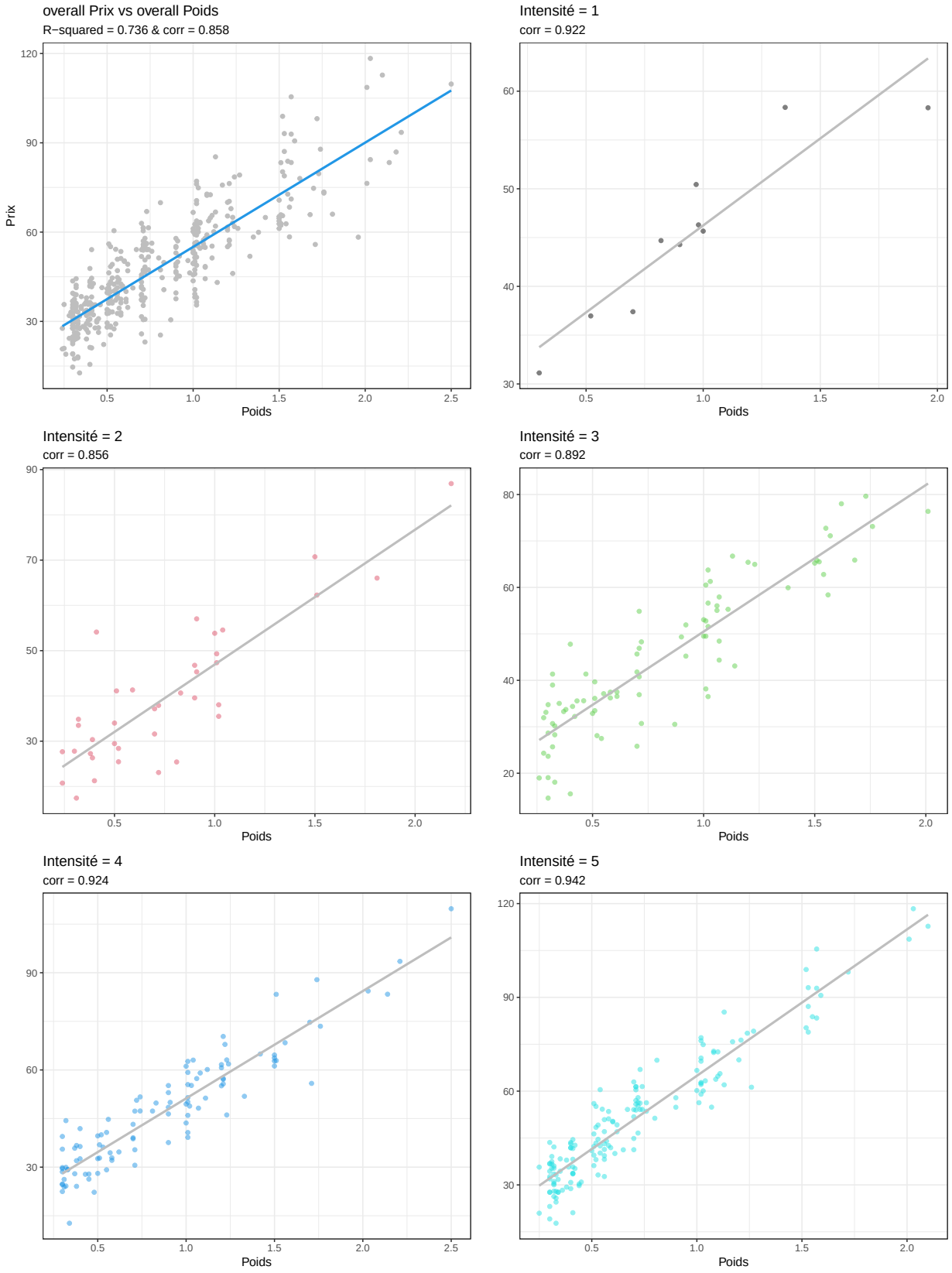


Figure 4: Prix et Poids de la graine

Sélection et étude du second echantillon

Methode de sélection du second échantillon

De l'ensemble des analyses précédentes, nous avons vu que le prix de la fleur varie en fonction de l'intensité de la graine qui donne ladite fleur. Ainsi, nous avons décidé de concevoir notre second échantillon en nous appuyant sur cette information. Et donc, nous avons stratifié à nouveau les graines restantes de notre base de données en fonction de l'intensité des graines. Ensuite, étant donné que l'objectif de l'estimation des prix des fleurs, nous avons procédé à une stratification optimale pour sélection les graines devant constituer notre second échantillon. Le but étant bien sûr d'avoir une variance minimale pour l'estimation à venir du prix moyen des fleur. Aussi, ne disposant pas la variance des prix au niveau population, nous avons utilisé les estimations obtenues à partir de notre premier échantillon.

Ainsi donc, notre premier échantillon nous a permis de récolter des informations pour mieux constituer notre second échantillon. C'est sur ce dernier que nous allons nous baser pour répondre aux préoccupations des Ziroboudons.

Une première choses à revérifier est de savoir si le budget fournit par les Ziroboudons est respecté :

Table 6: Coût du second échantillon

Couleur	Bleue	Jaune	Rouge	Total
Taille	679	437	627	1743
Coût (en euros)	4074	3059	5016	12149

Et bien evidemment, avec notre stratégie de sélection du second échantillon de taille $n = 1743$, nous respectons notre budget (2850 euros pour constituer l'echantillon_1 et 12150 euros pour l'échantillon_2) avec 1 euro restant.

Etude du second échantillon

Notre second échantillon est un échantillon de taille 1743 graines. Etant donné que nous avons opté pour une stratification selon l'intensité de la couleur des graines, voyons aussi la taille de nos strates et les taux de sondages y associés.

Soit donc U_h la strate associées à la catégories des graines d'intensité h avec $h \in \{1, 2, 3, 4, 5\}$

Table 7: Taux de sondage pour Echantillon_2

Intensité	1	2	3	4	5
N_h	3621	10980	27308	31057	48380
n_h	48	161	381	466	687
f_h	0.0133	0.0147	0.014	0.015	0.0142

Analyse description

Commençons alors par une analyse description de notre second échantillon.

Table 8: Satistique description de Echantillon_2

Prix	Poids	Couleur	Intensité	Régularité
Min. : 6.36	Min. :0.2300	Bleu :679	1: 48	1:790
1st Qu.: 31.45	1st Qu.:0.3700	Jaune:437	2:161	2:546
Median : 41.60	Median :0.7000	Rouge:627	3:381	3:407
Mean : 45.76	Mean :0.7796		4:466	
3rd Qu.: 56.37	3rd Qu.:1.0300		5:687	
Max. :129.13	Max. :3.0000			

En effet, le prix moyen des fleurs de notre second échantillon est 45.76 *euros* avec écart-type de 19.82 *euros*. Rappelons que ce prix moyen était de 47.43 *euros* dans l'échantillon_1 avec un écart-type estimé de 18.2 *euros*. Aussi un rapide t.test permet d'avoir une idée de la significativité de la différence entre les deux échantillons. En effet, la p-value du test de la comparaison des deux moyennes est **0.105** ; cette dernière étant supérieure à au seuil 5%, on ne rejettera pas l'hypothèse d'égalité des moyennes de nos deux échantillons.

Interessons nous maintenant à la distribution du prix des fleurs au sein des caractéristiques des graines.

Distribution du prix des fleurs en fonction des caractéristiques des graines

A la lecture du graphique, on peut remarquer que la distribution globale du prix des fleurs au sein des caractéristiques des graines semblent s'apparenter. Mais est-ce vraiment le cas? Pour répondre à cette question, allons plus en profondeur de nos données. D'abord, jettons un coup d'oeil aux boxplots avec les positions relatives des moyennes de chaque catégories des caractéristiques des graines.

On retient bien que :

- **Couleur des graines**

Le prix moyen au sein des différentes couleurs de la graine sont bien différent. En effet, les graines de couleur jaune apparaissent comme celles produisant des fleurs dont le prix de vente moyen supérieur est aux prix de vente moyen des fleus produites par les deux autres couleurs. Pendant ce même temps, le test de comparaison du prix moyen des fleurs produites par les graines *bleues* et *rouges* laisse conclure que ces prix ne sont pas significativement différents l'un de l'autre : $p.value_{\ (H_o:\ \mu_b=\mu_r)} = 0.54$ (>5%).

- **Intensité de la couleur des graines**

En analysant l'intensité de la couleur des graines en relation avec le prix de vente des fleurs produites par ces graines, la comparaison des prix moyens des 5 catégories résulte en un rejet de l'hypothèse d'égalité des 5 moyennes. Et donc, il y a bien un lien significatif entre l'intensité de la couleur d'une graine et le prix auquel la fleur produites par cette graine est vendu. Ainsi, en analysant plus en détails, les graines appartenant à la catégories d'intensité **4** et **5** présentent les prix moyens les plus élévés.

- **Régularité de la forme des graines**

Pour ce qui concerne la régularité, on retiendra que le prix des fleurs produites restent relativement identitquement distribué au sein des catégories. En d'autres termes, la *régularité de la forme des graines* n'a pas un impact significatif sur le prix de vente des fleurs produites par ces graines. Ceci étant la conclusion logique du résultat de l'analyse des variances réalisée sur la moyenne des 3 catégories de la régularité ($p.value = 0.697$).

- **Poids des graines**

Au niveau de ce dernier, le nuage de point traduit un lien linéaire entre le prix d'une fleur et le poids de la graine ayant donnée cette fleur. En effet, le R^2_{adj} de la régression linéaire donne *0.781* et le coefficient de corrélation linéaire entre ces deux variables est *0.884* (>0).

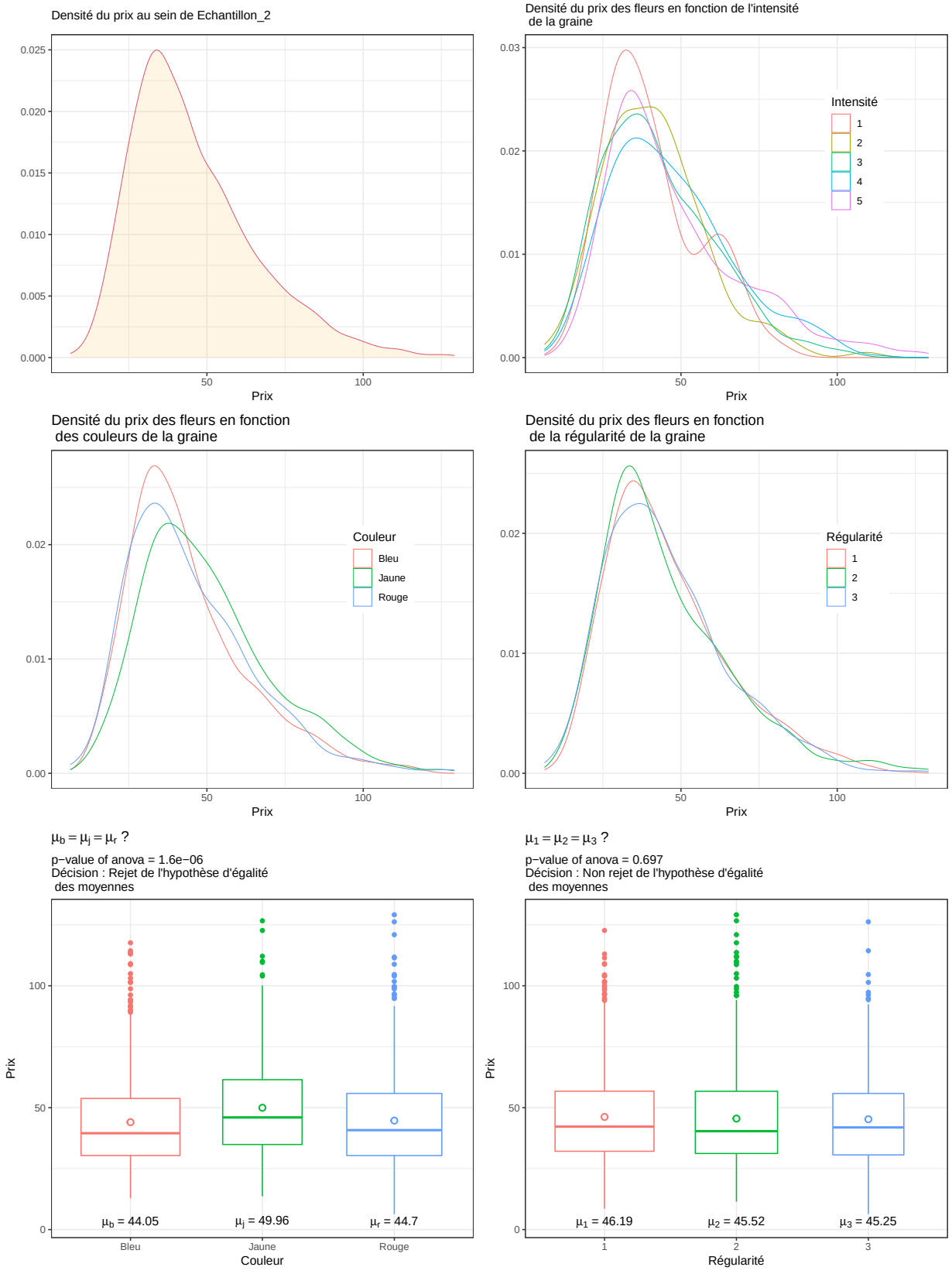


Figure 5: Densité du prix au sein des différents groupes de graines

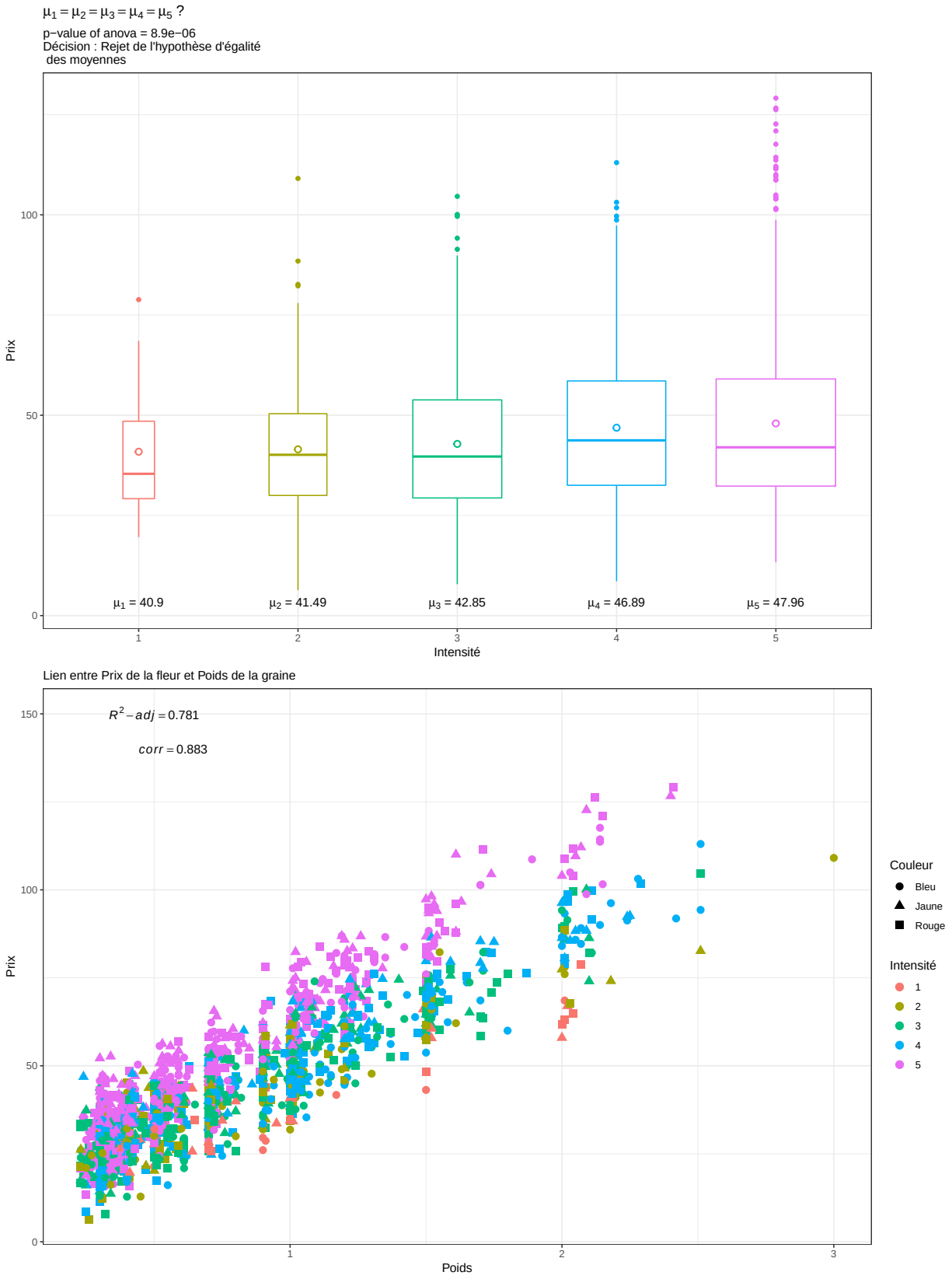


Figure 6: Densité du prix au sein des différents groupes de graines (suite)

Le coefficient de corrélation étant positif, on peut donc conclure que plus le poids de la graine est élévé, plus cher seraient vendus les fleurs qu'elle donnera.

Aussi en analysant plus en détails le scatterplot, on remarque que les graines d'intensité **5** et **4** donnent des fleurs dont les prix de vente sont élévés. Dans ce groupe, les graines d'intensité **1** sont les graines donnant des fleurs dont les prix de vente sont les bas.

Estimation des paramètres de la population

- **Estimation des paramètres**

Dans cette sous-section, nous nous focaliserons sur l'estimation des paramètres du prix des fleurs produites par nos graines sélectionnées. Cette estimation se basera principalement sur les données recueillies sur l'échantillon_2.

En effet, dans les sections précédentes, nous avons étudié nos deux échantillons et avons émis des conclusions sur lesdits échantilons tout en comparant les différents paramètres obtenus sur chacun des échantillons.

Rappelons que notre Echantillon_2 est issu d'une ***stratification optimale*** de notre base de donnée de $N = 121346$ graines en fonction de ***l'intensité des couleurs*** desdites graines. Et aussi, notre variable d'interêt Y est le prix de la fleur. Cette sur cette dernière que sera focalisée nos estimations.

Dans le tableau ci-dessous, sont résumés les paramètres de nos cinq strates.

Table 9: Paramètres de Echantillon_2 et estimation

Intensité	1	2	3	4	5	***Estimation***
N_h	3621	10980	27308	31057	48380	$N = 121346$
n_h	48	161	381	466	687	$n = 1743$
f_h	0.0133	0.0147	0.014	0.015	0.0142	$f = 0.0144$
$\overline{y}_h$	40.897	41.493	42.849	46.889	47.958	$\widehat{\mu} = 45.74$
$\widehat{Var}(\overline{y}_h)$	0.308	0.102	0.047	0.042	0.032	$\widehat{Var}(\widehat{\mu}) = 0.216$
$s_{h;corr}$	14.807	16.401	17.73	19.483	21.725	

Ainsi donc avec notre stratégie de sondage, nous avons une estimation du prix moyen de vente des fleurs qui s'éleverait à ***45.74*** euros avec une variance estimée de ***0.216***.

Ainsi donc avec un niveau de confiance de 95%, on est sûr que toute estimation du prix moyen des fleurs à partir d'un échantillon sera dans l'intervalle : **[44.83, 46.65]**

Retenons que ce prix est bien l'estimation au niveau global. Au niveau des diverses intensités des couleurs des graines, bien entendu, nous avons des disparités du prix des fleurs et donc on est tenté de construire un intervalle de confiance à 95% pour chacune de ces catégories. Cette information est résumée dans le tableau suivant:

Table 10: Estimation des paramètres des strates et IC à 95%

Intensité	1	2	3	4	5
$\widehat{\mu}_h$	40.897	41.493	42.849	46.889	47.958
$\widehat{Var}(\widehat{\mu}_h)$	4.507	1.646	0.814	0.802	0.677
$IC_{\mu_h}(95\%)$	**[36.74, 45.06]**	**[38.98, 44.01]**	**[41.08, 44.62]**	**[45.13, 48.64]**	**[46.35, 49.57]**

- **Redressement des estimations**

Ayant à disposition, la répartition de l'intensité des couleurs des graines au sein de la population, il est donc avantageux d'effectuer un calage de nos estimations pour un meilleures estimation. Bien évidemment

une stratification optimale nous permet d'avoir de bon estimateur, mais un meilleur estiamteur calé sur les données de la population serait idéale.

- **Maximisation de profit et bénéfice attendu**

A la lumière des analyses précédentes, il apparait évident que pour maximiser le profit pouvant être généré de cette activité, les ziroboudons ont intérêt à choisir prioritairement les graines de couleur ***Jaune*** et dont les l'intensité de la couleur est dans les catégories ***4*** et ***5***.

Aussi, dans le cas où les ziroboudons achetèrent tout le stock de graines sans contrainte de budget, le profit auquel ils peuvent prétendre, ceteris paribus, est évalué à 5.550366×10^6 euros. A ce montant viendra bien évidemment en déduction le coût de l'achat de graines.

Conclusion

Au vue des analyses précédentes, nous avons conclut que le prix de vente des fleurs restent fortement lié à l'intensité de la couleur des graines. Ceci, nous a conduit à réaliser une stratification optimale en fonction de cette caractéristique pour construire notre second échantillon. Certes, une stratification nous donnerait des estimations avec les variances les plus faible, mais ces estimations seront-elles sans biais ?? En effet, l'étude de l'effet de sondage de notre staratégie d'échantillonnage laisse paraitre d'une PESR aurait donné de meilleure variance.