Université Cergy Paris

Diplôme universitaire : Data Analyst

Examen
Statistiques

11 décembre 2022

Haury Fabien

Examen Statistiques

Haury Fabien

11 décembre 2022

Table des matières

1 Régression **3**

1.1 Quelle est la différence entre une régression linéaire simple et une régression linéaire multiple ? 3

1.2 Que signifie la multicolinéarité ? A quoi sert le facteur d'inflation de la variance (VIF en anglais) dans le cadre de la lutte contre la multicolinéarité ? 3

1.3 Que signifie l'homoscédasticité des résidus ? Pourquoi est-ce important ? 4

1.4 Quelles sont les hypothèses sous-jacentes à l'application du modèle linéaire (régression, etc.) ? 4

1.5 A quoi sert un qqplot au juste ? Pourquoi est-ce souvent préféré aux tests comme Shapiro ou Kolmogorov-Smirnov ? 4

2 Tests non paramétriques **5**

2.1 Quelle est la différence entre un test apparié et un test non-apparié ? Illustrer avec des exemples concrets 5

2.2 Dans quel contexte utilise-t-on des tests non paramétriques ? 6

2.3 Citer quatre exemple de tests non paramétriques, et les situations d'application correspondantes 7

2.4 D'où vient l'expression "test de rangs" utilisée pour désigner les tests nonparamétriques ? . . . 7

3 ANOVA **8**

3.1 Expliquer comment est calculée une somme de carrés pour une variable donnée 8

3.2 Comment calcule-t-on une statistique F pour une variable donnée ? 9

3.3 Comment passe-t-on du F à la p-value ? Illustrer obligatoirement avec un graphe 9

3.4 Que représente l'interaction entre deux facteurs dans une ANOVA ? Illustrer avec un cas concret 11

4 Régression logistique **13**

4.1 Dans quel contexte utilise-t-on des régressions de type "loi de Poisson" ? 13

4.2 Pourquoi au juste utilise-t-on le terme "Poisson" pour désigner ce type de régression logistique ? 13

4.3 Pourquoi au juste utilise-t-on le terme "binomiale" pour désigner un autre type de régression logistique ? 13

4.4 A quoi correspond un odd-ratio et en quoi les OR diffèrent-ils du risque relatif ? 14

4.5 Dans quelle condition risques relatifs et OR sont-ils approximativement les mêmes ? 14

Table des figures **15**

Liste des tableaux **16**

Chapitre 1

Régression

1.1 Quelle est la différence entre une régression linéaire simple et une régression linéaire multiple ?

Une régression linéaire se compose de deux parties :

- Une variable dépendante noté y. Elle correspond à la variable que nous cherchons à expliquer.
- Une ou plusieurs variables indépendantes noté $x_1, ..., x_n$. Elles correspondent aux variables servant à expliquer la variable dépendante.

La différence entre une régression simple et une régression multiple réside dans le nombre de variables indépendantes employées. Une régression simple est de la forme : $y = \beta_0 + \beta_1 x$, avec β_0 l'ordonnée à l'origine de la droite et β_1 son coefficient directeur. Celle-ci explique la relation entre la variable dépendante et la variable indépendante sélectionnée. Une régression multiple est de la forme : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$, elle permet d'expliquer le lien entre la variable dépendante et les variables indépendantes choisies. Il est possible de faire interagir les variables indépendantes entre elles.

1.2 Que signifie la multicolinéarité ? A quoi sert le facteur d'inflation de la variance (VIF en anglais) dans le cadre de la lutte contre la multicolinéarité ?

La multicolinéarité est l'occurrence de fortes inter-corrélations entre deux ou plusieurs variables indépendantes dans un modèle de régression multiple. La multicolinéarité peut conduire à des résultats faussés ou trompeurs lorsque l'on tente de déterminer dans quelle mesure chaque variable indépendante peut être utilisée le plus efficacement pour prédire ou comprendre la variable dépendante dans un modèle statistique. Le facteur d'inflation de la variance (FIV) ou variance inflation factor (VIF) est une mesure de l'importance de la multicolinéarité dans l'analyse de régression. La détection de la multicolinéarité est importante car si la multicolinéarité ne réduit pas le pouvoir explicatif du modèle, elle réduit la signification statistique des variables indépendantes. Un VIF élevé sur une variable indépendante indique une relation hautement colinéaire avec les autres variables qui doit être prise en compte ou ajustée dans la structure du modèle et la sélection des variables indépendantes.

1.3 Que signifie l'homoscédasticité des résidus ? Pourquoi est-ce important ?

L'homoscédasticité signifie que la variance des erreurs de la régression est identique. L'homoscédasticité est aussi appelée homogénéité de variance. Cela suggère un niveau de cohérence et facilite la modélisation et l'utilisation des données par régression ; cependant, l'absence d'homoscédasticité peut suggérer que le modèle de régression doit inclure des variables indépendantes supplémentaires pour expliquer la performance de la variable dépendante.
Elle permet d'assurer la validité des tests statistiques utilisés pour évaluer les résultats du modèle de régression. En effet, la plupart de ces tests sont basés sur l'hypothèse que les résidus ont une variance constante, et une violation de cette hypothèse peut entraîner des résultats erronés. Elle permet d'assurer la qualité des prédictions du modèle de régression. En effet, une variabilité constante des résidus signifie que les erreurs de prédiction du modèle sont uniformément distribuées autour de la droite de régression, ce qui peut garantir une meilleure précision des prédictions.

1.4 Quelles sont les hypothèses sous-jacentes à l'application du modèle linéaire (régression, etc.) ?

- Hypothèse de corrélation : il existe une relation de corrélation entre la variable dépendante et la/les variables indépendantes.
- Hypothèse de distribution gausienne des résidus : les résidus suivent une loi de distribution gausienne normale
- Hypothèse d'homoscedasticité des résidus : les résidus ont la même variance quel que soit le groupe considéré, ou quelle que soit la valeur de la variable explicative considérée.

1.5 A quoi sert un qqplot au juste ? Pourquoi est-ce souvent préféré aux tests comme Shapiro ou Kolmogorov-Smirnov ?

Un qqplot est utilisé pour comparer les formes des distributions, fournissant une vue graphique de la façon dont les propriétés telles que l'emplacement (location), l'échelle (scale) et l'asymétrie (skewness) sont similaires ou différentes dans les deux distributions.
Un qqplot est souvent préféré car il permet d'avoir un aperçu de la nature des déviations par rapport aux test de Shapiro/Kolmogorov-Smirnov. Ainsi, un qqplot peut montrer si la distribution présente une asymétrie (skewness), ou bien des outliers. Ils sont de même plus simples à interpréter et peuvent être utilisés avec une plus grande variété de distribution théorique.

Chapitre 2

Tests non paramétriques

2.1 Quelle est la différence entre un test apparié et un test non-apparié ? Illustrer avec des exemples concrets

Test apparié : Utilisé pour comparer deux échantillons de données qui ont été mesurés sur les mêmes sujets ou échantillons. Cela signifie que chaque échantillon est associé à un sujet ou un échantillon spécifique, de sorte que les différences entre les échantillons peuvent être attribuées à des différences entre les sujets ou les échantillons plutôt qu'à des différences aléatoires dans les mesures.

Test non apparié : Utilisé pour comparer deux échantillons de données qui n'ont pas été mesurés sur les mêmes sujets ou échantillons. Cela signifie que les échantillons ne sont pas associés à des sujets ou des échantillons spécifiques, de sorte que les différences entre les échantillons peuvent être attribuées à des différences aléatoires dans les mesures plutôt qu'à des différences entre les sujets ou les échantillons.

Supposons que l'on veuille déterminer si deux techniques d'étude différentes mènent ou non à des résultats moyens différents.
Pour effectuer un test apparié, nous recrutons 10 personnes et leur demandons d'utiliser une technique d'étude pendant un mois et de passer un examen, puis leur demander d'utiliser la deuxième technique d'étude pendant un mois et de passer un autre examen de difficulté égale.

Voici à quoi ressembleraient les données :

Studying Technique #1

	Exam Grade
Student #1	77
Student #2	79
Student #3	83
Student #4	84
Student #5	84
Student #6	87
Student #7	89
Student #8	90
Student #9	94
Student #10	95

Studying Technique #2

	Exam Grade
Student #1	79
Student #2	84
Student #3	80
Student #4	83
Student #5	83
Student #6	82
Student #7	80
Student #8	91
Student #9	92
Student #10	87

FIGURE 1 – Exemple d'illustration pour les tests appariés

Comme chaque personne figure dans chaque groupe, on utilisera un test apparié pour déterminer si les scores moyens sont différents entre les deux groupes.

Pour réaliser un test non apparié, recrutons 20 personnes au total et répartissons-les au hasard en deux groupes de 10. On demandera à un groupe d'utiliser une technique d'étude pendant un mois et demandera à l'autre groupe d'utiliser la deuxième technique d'étude pendant un mois. Tous les étudiants passeraient le même examen.

Voici à quoi ressembleraient les données :

Studying Technique #1

	Exam Grade
Student #1	77
Student #2	79
Student #3	83
Student #4	84
Student #5	84
Student #6	87
Student #7	89
Student #8	90
Student #9	94
Student #10	95

Studying Technique #2

	Exam Grade
Student #11	84
Student #12	78
Student #13	80
Student #14	76
Student #15	88
Student #16	89
Student #17	92
Student #18	93
Student #19	90
Student #20	86

FIGURE 2 – Exemple d'illustration pour les tests non appariés

Étant donné que les personnes d'un groupe sont totalement indépendants des personnes de l'autre groupe, on utilisera un test non apparié pour déterminer si les scores moyens sont différents entre les deux groupes.

2.2 Dans quel contexte utilise-t-on des tests non paramétriques ?

Un test non paramétrique est un test d'hypothèse qui n'exige pas que la distribution de la population soit caractérisée par certains paramètres. Par exemple, de nombreux tests d'hypothèse supposent que la population obéit à une loi normale. Comme les tests non paramétriques ne partent pas de cette hypothèse, ils s'avèrent utiles lorsque les données sont fortement non normales ou résistantes à transformation.

Les tests non paramétriques présentent les limites suivantes :

- Les tests non paramétriques sont généralement moins puissants que leurs équivalents paramétriques quand l'hypothèse de normalité est vérifiée. Ainsi, vous avez moins de chances de rejeter l'hypothèse nulle lorsqu'elle est fausse si les données obéissent à une loi normale.
- Ces tests requièrent souvent la modification des hypothèses. Par exemple, la plupart des tests non paramétriques relatifs au centre de la population utilisent la médiane au lieu de la moyenne. Le test ne répond pas à la même question que la procédure paramétrique correspondante si la population n'est pas symétrique.

2.3 Citer quatre exemple de tests non paramétriques, et les situations d'application correspondantes

- Test de Wilcoxon-Mann-Whitney utilise le rang de chaque observation pour tester si les groupes sont issus de la même population. Les tests de Mann-Whitney servent à vérifier que deux échantillons d'une population ont une position équivalente. Les observations des deux groupes sont combinées et ordonnées. Par exemple, on veut savoir si un nouveau type de plâtre utilisé pour réduire des fractures doit être porté plus longtemps que les anciens modèles. Dix personnes choisissent au hasard vont porter le nouveau modèle et dix autres l'ancien modèle. Le test U permettra de savoir si, en moyenne, les sujets portant le nouveau modèle ne doivent pas le porter aussi longtemps que les sujets portant l'ancien modèle.
- Le test de Kruskal-Wallis est utilisé pour déterminer s'il existe ou non une différence statistiquement significative entre les médianes de trois groupes indépendants ou plus. Par exemple, nous voulons savoir si trois médicaments ont des effets différents sur la douleur au genou. On recrute 30 personnes qui souffrent toutes de douleur au genou et les répartissons au hasard en trois groupes pour recevoir soit le médicament 1, soit le médicament 2, soit le médicament 3.
- Le test de la médiane de Mood est un test non paramétrique permettant de comparer les médianes de deux échantillons indépendants. Par exemple, on veut déterminer si la méthode de présentation utilisée par un enseignant a une influence sur la compréhension de son cours par les étudiants. On sélectionne des étudiants et leur assigne aléatoirement des cours utilisant l'une des trois méthodes de présentation suivantes : descriptions textuelles, photographies ou dessins.
- Le test de Friedman est un test statistique utilisé pour déterminer si trois mesures ou plus du même groupe de sujets sont significativement différentes les unes des autres sur une variable d'intérêt asymétrique. Par exemple, nous voulons savoir si le temps de réaction moyen des sujets est différent pour trois médicaments différents. Pour le vérifier, nous recrutons 10 patients et mesurons chacun de leurs temps de réaction (en secondes) sous l'effet des trois médicaments différents.

2.4 D'où vient l'expression "test de rangs" utilisée pour désigner les tests nonparamétriques ?

Contrairement aux test statistiques paramétriques qui se basent sur les valeurs des observations et la notion de barycentre (moyenne des observations), les tests non paramétrique se basent sur les rangs des observations et s'intéressent à l'ensemble de la distribution (somme des rangs). Par exemple, si une série de données comprend les valeurs 4, 6, 2, 8 et 3, les rangs correspondants seraient 2, 4, 1, 5 et 3. Les tests non paramétriques utilisent les rangs des données pour calculer des statistiques de test, comme le test-statistique ou la p-value, qui permettent de déterminer si les données suivent ou non une distribution théorique.

Chapitre 3

ANOVA

3.1 Expliquer comment est calculée une somme de carrés pour une variable donnée

La somme des carrés SC (Sum of Squares, SS) est une mesure de variation ou d'écart par rapport à la moyenne. Elle représente la somme des carrés des différences par rapport à la moyenne. Le calcul de la somme totale des carrés prend en compte les différences dues aux facteurs et de celles dues au hasard ou à l'erreur. La formule est la suivante :

$$SC = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

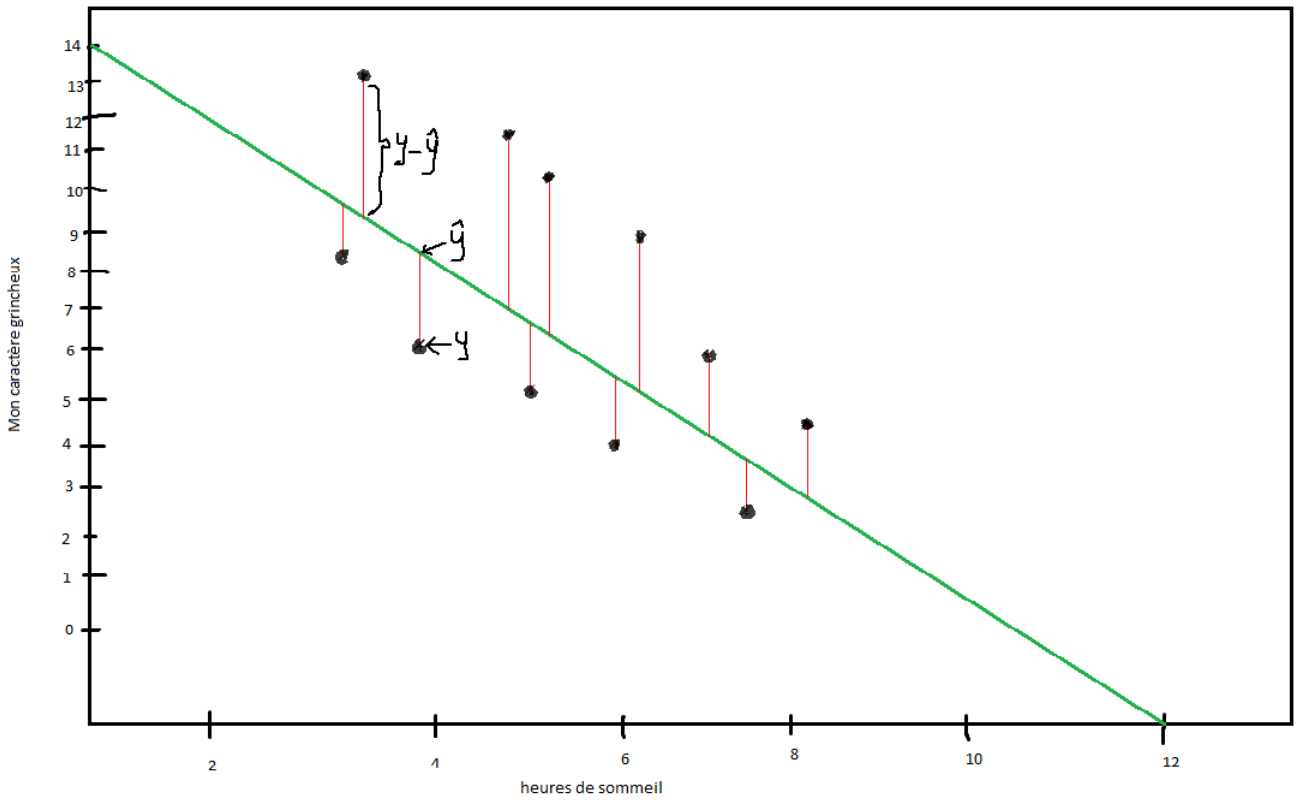

FIGURE 3 – Représentation de mon état grincheux au réveil par rapport à mon temps de sommeil

La différence entre la valeur prédite ($\hat{y}$) et la valeur actuelle (y) est calculée puis mit au carré. Ce processus est répété pour l'ensemble des points présent puis ils sont additionnés.

3.2 Comment calcule-t-on une statistique F pour une variable donnée ?

La statistique F est une mesure de la différence entre les groupes dans une analyse de variance (ANOVA). Elle est calculée en comparant la variabilité des données entre les groupes et la variabilité des données dans les groupes. La formule pour calculé une statistique F est la suivante :

$$F_s = \frac{MS_{between}}{MS_{within}} = \frac{\frac{SS_{between}}{df_{between}}}{\frac{SS_{within}}{df_{within}}} = \frac{\frac{SS_{between}}{k-1}}{\frac{SS_{within}}{n-k}}$$

MS between = variance entre les groupes,
MS within = variance au sein des groupes,
SS Between est la somme des carrés entre les moyennes des groupes et la grande moyenne,
SS within est la somme des carrés au sein des groupes,
df = degré de liberté,
n = Nombre total d'observations dans l'échantillon,
k = Degrés de liberté

3.3 Comment passe-t-on du F à la p-value ? Illustrer obligatoirement avec un graphe

Supposons que nous voulons une hypothèse nulle telle que $H_0 : \mu_1 = \mu_2 = \mu_3$.

	Df	Sum Sq	Mean Sq	F-value
Genre	2	1867	3721	2.53
IDH	30	7433	1467	
Total	5	9300		

TABLE 1 – Table ANOVA du nombre de vidéos vues par rapport au genre et IDH

La F-value est calculée par le rapport de $MeanSq_{Genre}$ sur $MeanSq_{IDH}$, soit 3721 / 1467 = 2.53.

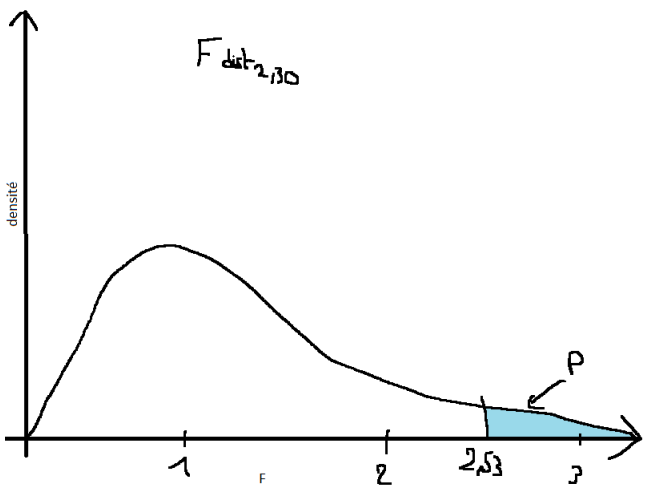

FIGURE 4 – Courbe de F distribution avec F-value

La P-value est la partie colorée en bleu dans la Figure 4. Nous recherchons donc via une table de F-value les valeurs p encadrant notre F-value.

TABLE E

F critical values (continued)

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

FIGURE 5 – Table F distribution

En suivant le même ordre que pour le calcul de la F-value, et utilisant les degrés de liberté, nous avons 2 comme numérateur et 30 comme dénominateur . Nous obtenons que notre P-value est compris $p = 0.1$ et $p = 0.05$. Autrement dit, $0.05 <$ P-value < 0.1

La Figure 6 représente la courbe F-distribution et les courbes des p-value. Le score F de ces deux courbes encadre le score F obtenue à la Table 1, nous indiquant que la P-value de notre score F se situe entre ces deux scores F.

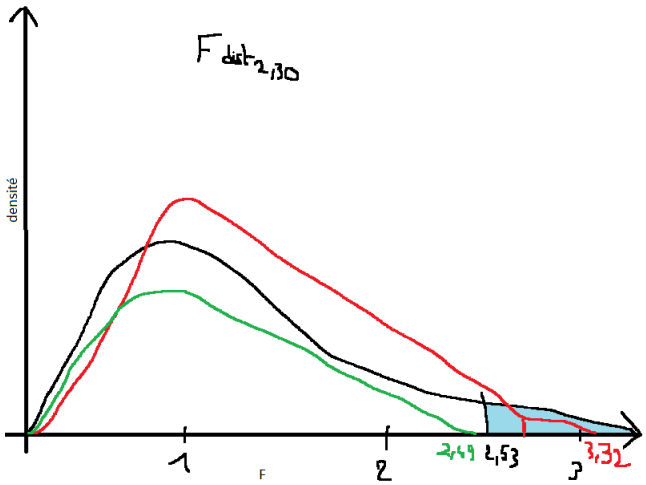

FIGURE 6 – Courbe de F distribution avec F-value et les courbes des p-value choisit

3.4 Que représente l'interaction entre deux facteurs dans une ANOVA ? Illustrer avec un cas concret

L'interaction entre deux facteurs est l'effet d'un facteur n'a pas le même effet sur la variable dépendante selon les modalités de l'autre facteur. Cela permet d'évaluer si des deux variables agissent conjointement sur la variable réponse, ou non. Si l'évolution de la réponse en fonction des différentes modalités de la première variable, ne dépend pas des modalités de la seconde variable, alors il n'existe pas d'interaction entre les deux variables. Si au contraire, on observe une modification de cette évolution, soit par une augmentation de l'effet de la première variable, soit par une diminution, alors il existe une interaction.

Prenons le cas de métiers à tisser, et regardons le nombre de rupture du fil de laine par rapport à la tension appliquée et ce pour deux types de laines différentes.

Sur la première représentation de la Figure 7, l'évolution du nombre de ruptures en fonction du niveau croissant de tension est identique pour les deux types de laines, puisque les profils sont parallèles. Il n'y a donc pas d'interaction.

Sur la seconde représentation, le nombre de ruptures en fonction du niveau croissant de tension, augmente plus rapidement pour la laine de type B. Il y a alors une interaction entre la tension du fil et le type de laine avec ici un effet synergique. Lorsque les profils ont la même direction, mais avec des "vitesses" différentes, on parle parfois d'interaction "quantitative".

Sur la dernière représentation, les évolutions du nombre de ruptures en fonction de la tension sont contraires. Lorsque les profils se croisent, l'interaction est parfois appelée "qualitative".

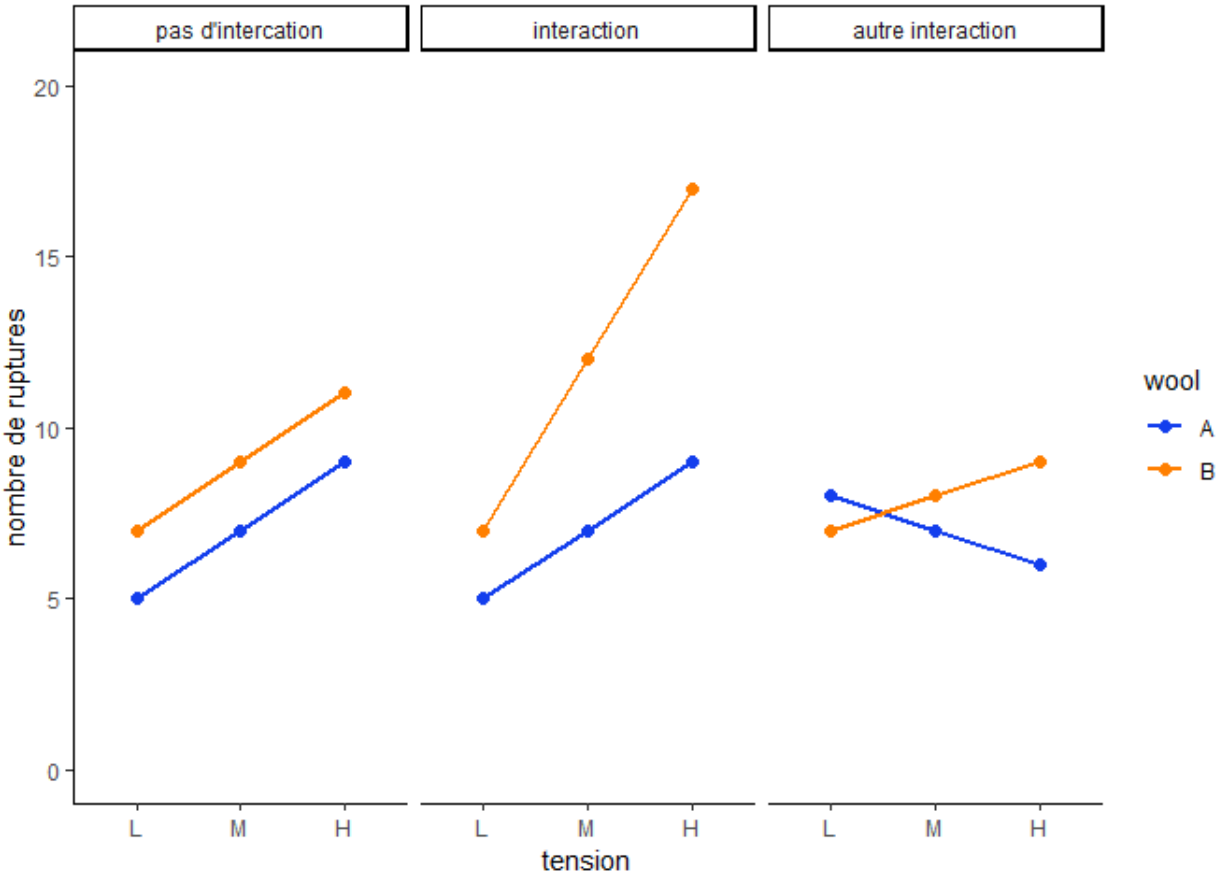

FIGURE 7 – Représentation des différentes interactions d'une ANOVA à 2 facteurs

Chapitre 4

Régression logistique

4.1 Dans quel contexte utilise-t-on des régressions de type ”loi de Poisson” ?

La régression de Poisson est utilisée pour prédire une variable dépendante constituée de "données de comptage" en fonction d'une ou plusieurs variables indépendantes et quand les données ne suivent pas une distribution normale. Par exemple, on peut utiliser la régression de Poisson pour examiner le nombre d'élèves suspendus par une école.

Par exemple, une régression de type loi de Poisson peut être utilisée pour modéliser le nombre de clients qui entrent dans un magasin au cours d'une journée donnée, en fonction de facteurs tels que la météo, les promotions en cours, la concurrence, etc. Dans ce cas, la variable dépendante (nombre de clients) suit une distribution de Poisson, et les facteurs explicatifs (météo, promotions, concurrence, etc.) sont utilisés pour expliquer les variations dans le nombre de clients.

4.2 Pourquoi au juste utilise-t-on le terme ”Poisson” pour désigner ce type de régression logistique ?

Le terme ”Poisson” est utilisé car une régression de type ”Poisson” suit une la loi de Poison ou de quasi-Poison. Cette loi porte le nom du mathématicien français Siméon Denis Poisson.

4.3 Pourquoi au juste utilise-t-on le terme ”binomiale” pour désigner un autre type de régression logistique ?

Une régression logistique binomiale, aussi appelée régression logistique, prédit la probabilité qu'une observation entre dans l'une des deux catégories d'une variable dépendante dichotomique en fonction d'une ou plusieurs variables indépendantes. Il s'agit de données prenant la forme de deux réponses possible, tel que 0 ou 1, homme ou femme etc.

	Two Class Classification	
$y \in \{0, 1\}$	**1 or Positive Class**	**0 or Negative Class**
Email	Spam	Not Spam
Tumor	Malignant	Benign
Transaction	Fraudulent	Not Fraudulent

FIGURE 8 – Exemple de table binomiale

4.4 A quoi correspond un odd-ratio et en quoi les OR diffèrent-ils du risque relatif ?

L'Odds Ratio, noté OR, également appelé rapport des chances ou rapport des cotes, est une approche non paramétrique permettant de mesurer l'association entre deux variables X_1, X_2 en déterminant la chance/le risque qu'un évènement de X_2 se produise sachant les valeurs de X_1. Le Risque Relatif, noté RR, est une approche non paramétrique permettant de mesurer l'association entre deux variables X_1, X_2 en déterminant la chance/le risque qu'un évènement de X_2 se produise dans l'un des deux groupes de X_1 par rapport à l'autre groupe de cette même variable. L'Odds Ratio est le rapport entre la cote d'un évènement chez le premier sous-groupe et la cote de ce même évènement chez le second groupe alors que le Risque Relatif est le rapport entre le risque de voir un évènement se produire chez le premier sous-groupe et le risque de voir ce même évènement se produire chez le second sous-groupe.

4.5 Dans quelle condition risques relatifs et OR sont-ils approximativement les mêmes ?

L'Odds Ratio converge vers le Risque Relatif quand le nombre d'évènements de l'Odds Ratio est faible.

Table des figures

1 Exemple d'illustration pour les tests appariés 5

2 Exemple d'illustration pour les tests non appariés 6

3 Représentation de mon état grincheux au réveil par rapport à mon temps de sommeil 8

4 Courbe de F distribution avec F-value 9

5 Table F distribution 10

6 Courbe de F distribution avec F-value et les courbes des p-value choisit 11

7 Représentation des différentes interactions d'une ANOVA à 2 facteurs 12

8 Exemple de table binomiale 14

Liste des tableaux

1 Table ANOVA du nombre de vidéos vues par rapport au genre et IDH 9