Analyse de Contenu et de Données
Etudes de variables deux à deux : corrélations

Guillaume Metzler
guillaume.metzler@live.fr

Univ. Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France

Printemps 2020

Résumé

Jusqu'à présent vous avez:

- appris à lire un tableau de données et vérifier la qualité de votre échantillon (cours 1)
- représenté sous forme de tableaux/graphiques les résultats d'une enquête (cours 2)
- analysé avec des outils statistiques (moyenne/variance) les résultats d'une enquête (cours 3)

$\rightarrow$ **Etude d'une variable indépendamment des autres variables !**

Mais ce qui est intéressant, c'est d'étudier les liens entre les variables entres les différentes variables !

Pour ce dernier cours :

- Quelques rappels sur la théorie des tests statistiques.
- Tests statistiques en fonction du type de variable.
- Etude de la corrélation (définition et dangers).
- Mise en pratique avec Sphinx.

Tests Statistiques

Pour effectuer des tests en statistiques il convient de :

1) Définir les variables concernées par le test ainsi que leur nature.
2) Formuler une hypothèse a priori vraie concernant nos variables.

Hypothèses

Les tests statistiques nécessitent de formuler une hypothèse H_0 qui sera considérée comme vraie et qui définira **la statistique de test**. Une hypothèse H_1 est également définie comme **hypothèse alternative** en cas de **rejet de l'hypothèse** H_0.

Hypothèses

Exemple

Dans le cadre de ce cours nous allons essentiellement étudier les corrélation entre deux variables. Les hypothèses formulées seront les suivantes :

- H_0 : les deux variables ne sont pas corrélées
- H_1 : les deux variables sont corrélées

D'autres exemples d'utilisations des tests statistiques : comparaison de moyenne (H_0 : les deux moyennes sont égales vs H_1 : les deux moyennes sont différentes.)
Ou encore pour tester la significativité des paramètres d'un modèles statistique.

Statistique de test

Définition

La **statistique de test** une **variable aléatoire** qui va permettre de formuler une réglé de décision dans le cadre d'un test statistique. Cette variable aléatoire est construite à partir de l'échantillon sur lequel est effectué le test.

La loi de la **statistique de test** est conditionnée à l'hypothèse nulle, i.e. elle dépend de H_0.

Coefficient de corrélation

Pour tester si la corrélation entre deux variables est significatives, la statistique de test S utilisée (sous H_0 : les deux variables ne sont pas corrélées) suit une loi de Student à $(n-2)$ degrés de libertés.

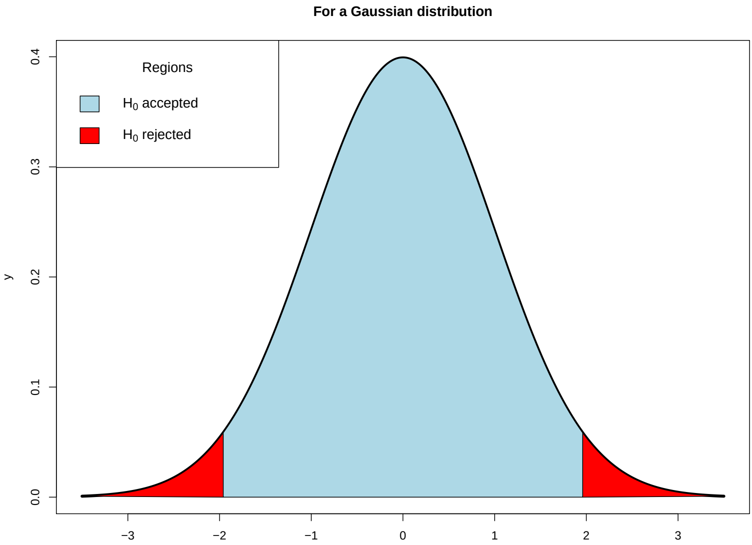
Région de rejet
Lemme de Neyman-Pearson (1931)

Lorsque l'hypothèse et la statistique de test sont définis, il convient maintenant de définir une **région de rejet** de l'hypothèse H_0, qui si la statistique de test S se trouve dans cette région là, alors l'hypothèse H_0 est rejetée.

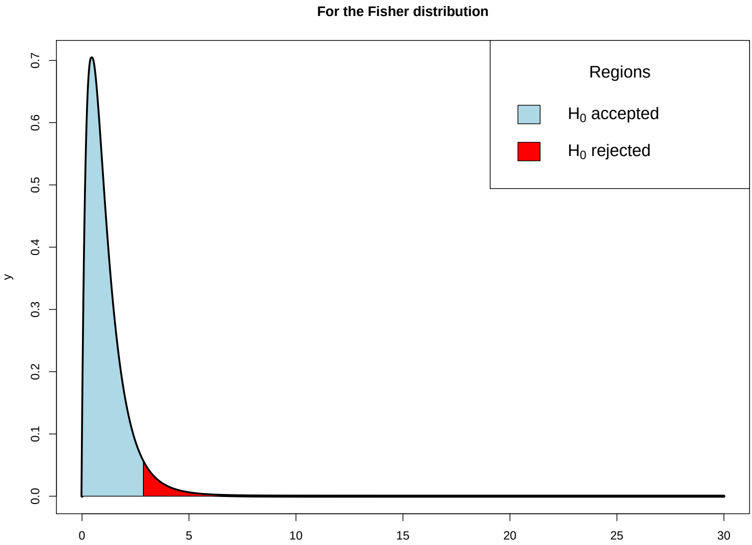
On définit des régions de rejet pour deux sortes de test :

- **Les tests bilatéraux :** on **rejette l'hypothèse** H_0 si la valeur observée de la statistique est **en dehors** d'un intervalle de la forme $\left[I_{\frac{\alpha}{2}}, I_{1-\frac{\alpha}{2}}\right]$.
- **Les tests unilatéraux** (gauche ou droite) : on **rejette l'hypothèse** H_0 si la valeur observée de la statistique se trouve dans un intervalle de la forme $[-\infty, I_\alpha]$ ou $[I_{1-\alpha}, +\infty]$.

Région de rejet
Distribution Gaussienne



Région de rejet
Distribution de Fisher



Probabilité Critique

On utilise très rarement la valeur de façon directe cette valeur pour conclure ou non au rejet de l'hypothèse H_0 car elle nécessite l'usage de table de statistique et de connaître les quantiles de notre distribution.

Un critère communément employé est la **p-value** (ou probabilité critique).

Probabilité critique

Il s'agit de la probabilité, calculée sous l'hypothèse H_0, d'observer **une valeur plus grande** (**ou plus petite**, selon la nature de la région de rejet) que la valeur observée de la statistique de test S, notée S_{obs}.
Plus formellement : p-value$= \mathbb{P}[X > S_{obs}]$ où X est une variable aléatoire suivant la même loi que S.

p-value et risque de première espèce α du test

Lorsque l'on effectue un test statistique, on se fixe un seuil de probabilité de **rejeter l'hypothèse** H_0 **à tort** notée α.

α et risque de première espèce

- La valeur α est appelée **risque de première espèce** de première espèce et correspond à la probabilité d'accepter l'hypothèse H_1 sachant que H_0 est vraie.
- La valeur $1 - \alpha$ est appelé **confiance du test**.

α et p-value

C'est ce risque de première espèce que nous allons le plus souvent utilisé pour valider ou non un test statistique ! Plus précisément, on va effectuer les étapes suivantes :

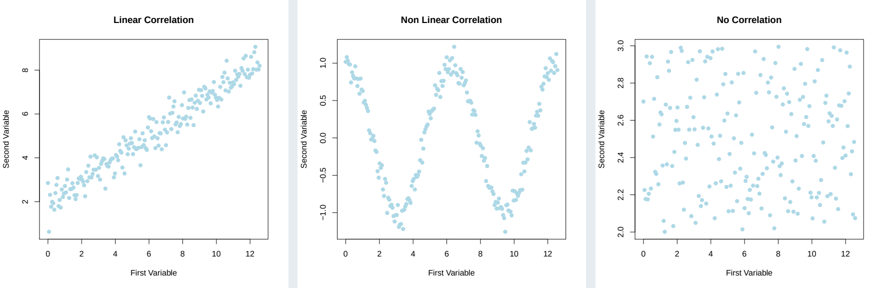
- Fixer H_0 notre hypothèse vraie *a priori*
- Fixer le risque α de rejeter l'hypothèse H_0 à tort (on prendra très souvent $\alpha = 0.05$).
- Calculer la statistique de test sur notre échantillon S_{obs}
- Calculer la p-value, i.e. $\mathbb{P}[X > S_{obs}]$ pour un test unilatéral à droite ou $\mathbb{P}[|X| > S_{obs}]$ pour un test bilatéral (cela se fait à l'aide de logiciels de statistiques comme SAS, SPSS ou R).
- Cette valeur est ensuite comparée à α. Si la p-value est **plus petite** que α alors on rejette l'hypothèse H_0. En effet, il est peu probable d'observer des valeurs plus rares (statistiquement) que celle observée. Dans le cas contraire, l'hypothèse H_0 est conservée.

Corrélations entre deux variables

Contexte

La théorie des tests statistiques va permettre d'affirmer si oui ou non, il y a bien **une corrélation** entre les deux variables étudiées, i.e. est-ce qu'il y a **une liaison** entre deux variables.

Plusieurs types de liaisons



Le calcul du **coefficient de corrélation de Pearson (1857-1927)** permet de déterminer comment la variable 1 agit sur la variable 2.

Définition

Le **coefficient de corrélation de Pearson** r entre deux variables **quantitatives** X et Y est définie par :

$$r = \frac{\mathbb{E}[(X - \mathbb{E}[X]])((Y - \mathbb{E}[Y]])}{\sqrt{Var(X)Var(Y)}}.$$

Si l'on dispose d'un échantillon de taille m on peut déterminer une valeur empirique de cette quantité :

$$\hat{r} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})\sum_{i=1}^{m}(y_i - \bar{y})^2}}.$$

Le coefficient de corrélation prend des valeurs dans $[-1, 1]$.

Interprétation

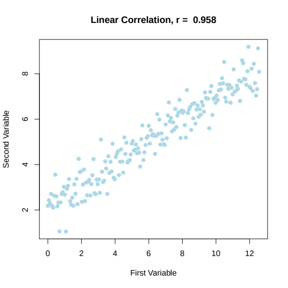
Lorsque le coefficient de corrélation de Pearson r entre deux variables numériques X et Y est :

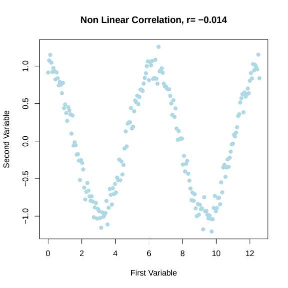
- **positif :** on dit que la **liaison est positive** entre les deux variables, i.e. des valeurs de X croissantes impliquent des valeurs croissantes de Y.
- **négatif :** on dit que la **liaison est négative** entre les deux variables, i.e. des valeurs de X croissantes impliquent des valeurs décroissantes de Y.

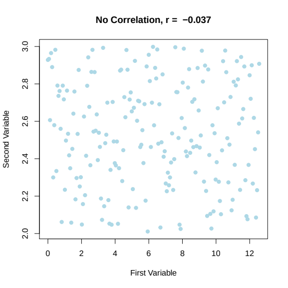
L'intensité de cette liaison est donnée par la **valeur absolue de** r, plus la valeur est proche de 1 plus la liaison est forte. En revanche plus elle proche de 0, moins les deux variables sont liées ou corrélées.

Coefficient de corrélation

Exemples et remarques





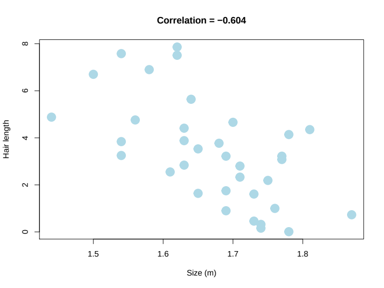


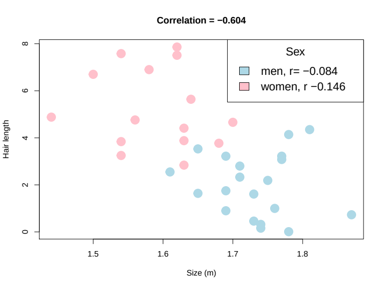
- Le coefficient de corrélation ne caractérise que les liaisons linéaires !
- Un r nul ne signifie pas qu'il n'y a pas de lien entre les deux variables mais absence de **liens linéaires**.
- **Corrélation n'implique pas forcément un lien de cause à effet.**

Corrélation et lien de cause à effet

Contre-exemple

Taille vs longueur de cheveux





La corrélation est globalement négative, mais lorsque l'on prend en compte le sexe de la personne, on remarque qu'au sein d'un même groupe, la corrélation est inexistante.

On parle de **facteur confondant**, i.e. la corrélation se cache dans une variable non prise en compte pour l'étude de corrélation

Etude de la corrélation selon les variables

Vous serez amenés à étudier les corrélations entre :

1) deux variables quantitatives
2) une variable quantitative - une variable qualitative
3) deux variables qualitatives

Selon la nature des variables, nous utiliserons différents tests statistiques :

1) **coefficient de corrélation** avec **un test de Student**
2) **une analyse de variance**
3) **un test du** χ^2 (lisez Khi deux)

Cas de deux variables quantitatives

On considère deux grandeurs x et y dont les valeurs sont les suivantes :

Données

x	9.0	10.0	16.0	12.5	19.0	22.5	16.0	12.5	17.0	28.0
y	14.0	12.0	11.1	14.2	9.7	7.5	11.1	11.4	11.8	10.1

- Calcul du coefficient de corrélation r.
  Dans notre exemple $\hat{r}$ est égal à -0.85, il y a donc une relation négative entre les deux variables (**a priori forte !**).
- On fixe l'hypothèse H_0 : les deux variables ne sont pas corrélées et le seuil de risque $\alpha = 0.5$.
- Il reste donc à tester si la valeur de r est significativement différente de 0, i.e. si la corrélation est significative.

Cas de deux variables quantitatives

Pour cela, on calcule la statistique de test $S_{obs} = \dfrac{\hat{r}}{\sqrt{(1-\hat{r}^2)/(n-2)}}$, où n représente le nombre d'individus. On a donc $S_{obs} = -5.19$.

- La statistique de test suit une loi de Student à $(n-2)$ degrés de liberté et la **p-value** de notre test est égale à $\mathbb{P}[S > |S_{obs}|] = 2.10^{-4}$.

Conclusion

On peut donc rejeter l'hypothèse H_0 selon laquelle les deux variables sont indépendantes car le coefficient de corrélation a une valeur significativement différente de 0. Les deux variables sont donc liées et **cette liaison est forte** et **négative**.

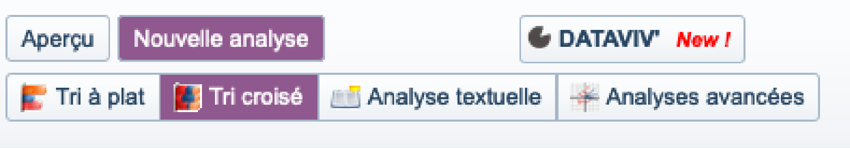
$\rightarrow$ **Dans la suite et pour les autres tests statistiques, Sphinx vous donnera directement les p-values**

Mise en pratique sur Sphinx

Procédure

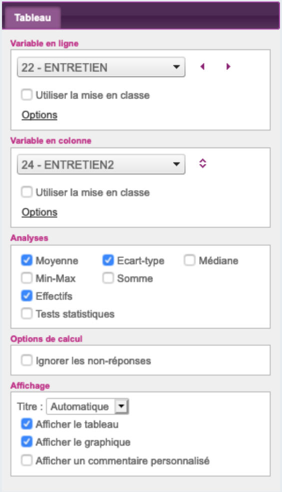
Pour cela reprenons l'enquête Automobiles et retournons dans l'onglet **Analyses**:

Pour effectuer une analyse bi-variée, cliquez ensuite sur **Tris Croisés**:



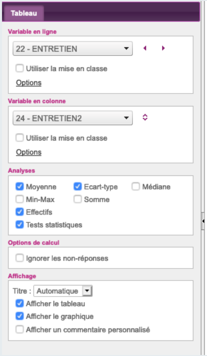
Procédure

- Le menu ci-contre devrait alors s'afficher sur la gauche de votre écran
- C'est ici que nous sélectionnerons les variables à étudier et que nous aurons la possibilité d'effectuer une analyse de corrélation
- Sélectionnons par exemple les variables **Entretien** et **Entretien2** et cochez la case *Tests Statistiques*



Mise en pratique

Résultat



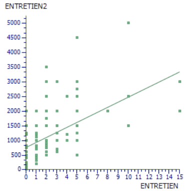
22 - ENTRETIEN / 24 - ENTRETIEN2

	Moyenne	Ecart-type	Effectif
ENTRETIEN	2,1	2,73	175
ENTRETIEN2	1132	812,74	175

Les valeurs en bleu / rouge sont significativement supérieures / inférieures à la grande moyenne (au seuil de risque de 5%).

Réponses effectives : 175 **Non-réponse(s) : 0** **Taux de réponse : 100%**
Corrélation = 0,58. L'ajustement entre les deux variables est plutôt faible (t = 9,27).

ENTRETIEN2 = 171 * ENTRETIEN + 771



Les deux variables que l'on étudie ici sont **quantitatives**. On s'intéresse donc au coefficient de corrélation.

On retrouve plusieurs éléments sur cette page.

Mise en pratique

Informations

- Notez tout d'abord l'ordre entre les variables : ici on étudie l'influence de **Entretien** (axe des abscisses) sur **Entretien2** (axe des ordonnées). En pratique, nous pourrions échanger l'ordre des variables cela ne changera rien à la valeur du coefficient de corrélation et donc du test.
- Un graphique représentant les données avec une **régression linéaire**
- Ce qui nous intéresse :

Réponses effectives : 175 Non-réponse(s) : 0 Taux de réponse : 100%
Corrélation = 0,58. L'ajustement entre les deux variables est plutôt faible (t = 9,27).

La valeur du coefficient de corrélation, 0.58 et la valeur de la statistique de test t égales à 9.27 (ce qui correspond à une p-value $\simeq 0$). **La relation est donc significative**.

Mise en pratique

Remarques

- Sphinx ne vous donnes pas la p-value dans la cas ou on étudie deux variables quantitatives. Donc on va se fier à la valeur de t. Mais ce n'est pas grave !
- Pour décider si la relation est significative, on regarde cette valeur t et le nombre d'exemples. Si on dispose d'au moins 50 observations et que $|t| > 2$, alors on peut rejeter l'hypothèse d'indépendance entre les deux variables : **la relation est donc significative**.

Mise en pratique

Regardons maintenant le cas de la corrélation entre une variable **quantitative : Entretien2** et une variable **qualitative : Marque**. La statistique de test suit cette fois-ci une **Loi de Fisher**.

Exemple

ENTRETIEN2 →	Moyenne	Ecart-type	Effectif
MARQUE ↓			
Renault	1026,39	684,89	36
Peugeot	1035,71	646,17	21
Décret	1108,33	884,16	12
BMW	2138,89	1179,98	9
Citroën	723,81	510,79	21
Gué	857,14	420,1	14
Mercedes-Benz	1906,25	778,48	8
Opel	965	558,79	10
Toyota	1350	1125,83	3
Volvo	733,33	404,15	3
Wolkswagen	1250	1088,04	18
Autre	1300	797,53	20
Total	**1132**	**812,74**	**175**

Les valeurs en bleu / rouge sont significativement supérieures / inférieures à la grande moyenne (au seuil de risque de 5%).

Réponses effectives : 175 **Non-réponse(s) : 0** **Taux de réponse : 100%**
p-value = < 0,01 ; Fisher = 3,27. La relation est très significative.

Mise en pratique

Analyse

Réponses effectives : 175 Non-réponse(s) : 0 Taux de réponse : 100%
p-value = < 0,01 ; Fisher = 3,27. La relation est très significative.

- On étudie ici la corrélation entre **Entretien2** et **Marque**.
- Dans la cas présent, le logiciel renvoie à la fois **statistique de test** de Fisher égale à 3.27 et la **p-value** < 0.01.
- Etant donné que la p-value a une valeur plus faible que le seuil de risque fixé pour effectuer la test ($\alpha = 0.05$). On peut donc rejeter l'hypothèse d'indépendance et conclure à une corrélation entre les deux variables.

Mise en pratique

Finalement, étudions le cas de la corrélation entre deux variables **qualitatives : Marque** et **Sexe** . La statistique de test employée suit une loi du χ^2.

Exemple

2 - MARQUE / 31 - SEXE

SEXE → / MARQUE ↓	Homme			Femme			Total	
	Eff.	% Obs.	Ecart	Eff.	% Obs.	Ecart	**Eff.**	**% Obs.**
Renault	16	44,4%		20	55,6%		**36**	**100%**
Peugeot	8	38,1%		13	61,9%		**21**	**100%**
Décret	5	41,7%		7	58,3%		**12**	**100%**
BMW	7	77,8%	+ PS	2	22,2%	- PS	**9**	**100%**
Citroën	11	52,4%		10	47,6%		**21**	**100%**
Gué	8	57,1%		6	42,9%		**14**	**100%**
Mercedes-Benz	6	75%		2	25%		**8**	**100%**
Opel	6	60%		4	40%		**10**	**100%**
Toyota	2	66,7%		1	33,3%		**3**	**100%**
Volvo	2	66,7%		1	33,3%		**3**	**100%**
Wolkswagen	8	44,4%		10	55,6%		**18**	**100%**
Autre	12	60%		8	40%		**20**	**100%**
Total	**91**	**52%**		**84**	**48%**		**175**	

Les pourcentages sont calculés par rapport au nombre d'observations en ligne.

Les valeurs en bleu / rouge sont significativement sur représentées / sous représentées (au seuil de risque de 5%).

Réponses effectives : 175 Non-réponse(s) : 0 Taux de réponse : 100%
p-value = 0,63 ; Khi2 = 8,90 ; ddl = 11,00. La relation n'est pas significative.

Mise en pratique

Analyse

- La table précédente est appelée **table de contingence**.
- Pour calculer la statistique de test, il faudra calculer les effectifs théoriques des éléments de cette table. Donc ... passons aux résultats.

Réponses effectives : 175 Non-réponse(s) : 0 Taux de réponse : 100%
p-value = 0,63 ; Khi2 = 8,90 ; ddl = 11,00. La relation n'est pas significative.

- Dans la cas présent, le logiciel renvoie à la fois **statistique de test** du χ^2 égale à 8.90 et la **p-value** est égale à 8.90.
- Etant donné que la p-value a une valeur **plus grande** que le seuil de risque fixé pour effectuer la test ($\alpha = 0.05$). **On peut ne peut donc pas rejeter l'hypothèse d'indépendance** entre les deux variables.