Cours Statistiques et Analyse de Données

Cours 4

FILIÈRE ISA - UV1

R. Billot et G. Coppin

2016 - 2017

Objectifs du cours

Objs:

1. Finir les tests
2. Jouer avec dépendance et indépendance de variables

Petits rappels

Les tests d'hypothèse statistiques se ramènent à :

1. modéliser le problème d'un point de vue statistique
2. définir une hypothèse nulle H_0 vis-à-vis du problème à traiter
3. choisir un test statistique ou - ce qui revient au même - une (variable) statistique pour le test : il s'agit d'une variable aléatoire qui doit permettre de choisir entre H_0 et H_1
4. définir la distribution de la variable statistique pour H_0
5. définir le niveau de signification du test ou risque
6. calculer à partir de l'échantillon la variable statistique
7. prendre une décision à partir du positionnement de la valeur (seuil associé au risque) ou à partir de la p-valeur obtenue

Petits rappels - questions (ii)

- Quel est le rapport entre risque et p-value (ou plus précisément comment utilise t on l'un ou l'autre ?)
- Que nous permet de faire le théorème central limite ?
- Comment estimer un intervalle de confiance autour d'une valeur estimée ?

Les calculs d'intervalle de confiance et les tests ne sont pas forcément neutres (I)

Une étude menée fin 2008 sur 298 logements parisiens choisis au hasard dans l'annuaire assure que le prix du loyer au mètre carré est de 18,4 euros, avec un écart-type mesuré de 3,2 euros.

- modéliser brièvement la situation
- on suppose que l'enquête est demandée par l'Observatoire des loyers parisiens. Comment exprimeriez-vous l'intervalle de confiance relatif au prix moyen du loyer au mètre carré à Paris.
- on suppose que l'enquête est demandée par le Collectif Jeudi Noir. Même question.
- on suppose que l'enquête est demandée par la Confédération nationale des propriétaires-bailleurs. Même question.

Les calculs d'intervalle de confiance et les tests ne sont pas forcément neutres (II)

- Population = ?
- Données / variables = ? Indépendantes ?
- Paramètre d'intérêt (estimateur) = ?
- Intervalle de confiance = ?

Les calculs d'intervalle de confiance et les tests ne sont pas forcément neutres (III)

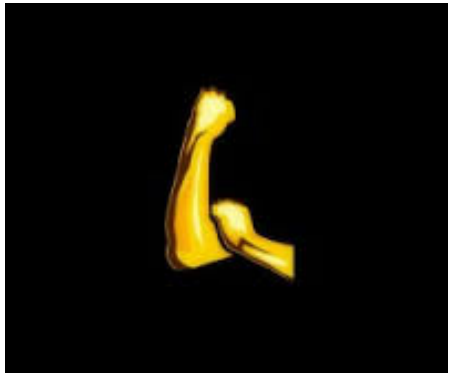
Vous avez obtenu un résultat de 9.5 à l'examen du terrible Professeur Patrick M. du département LUSSI, alors que sur 30 notes, la moyenne est de 10 pour un écart type de 3,2. Vous cherchez à négocier avec vos responsables de filières l'obtention de la totalité des crédits(parce que quand même ...) et vous vous procurez les résultats de l'année précédente (sur 30 notes également). Vous constatez que la moyenne mesurée est de 10,8 avec un écart type identique de 3,2. Pouvez vous argumenter et comment ?

Les calculs d'intervalle de confiance et les tests ne sont pas forcément neutres (IV)

Une seule chance, vous avez intérêt à montrer que l'examen était plus dur cette année, en considérant l'hypothèse H_0 $\mu_{2014} = \mu_{2013}$ et donc la statistique :

$$T = \sqrt{30}\frac{10,8 - 10}{3,2} = 1,37$$

Et effectivement, la valeur de $F^{-1}(1-\alpha)$ pour un test de Student à 90% est de 1,3 .. en unilatère. Vous avez donc intérêt à présenter l'hypothèse H_0 sous la forme unilatère. Sinon, Patrick M. ne vous rattrapera pas.
Maintenant, les enseignants de LUSSI ne sont pas non plus ignares en statistiques : que pensez vous que Patrick M. va vous répondre ?



Tests paramétriques et non paramétriques

- Paramétriques : on fait des hypothèses sur la loi sous-jacentes aux variables (et on ajuste les paramètres de cette loi). Ex : Test de Student sur une moyenne de lois normales $X_1, X_2, .., X_k$
- Non paramétriques : pas d'hypothèse sur la nature de la distribution (**distribution free**).

Question	Données	Hypothèse nulle	Exemple	Tests paramétriques	Equivalents non-paramétri
Comparaison d'une moyenne observée avec une tendance théorique	mesures sur 1 échantillon ; moyenne théorique (1 chiffre)	moyenne observée = moyenne théorique	Comparaison à une norme d'un taux de pollution mesuré	Test t pour un échantillon	
Comparaison de deux positions* observées (échantillons indépendants)	mesures sur 2 échantillons	Les positions* sont identiques	Comparaison de notes d'étudiants entre deux classes	Test t pour échantillons indépendants	Mann-Whitney
Comparaison de plusieurs positions* observées (échantillons indépendants)	mesures sur plusieurs échantillons	Les positions* sont identiques	Comparaison du rendement de maïs selon 4 engrais différents	ANOVA	Kruskal-Wallis
Comparaison de deux positions* observées (échantillons dépendants)	deux séries de mesures quanti sur les mêmes individus (avant-après)	Les positions* sont identiques	Comparaison du taux d'hémoglobine moyen avant / après l'application d'un traitmeent sur un groupe de patients	Test t pour échantillons appariés	Wilcoxon
Comparaison de plusieurs positions* observées (échantillons dépendants)	Plusieurs séries de mesures quanti sur les mêmes individus (avant-après)	Les positions* sont identiques	Suivi de la concentration d'un élément trace au cours du temps au sein d'un groupe de plantes	ANOVA à mesures répétées; modèles mixtes	Friedman
Comparaison de plusieurs séries de mesures binaires (échantillons dépendants)	Plusieurs séries de mesures binaires sur les mêmes individus (avant-après)	Les positions* sont identiques	Différents juges évaluent la présence/l'absence d'un attribut sur différents produits		Test Q de Cochran
Comparaison de 2 variances (peut être utilisé pour tester condition 3)	Mesures sur deux échantillons	variance(1) = variance(2)	Comparaison de la dispersion naturelle de la taille de 2 variétés d'un fruit	Test de Fisher	
Comparaison de plusieurs variances (peut être utilisé pour tester condition 3)	Mesures sur plusieurs échantillons	variance(1) = variance(2) = variance(n)	Comparaison de la dispersion naturelle de la taille de plusieurs variétés d'un fruit	Test de Levene	
Comparaison d'une proportion observée avec une proportion théorique	une proportion observée ; son effectif associé ; une proportion théorique	proportion observée = proportion théorique	Comparaison de la proportion de femelles à une proportion de 0.5 dans un échantillon	Test pour une proportion (khi²)	
Comparaison de plusieurs proportions observées	Effectif de chaque catégorie	proportion(1) = proportion(2) = proportion(n)	Comparaison des proportions de 3 couleurs d'yeux dans un échantillon	khi²	
Comparaison de proportions observées à des proportions théoriques	Proportion théorique et effectif associés à chaque catégorie	proportions observées = proportions théoriques	Comparer les proportions de génotypes obtenus par croisement F1xF1 à des proportions mendéliennes (1/2, 1/4, 1/2)	Test d'ajustement multinomial	
Test d'association entre deux variables qualitatives	Tableau de contingence	variable 1 et variable 2 sont indépendantes	La présence d'un attribut est-elle liée à la présence d'un autre attribut?	khi² sur un tableau de contingence	Test exact de Fisher ; métho Monte Carlo
Test d'association entre deux variables quantitatives	mesures de deux variables sur un échantillon	variable 1 et variable 2 sont indépendantes	La biomasse de plante change-t-elle avec la concentration de Pb?	Corrélation de Pearson	Corrélation de Spearman
Comparer une distribution observée à une distribution	Mesures d'une variable quantitative sur un échantillon;				

Comparer une distribution observée à une distribution théorique	Mesures d'une variable quantitative sur un échantillon paramètres de la distribution théorique	Les distributions observée et théorique sont les mêmes	Les salaires d'une société suivent-ils une distribution normale de moyenne 2500 et d'écart-type 150?		Kolmogorov-Smirnov
Comparer deux distributions observées	Mesures d'une variable quantitative sur deux échantillons	Les deux échantillons suivent la même distribution	Les distributions de poids humain sont-elles différentes entre ces deux régions?		Kolmogorov-Smirnov
Tests pour les valeurs extrêmes	Mesures sur un échantillon	L'échantillon ne comprend pas de valeur extrême (selon la distribution normale)	Cette donnée est-elle une valeur extrême?	Test de Dixon / test de Grubbs	Boxplot
Tests de normalité d'une série de mesures (peuvent être utilisés pour tester les conditions 2, 4, 7)	Mesures sur un échantillon	L'échantillon suit une distribution normale	La distribution observée s'écarte-t-elle d'une distribution normale?	Tests de normalité	

Rappel Student

On mesure les masses d'une équipe de bons gros gras mesurées avant et après un régime (draconien). On suppose que les lois sous-jacentes sont **normales**.

Sujet	1	2	3	4	5	6	7	8	9	10
Avant	86	92	75	84	66	75	97	67	99	68
Après	66	76	63	62	74	70	86	69	81	92
Différence	20	16	12	22	-8	5	11	-2	18	-24

On se ramène à une variable différence de Student (différence de deux lois normales divisée par écart-type). On calcule une moyenne de $\overline{D} = 7$ et $\sigma = 14,56$ et le calcul donne $t = \frac{7}{14,56\sqrt{10}} = 1,52$. La valeur critique d'un test de Student à 5% de risques vaut 2,269, donc ...

Rappel Student

On mesure les masses d'une équipe de bons gros gras mesurées avant et après un régime (draconien). On suppose que les lois sous-jacentes sont **normales**.

Sujet	1	2	3	4	5	6	7	8	9	10
Avant	86	92	75	84	66	75	97	67	99	68
Après	66	76	63	62	74	70	86	69	81	92
Différence	20	16	12	22	-8	5	11	-2	18	-24

On se ramène à une variable différence de Student (différence de deux lois normales divisée par écart-type). On calcule une moyenne de $\overline{D} = 7$ et $\sigma = 14,56$ et le calcul donne $t = \frac{7}{14,56\sqrt{10}} = 1,52$. La valeur critique d'un test de Student à 5% de risques vaut 2,269, donc ... on ne rejette pas l'hypothèse d'égalité des deux moyennes.

Tests non paramétrique : test des signes (I)

Mais, que se passe-t-il si :

- les lois initiales ne sont pas normales
- le nombre d'échantillons ne suffit pas pour pouvoir se raccrocher au théorême central limite ?

Le **test des signes** sert à ça : il sert à comparer deux séries de mesures sur une même population (données appariées) mais sans faire d'hypothèses sur la distribution. On compte le nombre de différences positives et négatives entre les paires. Si les moyennes des deux séries de mesures sont égales, on devrait avoir une probabilité équivalente entre les deux configurations (loi binomiale $B(n, \frac{1}{2})$).

Test non paramétrique : test des signes (II)

Sujet	1	2	3	4	5	6	7	8	9	10
Avant	86	92	75	84	66	75	97	67	99	68
Après	66	76	63	62	74	70	86	69	81	92
Différence	**20**	**16**	**12**	**22**	-8	**5**	**11**	-2	**18**	-24

L'hypothèse nulle est que ces tirages peuvent être obtenus par hasard. Puisque l'on a 7 différences positives, on évalue

$$P(B(10, 0.5) < 8) = 0.9453$$

ce qui est acceptable avec $\alpha = 5\%$. On ne peut pas rejeter l'hypothèse nulle, donc on considère qu'il n'y a pas de résultats significatifs malgré les efforts ...

Tests non paramétriques : test de Wilcoxon (I)

Le test de Wilcoxon traite du même problème de façon un peu plus robuste. On classe les différences par ordre de valeurs absolues.

Rang	10	9	8	7	6	5	4	3	2	1
Différence	-24	22	20	18	16	12	11	-8	5	-2

et on calcule la somme des rangs des différences positives, soit $W_+ = 2 + 4 + 5 + 6 + 7 + 8 + 9 = 41$. Ici, on va tester le fait que

les sommes des rangs positifs et des rangs négatifs devraient être équivalentes.

Test de Wilcoxon (II)

Si les différences en valeur absolue sont rangées dans un ordre croissant, chacune d'elle, quelque soit son rang, a une chance sur deux d'être positive : le rang 1 a une chance sur deux de porter le signe +, le rang 2 a une chance sur deux de porter le signe +, etc.

Test de Wilcoxon (III)

La statistique de Wilcoxon est ainsi définie par :

$$W_+ = \sum_{i=1}^{n} r_i Z_i \text{ avec } E(Z_i) = \frac{1}{2} \text{ et } Var(Z_i) = \frac{1}{4}.$$

$$E(W_+/R) = \sum_{i=1}^{n} r_i E(Z_i) = \frac{1}{2}\sum_{i=1}^{n} r_i = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$$

$$Var(W_+/R) = \sum_{i=1}^{n} r_i^2 Var(Z_i) = \frac{1}{4}\sum_{i=1}^{n} r_i^2 = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}$$

On montre que W_+ peut être approximée et donc testée par une loi normale à partir de $n = 10$ (pour certains auteurs $n = 25$?).
Il suffit donc de tester W_+ sur $N(E(W_+/R), Var(W_+/R))$.

Test de Wilcoxon (IV)

Appliqué à l'exemple

Rang	10	9	8	7	6	5	4	3	2	1
Différence	-24	22	20	18	16	12	11	-8	5	-2

$$W_+ = 41$$

$$E(W_+/R) = 27.5$$

$$Var(W_+/R) = 96.25$$

ce qui aboutit à $Z = 0.14$
L'hypothèse nulle peut être conservée avec $\alpha = 5\%$. Pourquoi ?

La foule demande l'exemple en R, alors je m'exécute ...

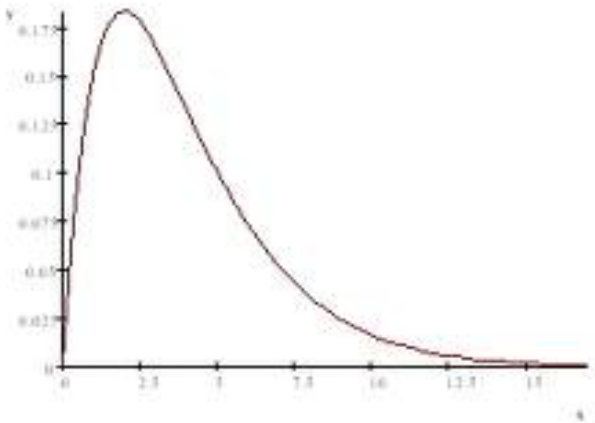
```
> wilcox.test(c(20, 16, 12, 22, -8, 5, 11, -2, 18, -24))

Wilcoxon signed rank test

data: c(20, 16, 12, 22, -8, 5, 11, -2, 18, -24)
V = 41, p-value = 0.1934
alternative hypothesis: true location is not equal to 0
```

Test du χ^2

Rappel :

Definition

si $X_1, X_2, .., X_k$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi $N(0,1)$, alors la loi de $X_1^2 + X_2^2 + .. + X_k^2$ est une loi dite du χ^2 à k degrés de libertés et on la note χ_k^2.



$n = 4$

Utilisation du χ^2

La loi (et le test) du χ^2 est utilisée en présence de variables qualitatives **catégorielles** (loi discrète ou loi continue avec les échantillons regroupés en classes). Elle permet d'effectuer des tests d'hypothèse sur :

- l'égalité de distributions observées (**test homogénéité**) - type de question traitée : la distribution des pointures de chaussures dépend-elle du département considéré ?
- la dépendance entre deux caractères qualitatifs (**test d'indépendance**) - type de question traitée : y a t il une dépendance entre la couleur des yeux et la couleur des dents ?
- la conformité à une distribution connue (**test d'ajustement**) - type de question traitée : les naissances suivent elles une loi équirépartie ?

Rappelez vous ? l'effet de la lune sur les naissances

On souhaite étudier les effets de la lune sur les naissances (plus précisément l'effet supposé de la pleine lune sur l'augmentation des naissances). On relève dans une maternité les données suivantes:

Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif	76	88	100	96	360
Fréquence	0,211	0,244	0,278	0,267	1

Peut-on valider l'hypothèse à partir de ces données ?

Comparaison de la distribution observée avec la distribution équiprobable

- On pose l'hypothèse nulle H_0: les naissances sont équiprobables par rapport aux phases de la lune.
- Ceci peut se traduite par:

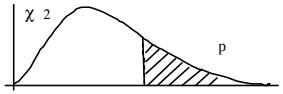
Phase	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Effectif observé	89	88	92	91	360
Effectif théorique	90	90	90	90	360

Valeur du χ^2

- Dans notre cas, on peut "comparer" les distributions à l'aide d'une mesure globale $M = \sum_1^4 (Obs. - Theo)^2 / Theo$
- On suppose les distributions normales, et la mesure M est donc une variable aléatoire de type χ^2_3
- M peut donc être comparée à la valeur de référence (seuil) définie dans la table du χ^2 en fonction du nombre de degrés de libertés des données (ν égal au nombre de classes - 1, soit ici 3) et d'une marge d'erreur classique de 5%.
- Si la mesure est inférieure au seuil, on ne rejette pas l'hypothèse nulle

Table du χ^2

TABLE DU CHI-DEUX : $\chi^2(n)$

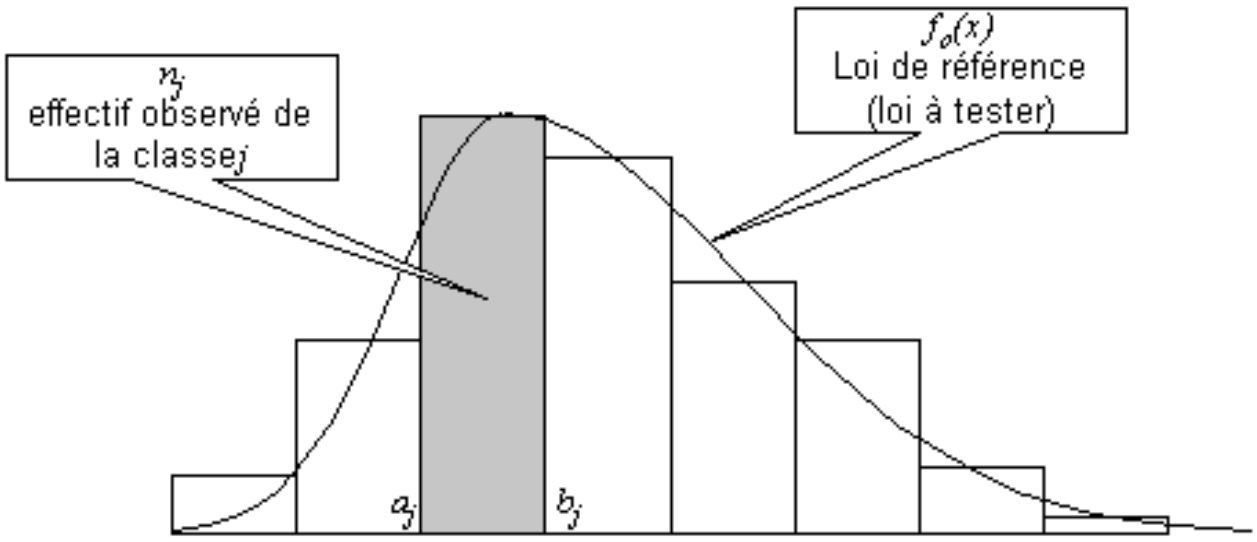


n p	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

Pour n > 30, on peut admettre que $\sqrt{2\chi^2} - \sqrt{2n-1} \approx N(0,1)$

le point critique vaut 7,82 et notre mesure vaut 3,83 et *l'écart est assez petit pour être justifié par le hasard, on conservera donc l'hypothèse nulle*.

Comparaison de distribution empirique et théorique



On ne conserve pas des intervalles trop peu peuplés (<5) et on regroupe donc les classes si nécessaire.

Test d'ajustement simple à une loi de référence

- Données : $x_1, x_2, .., x_n \in \{1, .., k\}$
- Modélisation : observations $X_1, X_2, ..., X_n$ indépendantes et suivant une loi $\mathbf{p}$ sur $\{1, .., k\}$
- Hypothèse H_0 : $\mathbf{p} = \mathbf{p_{ref}}$
- Statistique de test (fréquences) $C = n \sum_{j=1}^{k} \frac{(\widehat{p}_{j,n} - p_j^{ref})^2}{p_j^{ref}}$
- Statistiques de test (effectifs) $C = \sum_{j=1}^{k} \frac{(N_{j,n} - np_j^{ref})^2}{np_j^{ref}}$
- Sous l'hypothèse H_0, la statistique C tend vers une loi du χ^2_{k-1}, sinon elle tend vers l'infini (donc prend des valeurs plus grandes).

Autre exemple

Un agent immobilier souhaite pouvoir embaucher un stagiaire sur la période de printemps, en avançant l'argument que les ventes se font le plus souvent à cette période. Il relève les résultats suivants pour l'année passée :

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Ventes	1	3	4	6	6	5	3	1	2	1	2	2

Qu'en pensez-vous ?

Autre exemple (II)

L'hypothèse nulle H_0 est que les saisons sont équivalentes pour les ventes.
On effectue les regroupements par saison :

saison	hiver	printemps	été	automne
ventes	8	17	6	5
ventes théoriques	9	9	9	9
freq. théoriques	25%	25%	25%	25%
freq. mesurées	22.2%	47.2%	16.7%	13.9%

La réalisation de la variable statistique est égale à 10. On a 3 degrés de libertés, donc la lecture de la table permet de conclure que ...

Un jour, ma soeur a été mordue par un élan ...

```
> chisq.test(c(8,17,6,5))

Chi-squared test for given probabilities

data: c(8, 17, 6, 5)
X-squared = 10, df = 3, p-value = 0.01857
```

Test d'indépendance de couples de données

- Données : couples $(x_1, y_1), .., (x_n, y_n) \in \{1, .., r\} \times \{1, .., s\}$
- Modélisation : couples d'observations $(X_1, Y_1), ..., (X_n, Y_n)$ indépendants et suivant une loi $\mathbf{p}$ sur $\{1, .., r\} \times \{1, .., s\}$
- Hypothèse H_0 : les X_i sont indépendantes des Y_i, donc leur loi produit $\mathbf{p}$ est une loi égale au produit de ses marginales
- Statistique de test (fréquences)
  $C = \sum_{x=1}^{r} \sum_{y=1}^{s} \frac{(N_{x,y} - n\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y))^2}{n\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y)}$
- Sous l'hypothèse H_0, la statistique C tend vers une loi du $\chi^2_{(r-1)(s-1)}$, sinon elle tend vers l'infini (donc prend des valeurs plus grandes)

Quelques explications

On considère les données comme le produit des modalités en x et en y (donc on forme des couples). r et s sont les cardinaux de nos deux ensemble de modalités $\mathbf{X}$ et $\mathbf{Y}$. Les lois marginales correspondent aux probabilités "en colonnes" et "en lignes", donc estimées à :

$$\widehat{p}_{\mathbf{X}} = (\frac{N_{1.}}{n}, ..., \frac{N_{r.}}{n})$$

et

$$\widehat{p}_{\mathbf{Y}} = (\frac{N_{.\mathbf{1}}}{n}, ..., \frac{N_{.\mathbf{s}}}{n})$$

Par ailleurs, $\frac{N_{x,y}}{n}$ et $\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y)$ sont des estimées de $\mathbf{p}(x,y)$ et de $p_{\mathbf{X}}(x)p_{\mathbf{Y}}(y)$. Le test d'indépendance consiste à vérifier que ces deux quantités sont proches.

Indépendance de variables : un petit exemple pour la route

Patrick M. et Gilles C. sont deux enseignants chercheurs du département LUSSI (on a préféré ici conserver l'anonymat). Les bruits courent que Patrick M. note (entre A et F) comme une hyène et que Gilles C. lui note comme un bon bisounours. Les données sont les suivantes :

Notes	A	B	C	D	E	F	Total
Patrick M.	14	15	26	18	17	5	95
Gilles C.	21	18	24	19	15	2	99
Total	35	33	50	37	32	7	194

Prof. Grincheux contre Prof. Simplet

On regroupe les classes E et F (effectifs trop faibles). On obtient :

Notes	A	B	C	D	G	Total
Patrick M.	14	15	26	18	22	95
Gilles C.	21	18	24	19	17	99
Total	35	33	50	37	39	194

Calcul d'effectifs croisés

L'effectif attendu pour le méchant Patrick M. pour la note B est égal à $194 \times \widehat{p}_{\mathbf{X}}(1) \times \widehat{p}_{\mathbf{Y}}(2)$ soit :

$$194 \times \frac{95}{194} \times \frac{33}{194} = 16,2$$

Tableaux d'effectifs croisés

On obtient :

Notes	A	B	C	D	G	Total
Patrick M.	14	15	26	18	22	95
Patrick M. théo	17,1	16,2	24,5	18,1	19,1	95
Gilles C.	21	18	24	19	17	99
Gilles C. théo	17,9	16,8	25,5	18,9	19,9	99
Total	35	33	50	37	39	194
Total	35	33	50	37	39	194

Le dénouement

On effectue le calcul de l'écart entre effectifs théoriques et effectifs observés. Ici,

$$C = \frac{(14 - 17.1)^2}{17.1} + .. + \frac{(17 - 19.9)^2}{19.9} = 2.34$$

La loi de référence est un χ^2 à $(r-1)(s-1) = 4$ degrés de libertés. La p-value obtenue est de 0.674, donc ... Patrick M. n'est pas aussi terrible qu'il en a l'air (et Gilles C. n'est pas aussi bonne poire qu'il ne le dit).

Encore une fois, merci, et bonne chance ...

```
> chisq.test(toto)

Pearson's Chi-squared test

data: toto
X-squared = 4.5683, df = 4, p-value = 0.3345

> toto <- matrix(c(14, 15, 36, 18, 17, 5, 21, 18, 24, 19, 15, 2), nrow =2, byrow = TRUE)
> chisq.test(toto)

Pearson's Chi-squared test

data: toto
X-squared = 5.3386, df = 5, p-value = 0.376

Message d'avis :
In chisq.test(toto) : l'approximation du Chi-2 est peut-être incorrecte
```

Quand ne pas appliquer le test du χ^2 ?

Le test du χ^2, c'est beau mais

- quand il y a seulement deux classes et qu'on veut ajuster une distribution ("si on tombe 212 fois sur un 6 sur 1000 tirages de dé, est-il truqué ?"), il faut appliquer .. un test d'égalité de proportion avec une valeur de référence
- quand on a deux fois deux cases dans le tableau de contingence, il faut appliquer .. un test d'égalité de deux proportions entre elles (sauf quand les échantillons sont appariés !! voir ci-après).

Appliquer un test du χ^2 dans ce cas peut notamment poser des problèmes liés au caractère unilatère du test.

Test de McNemar : comparaison de pourcentage sur un même échantillon

Le test de McNemar concerne l'évaluation de l'évolution d'une proportion dans le temps. Les populations ne sont plus indépendantes, il faut se ramener à un test du χ^2 portant sur les effectifs d'individus ayant changé d'avis entre les deux enquêtes. La statistique de test se ramène à :

$$\frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

à comparer avec un χ^2_1

Un petit exemple

On fait deux sondages successifs T1 et T2 auprès des étudiants de TB sur leur satisfaction de se trouver dans le plus belle école du Mooonde, un premier à leur arrivée à l'école, un second en deuxième année un Jeudi soir après la soirée Mousse. On obtient le résultat suivant :

		T2	T2	
		oui	non	
T1	oui	200	50	250
T1	non	80	270	350
		280	320	600

La proportion de satisfaits est passée de 41.7% à 46.7%. Peu significatif. Mais les échantillons ne sont pas indépendants.

Un petit exemple (ii)

Il faut en fait identifier les changements d'état.

	T2 oui	T2 non	
T1 oui	p_{11}	p_{12}	$p_{1.}$
T1 non	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	

Avec l'hypothèse nulle, $p_{12} = p_{21}$ est estimé par $\frac{p_{12}+p_{21}}{2}$. La statistique est :

$$D = \frac{(n_{12} - \frac{n_{12}+n_{21}}{2})^2 + (n_{21} - \frac{n_{12}+n_{21}}{2})^2}{\frac{n_{12}+n_{21}}{2}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

On a pour $n_{12} + n_{21} \geq 25$ une statistique du χ^2_1

On obtient $\sqrt{D} = 2.63$ ce qui signifie une augmentation sensible de la satisfaction pour un risque de 5%. Au fait, de quel type est la variable $\sqrt{D}$ pour pouvoir avancer ce résultat ?

Tests de normalité : Kolmogorov

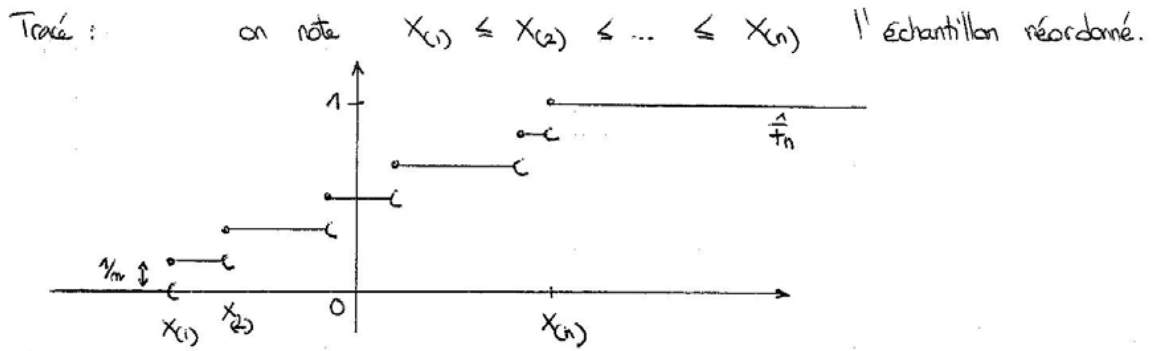
Le test de Kolmogorov-Smirnov consiste à mesurer, pour une variable aléatoire continue, la plus grande distance entre la distribution théorique $F_0(x)$ et la distribution expérimentale $F(x)$. On évalue la fonction de répartition empirique définie par

- 0 pour x plus petit que X_0
- $F(x) = \frac{i}{n}$ pour x compris entre X_i et X_{i+1}
- 1 pour x supérieur à X_n

Test de Kolmogorov (II)

Kolmogorov a proposé la distance entre fonction de répartitions :

$$D_{ks}(F_0, F) = max_{i=1,..,n}\{ \mid F_0(X_i) - \frac{i}{n} \mid, \mid F_0(X_i) - \frac{i-1}{n} \mid \}$$



Test de Kolmogorov (III)

Sous l'hypothèse H_0 (donc de normalité), on sait approximer cette statistique par :

$$lim_{\infty} P[\sqrt{n} D_{ks}(F_0, F) \leq t] = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} exp(-2k^2 t^2)$$

Le calcul de la p-value à partir de cette statistique (qui peut être tabulée) fait le reste.

Test de Kolmogorov (IV)

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35831	.40962	.45427	.50654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466

Tests de normalité : Shapiro-Wilks

On compare les quantiles de la loi observée avec les quantiles générés par une "vraie" loi normale. La corrélation à ces quantiles peut s'écrire :

$$W = \frac{[\Sigma_{i=1}^{[\frac{n}{2}]} a_i (x_{(n-i+1)} - x_{(i)})]^2}{\Sigma_i (x_i - \bar{x})^2}$$

où

- $x_{(i)}$ sont les données en rang
- les a_i sont des constantes générées à partir de la moyenne et de la matrice de covariance des quantiles d'un échantillon de taille n suivant une loi normale

La loi W est tabulée et on décide de la normalité d'un échantillon si la réalisation de W dépasse la valeur critique W_{crit} trouvée dans la table.

Tests de normalité : Shapiro-Wilks (II)

Table 4b : table des valeurs limites W_α de $W = \frac{b^2}{Z^2}$

pour les risques α = 5 % et 1 %

(Biometrika 1965)

n	Risque 5 %	Risque 1 %
	$W_{0,05}$	$W_{0,01}$
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805
13	0,866	0,814
14	0,874	0,825
15	0,881	0,835
16	0,887	0,844

Alors l'éléphant met un pied dans l'eau ...

```
> shapiro.test(rnorm(1000))

Shapiro-Wilk normality test

data: rnorm(1000)
W = 0.9984, p-value = 0.4822

> shapiro.test(rnorm(1000))

Shapiro-Wilk normality test

data: rnorm(1000)
W = 0.9957, p-value = 0.00642
```

Coefficient de corrélation de Pearson

La formule classique de ce coefficient est:

$$r = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

Ce coefficient mesure la corrélation linéaire sur des variables numériques.

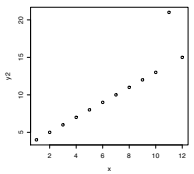
r est très sensible aux points extrêmes et en ce sens n'est pas très robuste. La relation de corrélation n'est pas transitive.

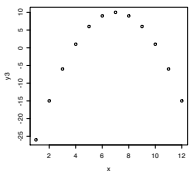
Coefficient de corrélation (II)

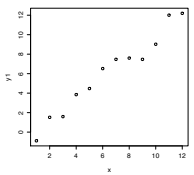
- r est toujours compris entre -1et 1
- 1 et -1 dénotent une corrélation parfaite entre x et y
- si x et y sont indépendantes, alors $r = 0$ mais l'inverse n'est pas vraie (mais la dépendance n'est alors pas linéaire)

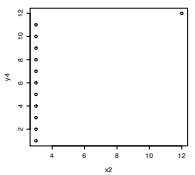
Coefficient de corrélation (II)

Les figures suivantes correspondent toutes à des nuages de même moyenne, même variance et ... même coefficient de corrélation $r = 0.82$. Pour quelle figure le coefficient est-il vraiment significatif ?









Validité de la corrélation

On montre que

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

suit une loi de Student à $(n-2)$ degrés de liberté. De façon pratique, on rejette l'hypothèse d'indépendance avec un risque 5% quand T est à l'extérieur de l'intervalle -2, 2.
Dans le cas de l'exemple, on calcule

$$r = 0.87$$

$$T = 11.02$$

et donc on ne peut pas attribuer la dépendance au hasard.

Validité de la régression

Avec un échantillon de taille 30, peut on déclarer que deux variables sont réellement indépendantes avec:

- $r = 0.1$ -> $T = 0.53$
- $r = 0.2$ -> $T = 1.08$
- $r = 0.3$ -> $T = 1.66$
- $r = 0.4$ -> $T = 2.31$
- $r = -0.2$ -> $T = -1.08$
- $r = -0.5$ -> $T = -3.06$

Coefficient de Spearman

Il est courant de ne disposer que d'un ordre sur les individus et non de variables numériques (ordre de classement, préférences, mesures non directement utilisables sur une échelle, etc..). On affecte un rang à chaque individu.

Objet	1	2	....	n
Rang 1	r_1	r_2		r_n
Rang 2	s_1	s_2		s_n

Coefficient de Spearman (II)

Le coefficient de Spearman est défini par:

$$r_s = \frac{cov(r,s)}{s_r s_s}$$

Comme les rangs sont des permutations de 1 à n, on sait que $\bar{r} = \bar{s} = \frac{n+1}{2}$. Après quelques premiers calculs, on obtient:

$$r_s = \frac{\frac{1}{n}\sum_i r_i s_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}}$$

soit

Coefficient de Spearman (II)

Le coefficient de Spearman est défini par:

$$r_s = \frac{cov(r,s)}{s_r s_s}$$

Comme les rangs sont des permutations de 1 à n, on sait que $\bar{r} = \bar{s} = \frac{n+1}{2}$. Après quelques premiers calculs, on obtient:

$$r_s = \frac{\frac{1}{n}\sum_i r_i s_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}}$$

soit

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

avec $d_i = r_i - s_i$

Coefficient de Spearman (III)

Une autre expression du coefficient est :

$$r_s = 12\left(\frac{\Sigma r_i s_i}{n^3 - n} - \frac{n+1}{4(n-1)}\right)$$

Coefficient de Spearman (IV)

Lorsque:

- $r_s = 1$, les deux classements sont identiques
- $r_s = -1$, les deux classements sont inverses l'un de l'autre
- $r_s = 0$, les deux classements sont indépendants

Coefficient de Spearman (V)

Neuf étudiants ont subi (c'est le mot, les pauvres) deux examens de statistiques et d'aide à la décision. Les résultats sont les suivants :

Stats	50	23	28	34	14	54	46	52	53
Décision	38	28	14	26	18	40	23	30	27

A-t-on corrélation entre les examens ?

Coefficient de Spearman (VI)

On calcule le tableau des rangs

Stats	6	2	3	4	1	9	5	7	8
Décision	8	6	1	4	2	9	3	7	5

et on calcule $\Sigma r_i s_i = 6 \times 8 + .. + 8 \times 5 = 266$, et

$$r_s = 12(\frac{266}{9^3 - 9} - \frac{10}{32}) = 0.6833$$

La valeur critique est de 0.683, on rejette tout juste l'indépendance.

Coefficient de corrélation des rangs τ de Kendall

Pour savoir si deux variables théoriques varient dans le même sens, on considère le signe de $(X_1 - X_2)(Y_1 - Y_2)$ avec (X_1, Y_1) et (X_2, Y_2) deux réalisations indépendantes de (X, Y). On définit le coefficient théorique τ par:

$$\tau = 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1$$

Ce coefficient est également compris entre -1 et 1 et s'annule quand les variables sont indépendantes.
On montre que si X et Y sont gaussiennes de coefficient de corrélation ρ, alors $\tau = \frac{2}{\pi} Arc(sin(\rho))$ (rq: $\tau \leq \rho$).

Concrètement ..

On note les concordances et les discordances des variables X et Y (soit 1 si $x_i < x_j$ et $y_i < y_j$, -1 sinon). On somme sur S les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples distincts, donc $S_{max} = \frac{n(n-1)}{2}$.On aura:

$$\tau = \frac{2S}{n(n-1)}$$

Si $\tau = 1$ les classements sont identiques, si $\tau = -1$ les classements sont inversés.

Encore plus concrètement ...

- on ordonne les x_i de 1 à n.
- on compte pour chaque x_i le nombre de $y_j > y_i$ pour les $j > i$ ce qui donne R
- $S = 2R - \frac{n(n-1)}{2}$ et
- $\tau = \frac{4R}{n(n-1)} - 1$

Exemple

On a les classements suivants:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	1	4	2	6	5	9	8	10	7

Le coefficient de Spearman vaut:

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 0.84$$

Le coefficient de Kendall se calcule par:

$$R = 7 + 8 + 6 + 6 + 4 + 4 + 1 + 1 = 37$$

$$S = 74 - 45 = 29$$

donc $\tau = 0.64$

Quelle validité pour les coefficients?

On peut tester les deux coefficients à partir:

- d'une table de validité du coefficient de Spearman (établie à partir de l'hypothèse de permutations équiprobables dès lors que les variables seraient indépendantes). La table est indexée en α et en n.
- de l'approximation $\tau \simeq N(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$ dès que $n > 8$.

Pour notre exemple...

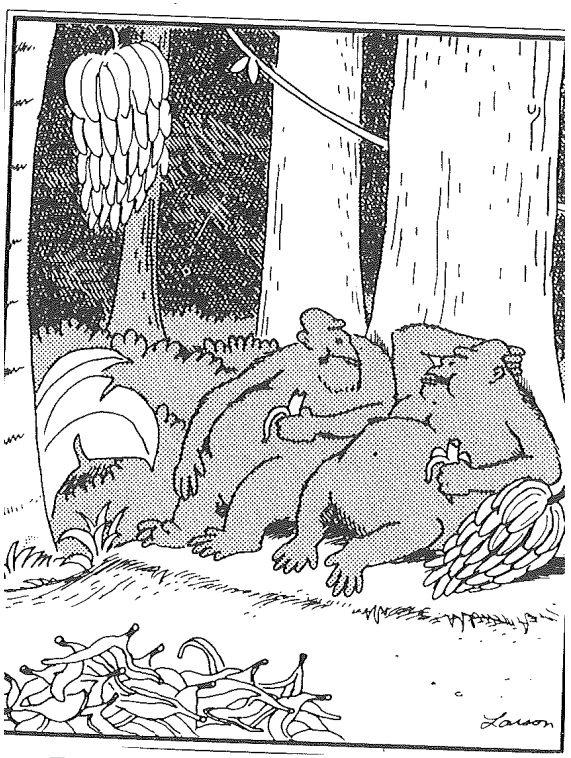
- Pour Spearman, on obtient dans la table $r_{s,critique} = \pm 0.648$
- Pour Kendall, $\tau_{critique} = \pm 1.96\sqrt{\frac{50}{90.9}} = \pm 0.49$

On a donc une liaison significative entre les classements puisque les valeurs réalisées sont supérieures au seuil et qu'on peut rejeter l'hypothèse nulle d'indépendance.

et hop

```
> x <- c(50, 23, 28, 34, 14, 54, 46, 52, 53)
> y <- c(38, 28, 14, 26, 18, 40, 23, 30, 27)
> cor(x,y, method = "pearson")
0.6794456
> cor(x,y, method = "spearman")
0.6833333
> cor(x,y, method = "kendall")
0.5
```

Bon appétit !



— Tu sais quoi, Zog, j'aime *vraiment* les bananes.