Scientific Method

Class 4

MASTER INFORMATIQUE - SIIA

G. Coppin

2021-2022

Class objectives

Objs:

1. End with the tests
2. Play with dependency and independence of variables

Small reminders

The statistical hypothesis tests boil down to :

1. model the problem from a statistical point of view
2. define a null hypothesis H_0 with respect to the problem to be treated
3. choose a statistical test or - which amounts to the same thing - a statistical (variable) for the test : this is a random variable which must allow you to choose between H_0 and H_1
4. define the distribution of the statistical variable for H_0
5. define the level of significance of the test or risk
6. calculate from the sample the statistical variable
7. make a decision from the positioning of the value (threshold associated with the risk) or from the p-value obtained

Small reminders - questions (ii)

- What is the relationship between risk and p-value (or more precisely how do you use one or the other?)
- What does the central limit theorem allow us to do?
- How to estimate a confidence interval around an estimated value?

Confidence interval calculations and tests are not necessarily neutral (I)

A study carried out at the end of 2008 on 298 Parisian dwellings chosen at random from the directory ensures that the price of rent per square meter is 18.4 euros, with a measured standard deviation of 3.2 euros.

- briefly model the situation
- we assume that the survey is requested by the Paris Rent Observatory. How would you express the confidence interval for the average rental price per square meter in Paris.
- we assume that the survey is requested by the Black Thursday Collective. Same question.
- it is assumed that the survey is requested by the National Confederation of Landlords. Same question.

Confidence interval calculations and tests are not necessarily neutral (II)

- Population =?
- Data / variables =? Independent?
- Parameter of interest (estimator) =?
- Confidence interval =?

Parametric and non-parametric tests

- Parametric : we make assumptions about the law underlying the variables (and we adjust the parameters of this law). Ex : Student's test on a mean of normal distributions $X_1, X_2, .., X_k$
- Nonparametric : no hypothesis on the nature of the 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	

Reminder Student

We measure the masses of a team of good fat measured before and after a diet (draconian). We assume that the underlying laws are **normal**.

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

We come back to a Student difference variable (difference of two normal laws divided by standard deviation). We calculate an average of $\overline{D} = 7$ and $\sigma = 14.56$ and the calculation gives $t = \frac{7}{14.56\sqrt{10}} = 1.52$. The critical value of a Student's test at 5% risk is 2.269, so ...

Reminder Student

We measure the masses of a team of good fat measured before and after a diet (draconian). We assume that the underlying laws are **normal**.

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

We come back to a Student difference variable (difference of two normal laws divided by standard deviation). We calculate an average of $\overline{D} = 7$ and $\sigma = 14.56$ and the calculation gives $t = \frac{7}{14.56\sqrt{10}} = 1.52$. The critical value of a Student's test at 5 % risk is 2.269, so ... we do not reject the hypothesis of equality of the two means.

Non parametric tests: sign test (I)

But, what if :

- the initial laws are not normal
- the number of samples is not enough to be able to stick to the central limit theorem?

The **sign test** is used for that : it is used to compare two series of measurements on the same population (paired data) but without making assumptions about the distribution. We count the number of positive and negative differences between the pairs. If the means of the two series of measurements are equal, we should have an equivalent probability between the two configurations (binomial distribution $B(n, \frac{1}{2})$).

Non parametric test : sign test (II)

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

The null hypothesis is that these prints can be obtained by chance. Since we have 7 positive differences, we evaluate

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

which is acceptable with $\alpha = 5\%$. We cannot reject the null hypothesis, so we consider that there are no significant results despite the efforts ...

Non parametric tests: Wilcoxon test (I)

Wilcoxon's test deals with the same problem in a somewhat more robust fashion. The differences are classified in order of absolute values.

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

and we calculate the sum of the ranks of the positive differences, i.e. $W_+ = 2+4+5+6+7+8+9 = 41$. Here, we will test the fact that

the sums of positive and negative ranks should be equal.

Wilcoxon test (II)

If the differences in absolute value are arranged in ascending order, each of them, whatever its rank, has a one in two chance of being positive : rank 1 has a one in two chance of carrying the + sign, the rank 2 has a 50/50 chance of carrying the + sign, etc.

Wilcoxon test (III)

The Wilcoxon statistic is thus defined by :

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

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

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

We show that W_+ can be approximated and therefore tested by a normal distribution from $n = 10$ (for some authors $n = 25$?).
So we just have to test W_+ on $N(E(W_+/R), Var(W_+/R))$.

Wilcoxon test (IV)

Applied to the example

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

$$W_+ = 41$$

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

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

which results in $Z = 0.14$
The null hypothesis can be kept with $\alpha = 5\%$. Why ?

The crowd asks for the example in R, so I do ...

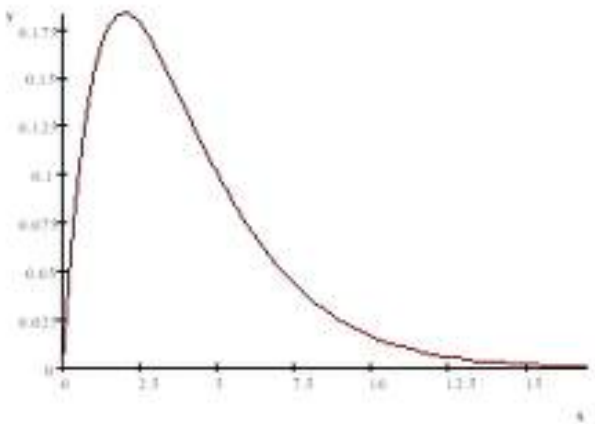
```
> wilcox.test (c (20, 16, 12, 22, -8, 5, 11, -2, 18, -24))

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 of χ^2

Reminder :

Definition

if $X_1, X_2, .., X_k$ are independent and identically distributed random variables according to a distribution $N(0,1)$, then the distribution of $X_1^2 + X_2^2 + .. + X_k^2$ is a so-called χ^2 law with k degrees of freedom and we denote it χ_k^2.



$n = 4$

Using the χ^2

The χ^2 law (and test) is used in the presence of qualitative **categorial** variables (discrete law or continuous law with the samples grouped into classes). It allows you to perform hypothesis tests on :

- equality of observed distributions (**homogeneity test**) - type of question addressed : does the shoe size distribution depend on the department considered?
- the dependence between two characters qualitative (**independence test**) - type of question answered : is there a dependence between the color of the eyes and the color of the teeth?
- conformity to a known distribution (**goodness-of-fit test**) - type of question addressed : do births follow an even-distributed law?

Remember? the effect of the moon on births

We want to study the effects of the moon on births (more precisely the supposed effect of the full moon on the increase in births). We note in a maternity the following data:

Phase	New moon	First quarter	Full moon	Last quarter	Total
Workforce	76	88	100	96	360
Frequency	0.211	0.244	0.278	0.267	1

Can we validate the hypothesis from these data?

Comparison of the observed distribution with the equiprobable distribution

- We set the null hypothesis H_0: births are equiprobable with respect to the phases of the moon.
- This can result in:

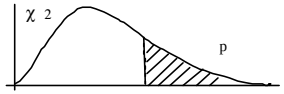
Phase	New moon	First quarter	Full moon	Last quarter	Total
Number observed	89	88	92	91	360
Theoretical workforce	90	90	90	90	360

Value of χ^2

- In our case, we can "compare" the distributions using a global measure $M = \sum_1^4 (Obs. - Theo)^2 / Theo$
- We assume normal distributions, and the measure M is therefore a random variable of type χ^2_3
- M can therefore be compared to the reference value (threshold) defined in the table of χ^2 according to the number of degrees of freedom of the data (ν equal to the number of classes - 1 , or here 3 and a classic margin of error of 5%.
- If the measure is less than the threshold, we do not reject the null hypothesis

Table of χ^2

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

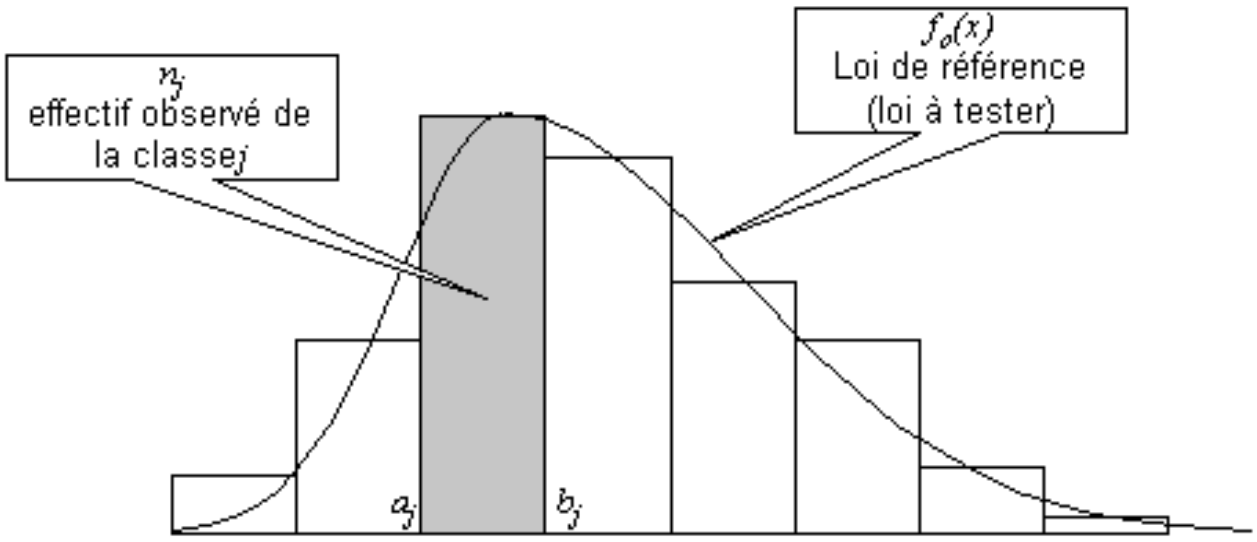


n \ p	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	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 $\sqrt{2\chi^2} - \sqrt{2n-1} \approx N(0,1)$

the critical point is 7.82 and our measurement is 3.83 and *the difference is small enough to be justified by chance, so we will keep the null hypothesis*.

Empirical and theoretical distribution comparison



We do not keep intervals that are too sparsely populated (<5) and we therefore regroup the classes if necessary.

Simple test of fit to a reference distribution

- Data : $x_1, x_2, .., x_n \in \{1, .., k\}$
- Modeling : observations $X_1, X_2, ..., X_n$ independent and following a law $\mathbf{p}$ on $\{1, .., k\}$
- Hypothesis H_0 : $\mathbf{p} = \mathbf{p_{ref}}$
- Test statistic (frequencies) $C = n \sum_{j=1}^{k} \frac{(\widehat{p}_{j,n} - p_j^{ref})^2}{p_j^{ref}}$
- Test statistics (effective) $C = \sum_{j=1}^{k} \frac{(N_{j,n} - np_j^{ref})^2}{np_j^{ref}}$
- Under the hypothesis H_0, the statistic C tends to a distribution of χ^2_{k-1}, otherwise it tends to infinity (therefore takes larger values).

Another example

A real estate agent wants to be able to hire an intern during the spring period, arguing that sales are most often made during this period. He noted the following results for the past year :

Month	J	F	M	A	M	J	J	A	S	O	N	D
Sales	1	3	4	6	6	5	3	1	2	1	2	2

What do you think?

Another example (II)

The null hypothesis H_0 is that the seasons are equivalent for sales. We carry out the groupings by season :

season	winter	spring	summer	autumn
sales	8	17	6	5
theoretical sales	9	9	9	9
freq. theoretical	25 %	25 %	25 %	25 %
freq. measured	22.2 %	47.2 %	16.7 %	13.9 %

The realization of the statistical variable is equal to 10. We have 3 degrees of freedom, so reading the table leads to the conclusion that ...

One day my sister was bitten by a moose ...

```
> 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
```

Data pair independence test

- Data : couples $(x_1, y_1), .., (x_n, y_n) \in \{1, .., r\} \times \{1, .., s\}$
- Modeling : pairs of observations $(X_1, Y_1), ..., (X_n, Y_n)$ independent and according to a law $\mathbf{p}$ on $\{1, .., r\} \times \{1, .., s\}$
- Hypothesis H_0 : the X_i are independent of the Y_i, so their product law $\mathbf{p}$ is a law equal to the product of its marginals
- Test statistic (frequencies) $C = \sum_{x=1}^{r} \sum_{y=1}^{s} \frac{(N_{x,y} - n\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y))^2}{n\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y)}$
- Under the hypothesis H_0, the statistic C tends to a distribution of $\chi^2_{(r-1)(s-1)}$, otherwise it tends to infinity (therefore takes larger values)

Some explanations

We consider the data as the product of the modalities in x and in y (so we form pairs). r and s are the cardinals of our two sets of modalities $\mathbf{X}$ and $\mathbf{Y}$. The marginal laws correspond to the probabilities "in columns" and "in rows", therefore estimated at :

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

and

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

In addition, $\frac{N_{x,y}}{n}$ and $\widehat{p}_{\mathbf{X}}(x)\widehat{p}_{\mathbf{Y}}(y)$ are estimates of $\mathbf{p}(x,y)$ and of $p_{\mathbf{X}}(x)p_{\mathbf{Y}}(y)$. The independence test consists in verifying that these two quantities are close.

Variable independence : a small example for the route

Jeremy R.. and Gilles C. are two research professors from SIIA Master (we preferred to keep anonymity here). There are rumors that Jeremy R. gives bad marks (between A and F) and that Gilles C. behaves as a good egg. The data is as follows :

Notes	A	B	C	D	E	F	Total
Jeremy R.	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. Grumpy vs. Prof. Dumb

We group together the classes E and F (too small numbers). We get :

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

Cross-counting calculation

The expected headcount for the villainous Jeremy R. for the B grade is equal to $194 \times \widehat{p}_{\mathbf{X}}(1) \times \widehat{p}_{\mathbf{Y}}(2)$ or :

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

Cross-sectional count tables

We get :

Notes	A	B	C	D	G	Total
Jeremy R.	14	15	26	18	22	95
Jeremy R. 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

The outcome

The difference between theoretical and observed numbers is calculated. Here,

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

The reference law is a χ^2 at $(r-1)(s-1) = 4$ degrees of freedom. The p-value obtained is 0.674, so ... Jeremy R. is not as bad as he looks (and Gilles C. is not a good egg as he says).

Once again, thank you, and good luck ...

```
> 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) : the Chi-square approximation may be incorrect
```

Normality tests : Kolmogorov

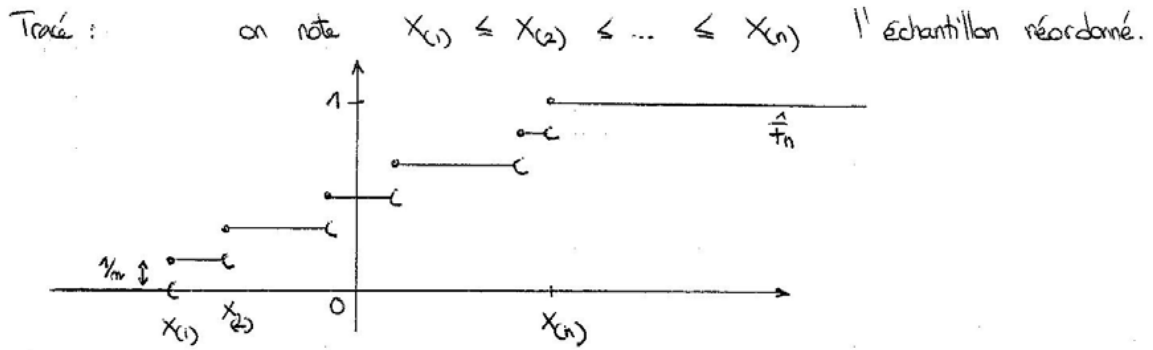
The Kolmogorov-Smirnov test consists in measuring, for a continuous random variable, the greatest distance between the theoretical distribution $F_0(x)$ and the experimental distribution $F(x)$. We evaluate the empirical distribution function defined by

- 0 for x less than X_0
- $F(x) = \frac{i}{n}$ for x between X_i and X_{i+1}
- 1 for x greater than X_n

Kolmogorov test (II)

Kolmogorov proposed the distance between distribution functions :

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



Kolmogorov test (III)

Under the hypothesis H_0 (therefore of normality), we know how to approximate this statistic by :

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

Calculating the p-value from this statistic (which can be tabulated) does the rest.

Kolmogorov test (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

Normality tests : Shapiro-Wilks

We compare the quantiles of the observed law with the quantiles generated by a "true" normal law. The correlation to these quantiles can be written:

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

or

- $x_{(i)}$ are row data
- the a_i are constants generated from the mean and the covariance matrix of the quantiles of a sample of size n following a normal distribution

The distribution W is tabulated and the normality of a sample is decided if the realization of W exceeds the critical value W_{crit} found in the table.

Normality tests : 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

So the elephant puts one foot in the water ...

```
> 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
```

Pearson correlation coefficient

The classic formula for this coefficient is:

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

This coefficient measures the linear correlation on numeric variables.

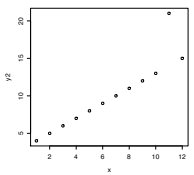
r is very sensitive to extreme points and in this sense is not very robust. The correlation relation is not transitive.

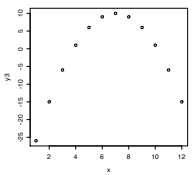
Correlation coefficient (II)

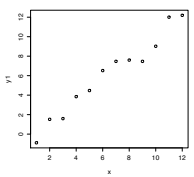
- r is always between -1 and 1
- 1 and -1 denote a perfect correlation between x and y
- if x and y are independent, then $r = 0$ but the reverse is not true (but the dependency is then not linear)

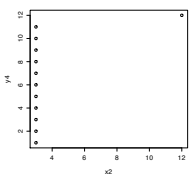
Correlation coefficient (II)

The following figures all correspond to clouds with the same mean, same variance and ... same correlation coefficient $r = 0.82$. For which figure is the coefficient really significant?









Correlation validity

We show that

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

follows a Student's law with $(n - 2)$ degrees of freedom. In a practical way, we reject the independence hypothesis with a risk of 5 % when T is outside the interval -2, 2.
In the case of the example, we calculate

$$r = 0.87$$

$$T = 11.02$$

and therefore one cannot attribute the dependence to chance.

Regression validity

With a sample of size 30, we can declare that two variables are really independent with:

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

Spearman coefficient

It is common to have only an order on individuals and not numerical variables (order of classification, preferences, measures not directly usable on a scale, etc.). We assign a rank to each individual.

Subject	1	2	....	n
Row 1	r_1	r_2		r_n
Row 2	s_1	s_2		s_n

Spearman coefficient (II)

The Spearman coefficient is defined by:

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

As the ranks are permutations from 1 to n, we know that $\bar{r} = \bar{s} = \frac{n+1}{2}$.
After a few initial calculations, we obtain:

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

either

Spearman coefficient (II)

The Spearman coefficient is defined by:

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

As the ranks are permutations from 1 to n, we know that $\bar{r} = \bar{s} = \frac{n+1}{2}$.
After a few initial calculations, we obtain:

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

either

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

with $d_i = r_i - s_i$

Spearman coefficient (III)

Another expression for the coefficient is :

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

Spearman coefficient (IV)

When:

- $r_s = 1$, the two rankings are identical
- $r_s = -1$, the two rankings are opposite of each other
- $r_s = 0$, the two rankings are independent

Spearman coefficient (V)

Nine students underwent (that's the word, the poor) two statistical and decision support exams. The results are as follows:

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

Is there a correlation between the exams?

Spearman coefficient (VI)

We calculate the row table

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

and we calculate $\Sigma r_i s_i = 6 \times 8 + .. + 8 \times 5 = 266$, and

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

The critical value is 0.683, we just reject independence.

Kendall's τ rank correlation coefficient

To know if two theoretical variables vary in the same direction, we consider the sign of $(X_1 - X_2)(Y_1 - Y_2)$ with (X_1, Y_1) and (X_2, Y_2) two independent realizations of (X, Y). We define the theoretical coefficient τ by:

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

This coefficient is also between -1 and 1 and vanishes when the variables are independent.
We show that if X and Y are Gaussians with correlation coefficient ρ, then $\tau = \frac{2}{\pi} Arc(sin(\rho))$ (rq: $\tau \leq \rho$).

Concretely ..

We note the concordances and the discrepancies of the variables X and Y (ie 1 if $x_i < x_j$ and $y_i < y_j$, -1 otherwise). We sum over S the values ??obtained for the $\frac{n(n-1)}{2}$ distinct pairs, so $S_{max} = \frac{n(n-1)}{2}$ .We'll have:

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

If $\tau = 1$ the classifications are identical, if $\tau = -1$ the classifications are reversed.

Even more concretely ...

- we order the x_i from 1 to n.
- we count for each x_i the number of $y_j > y_i$ for the $j > i$ which gives R
- $S = 2R - \frac{n(n-1)}{2}$ and
- $\tau = \frac{4R}{n(n-1)} - 1$

Example

We have the following classifications:

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

The coefficient of Spearman is worth:

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

The Kendall coefficient is calculated by:

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

$$S = 74 - 45 = 29$$

so $\tau = 0.64$

Which validity for the coefficients?

We can test the two coefficients from:

- of a validity table of the Spearman coefficient (established from the assumption of equiprobable permutations when the variables are independent). The table is indexed in α and in n.
- of the approximation $\tau \simeq N(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$ as soon as $n > 8$.

For our example...

- For Spearman, we get in the table $r_{s,critical} = \pm 0.648$
- For Kendall, $\tau_{critical} = \pm 1.96\sqrt{\frac{50}{90.9}} = \pm 0.49$

We therefore have a significant link between the classifications since the values achieved are greater than the threshold and we can reject the null hypothesis of independence.

and here it is

```
> 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
```

Democracy is important