Séance 3 : Introduction aux tests d'hypothèse

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

Introduction

Principe de la statistique inférentielle fréquentiste

Inférence = généralisation des résultats obtenus dans l'échantillon à la population dans son ensemble.

Dans la mesure où un échantillon, même aléatoire, est une représentation *imparfaite* de la population, à quelle condition pouvons-nous affirmer que les liaisons constatées entre deux variables de l'échantillon correspondent à des liaisons véritables dans la population ?

La statistique classique (qualifiée de "fréquentiste" par ses critiques) s'appuie pour l'inférence sur une technique particulière, le test de significativité de l'hypothèse nulle (null hypothesis significance testing, NHST).

Le test d'hypothèse

Principes du NHST

Les **tests d'hypothèse** (comme le test du χ^2 ou le test de Student) mesurent la vraisemblance d'une distribution statistique observée par rapport à un état fictif, l'hypothèse nulle.

Ils suivent les étapes de raisonnement suivantes :

1. Déterminer l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1).
2. Fixer un **seuil de significativité** (α).
3. Déterminer la probabilité de la distribution observée sous H_0 (p).
4. Comparer cette probabilité au seuil de significativité :
   a. Si $p > \alpha$, conserver H_0.
   b. Si $p \leqslant \alpha$, rejeter H_0.

Hypothèse nulle et hypothèse alternative

Nous testons ici la liaison entre deux variables dans un échantillon. Dans de tels tests, on considère toujours que :

- H_0 = il n'existe aucun lien entre les deux variables testées (elles sont indépendantes).
- H_1 = il existe un lien entre les deux variables testées (elles sont dépendantes l'une de l'autre).

Nous devons donc partir du principe que, jusqu'à ce que nous ayons de forts indices du contraire, les variables de notre échantillon sont indépendantes les unes des autres. On cherche à "démontrer" par le test que ça n'est pas le cas (en réalité, il ne s'agit pas d'une démonstration, car il existe toujours une marge d'erreur possible).

On considère habituellement le seuil $\alpha = 0.05 = 5\%$ comme suffisant. On rejettera donc l'hypothèse nulle si la distribution observée a moins de 5% de chance d'être observée si les variables sont aléatoires.

Erreurs

On distingue :

- l'**erreur de type I** : on rejette H_0 à tort (*faux positif*)
- l'**erreur de type II** : on ne rejette pas H_0 à tort (*faux négatif*)

La sélection d'un seuil de significativité (α) joue sur la probabilité de chaque type d'erreur.

- Plus on choisit un α élevé et plus on s'expose à des erreurs de type I.
- Plus on choisit un α faible et plus on s'expose à des erreurs de type II.

Significativité

Les tests d'hypothèses permettent de déterminer la **significativité** d'un résultat statistique. La grande majorité des techniques d'analyse statistique mobilisent une telle mesure. Elle désigne la vraisemblance d'un résultat statistique.

La significativité est souvent présentée sous la forme d'une **p-value**, qui peut être interprétée, dans un premier temps, comme la probabilité qu'un résultat statistique soit du au hasard.

Association entre deux variables catégorielles : le test du χ^2

Schéma général du test de χ^2

- H_0 : indépendance des deux variables
- H_1 : les deux variables sont dépendantes.
- le calcul de la distance entre distribution observé et distribution théorique se fait par la mesure de la **distance du χ^2**
- on compare cette distance à la **distribution de la loi du χ^2**
- on conclut à l'indépendance ou au lien entre les variables.

Plus belle la vie : Une écoute genrée ?

Suivez-vous régulièrement ou avez-vous suivi régulièrement à un moment de votre vie une série ou un feuilleton diffusé à la télévision (3 réponses possibles) ? (PCF 2008)

TAB. 1 : Effectifs observés

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	2185	209	2394
Femme	2233	377	2610
Sum	4418	586	5004

Plus belle la vie : Une écoute genrée ?

TAB. 2 : Fréquences observées

	Ne cite pas PBLV	Cite PBLV	Total
Homme	91.27	8.73	100
Femme	85.56	14.44	100
All	88.29	11.71	100

- H_0 : Sexe et écoute de PBLV sont indépendants.
- H_1 : Sexe et écoute de PBLV sont liés.

Situation d'indépendance

Que se passerait-il si les variables étaient indépendantes, comme on en fait l'hypothèse (H_0) ?

Dans ce cas, les modalités seraient équitablement réparties : la proportion d'hommes regardant PBLV serait égale à la proportion de femmes regardant PBLV et donc à la proportion moyenne de personnes regardant PBLV.

Pour décrire cette situation d'indépendance, on calcule les **effectifs théoriques** : la distribution parfaitement indépendante des variables.

Plus belle la vie : Une écoute genrée ?

TAB. 3 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme			2394
Femme			2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

TAB. 4 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	$\frac{4418*2394}{5004}$		2394
Femme			2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

TAB. 5 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	2114		2394
Femme			2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

TAB. 6 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	2114	$\frac{586*2394}{5004}$	2394
Femme			2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

TAB. 7 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	2114	280	2394
Femme			2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

TAB. 8 : Effectifs théoriques

	Ne cite pas PBLV	Cite PBLV	Sum
Homme	2114	280	2394
Femme	2304	306	2610
Sum	4418	586	5004

$$T_{ij} = \frac{O_{.j} \times O_{i.}}{N}$$

Plus belle la vie : Une écoute genrée ?

Effectifs observés et théoriques

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
Total		4418	586	5004

Plus belle la vie : Une écoute genrée ?

Écarts entre effectifs observés et théoriques

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
	$O - T$	71	-71	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
	$O - T$	-71	71	
Total		4418	586	5004

Distance du chi-2

La **distance du chi-2** donne son nom au **test du chi-2**. Méthode pour mesurer la différence entre deux tableaux de contingence.

Distance du chi-2 entre deux tableaux =

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{i,j} - T_{i,j})^2}{T_{i,j}}$$

Plus belle la vie : Une écoute genrée ?

Écarts entre effectifs observés et théoriques

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
	$O-T$	71	-71	
	$(O-T)^2$	5041	5041	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
	$O-T$	-71	71	
	$(O-T)^2$	5041	5041	
Total		4418	586	5004

Plus belle la vie : Une écoute genrée ?

Écarts et χ^2 par case

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
	$O-T$	71	-71	
	χ^2	$\frac{71^2}{2114} = 2.4$	$\frac{-71^2}{280} = 18,2$	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
	$O-T$	-71	71	
	χ^2	$\frac{-71^2}{2304} = 2,2$	$\frac{71^2}{306} = 16,7$	
Total		4418	586	5004

$$\chi^2_{ij} = \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

$$\chi^2 = \sum_{i,j} \chi^2_{ij} = 39,4$$

Plus belle la vie : Une écoute genrée ?

Écarts et χ^2 par case

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
	$O-T$	71	-71	
	χ^2	2,4	18,2	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
	$O-T$	-71	71	
	χ^2	2,2	16,7	
Total		4418	586	5004

$$\chi^2_{ij} = \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \qquad \chi^2 = \sum_{i,j} \chi^2_{ij} = 39,4$$

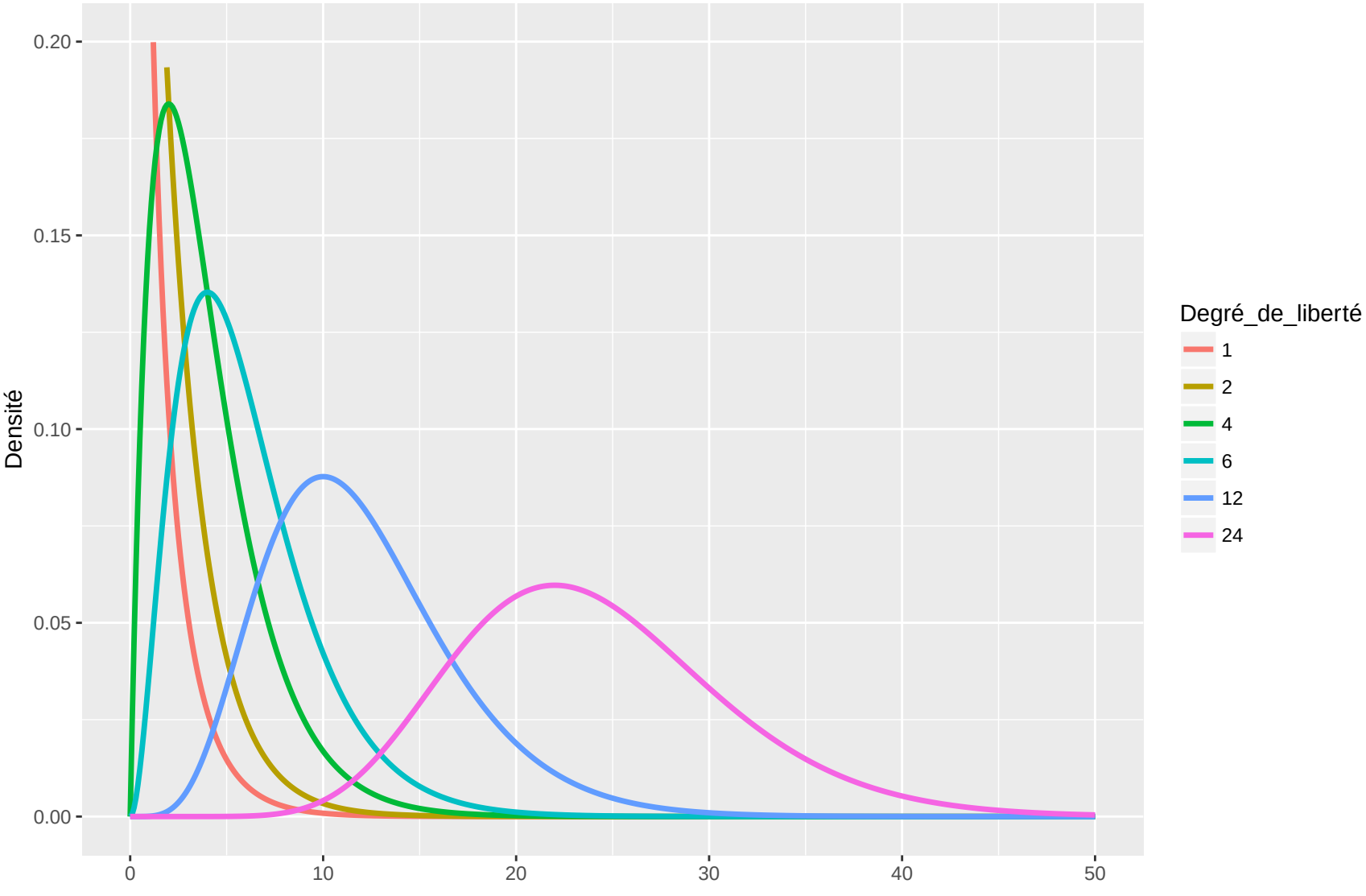
La loi du χ^2

La loi du χ^2 est une loi de probabilité. Soit k variables aléatoires (suivant une loi normale), indépendantes entre elles, de moyenne μ_i et d'écart-type σ_i. La variable Y telle que :

$$Y = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

suit la loi du χ^2. Cette loi a donc un seul paramètre, k, appelé **degré de liberté**. La densité de la loi du χ^2 suit la courbe suivante :

La loi du χ^2



La loi du χ^2

Si deux variables sont indépendantes, la distance du χ^2 qui sépare la distribution observée de la distribution théorique suit la loi du χ^2. La loi du χ^2 permet donc d'estimer la vraisemblance de cette hypothèse

- Si la distance du χ^2 est une valeur habituellement prise par la loi du χ^2, alors on peut penser que les deux variables sont bien indépendantes : on conserve H_0.
- Si la distance du χ^2 est une valeur inhabituellement élevée pour la loi du χ^2, alors, on peut penser que les deux variables ne sont pas indépendantes : on admet alors H_1.

Degré de liberté dans un test de χ^2

La loi du χ^2 prend pour paramètre le degré de liberté. Pour savoir à quelle loi du χ^2 comparer la distance du χ^2, il est d'abord nécecessaire de calculer le degré de liberté d'un tableau croisé.

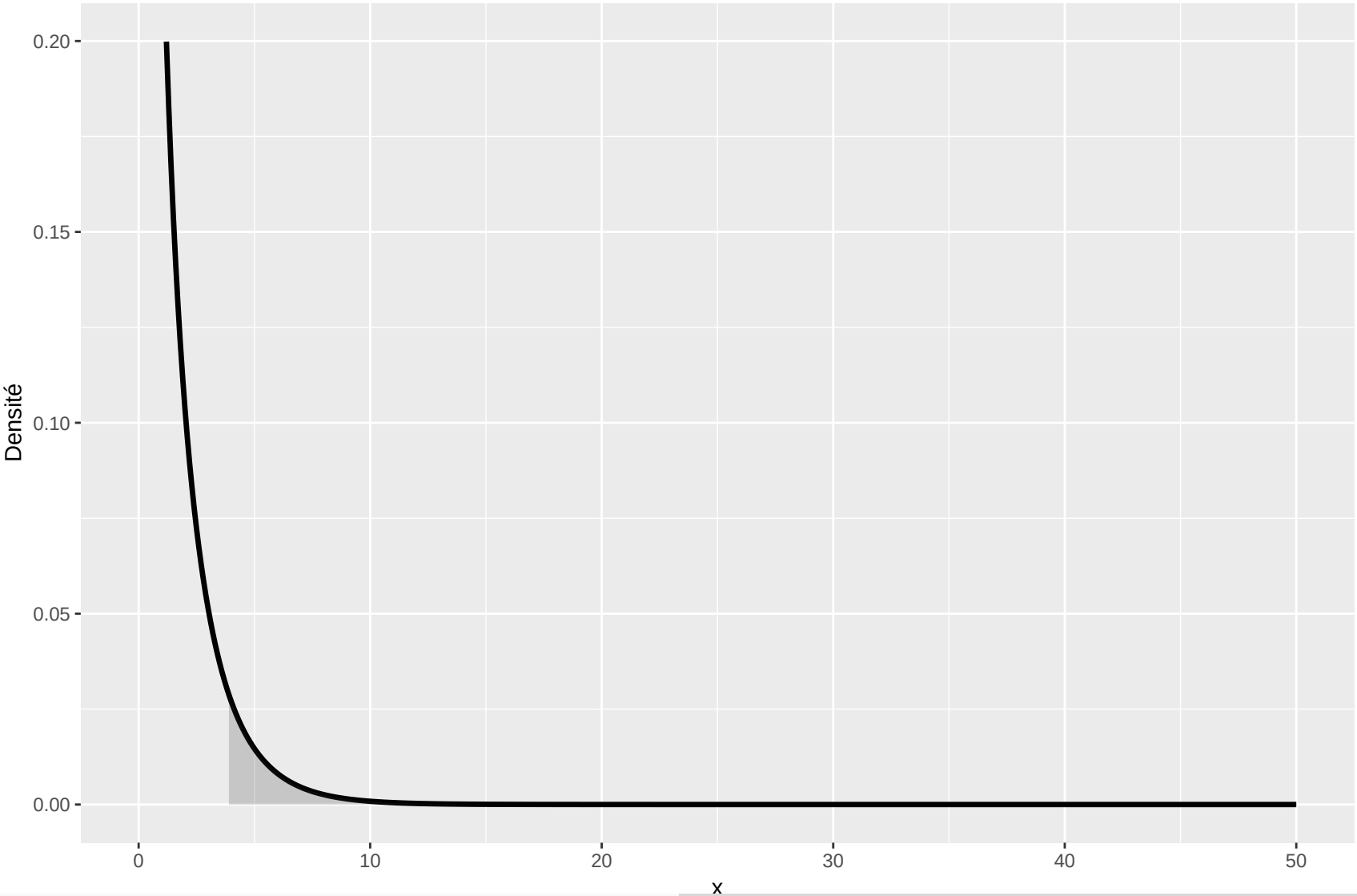
Soit un tableau de contingence contenant i lignes et j colonnes. Alors, le nombre de degré de liberté du tableau est de :

$$dl = (i-1)(j-1)$$

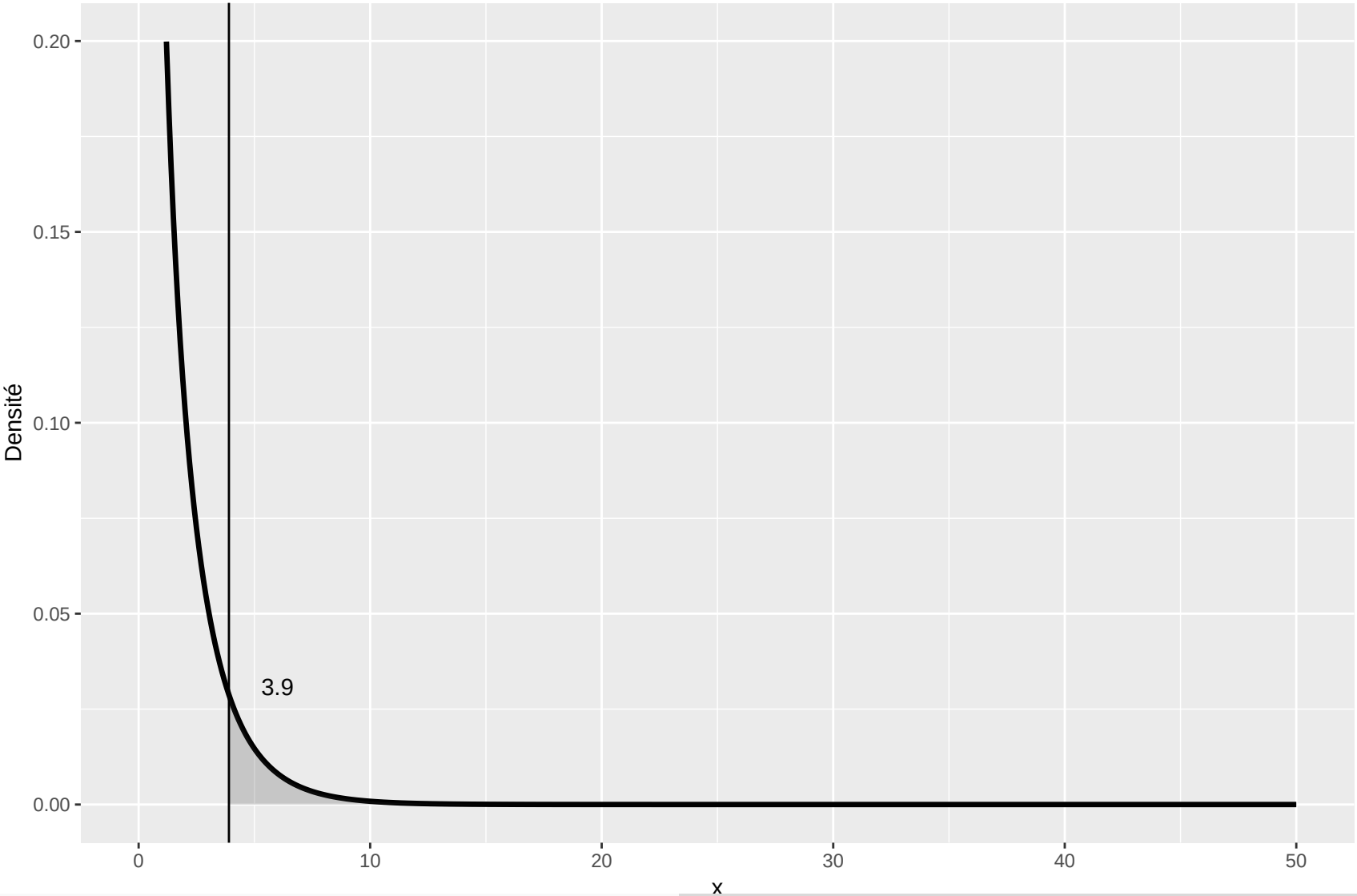
Détermination d'un seuil critique

On fixe $\alpha = 0.05$ (5%). Avec la loi du χ^2, on peut déterminer quelle est la valeur minimale de distance au χ^2 après laquelle on dispose de moins de χ^2 chances d'observer une telle distance si les variables sont indépendantes. Dans le graphique suivant, cela correspond à la zone grisée.

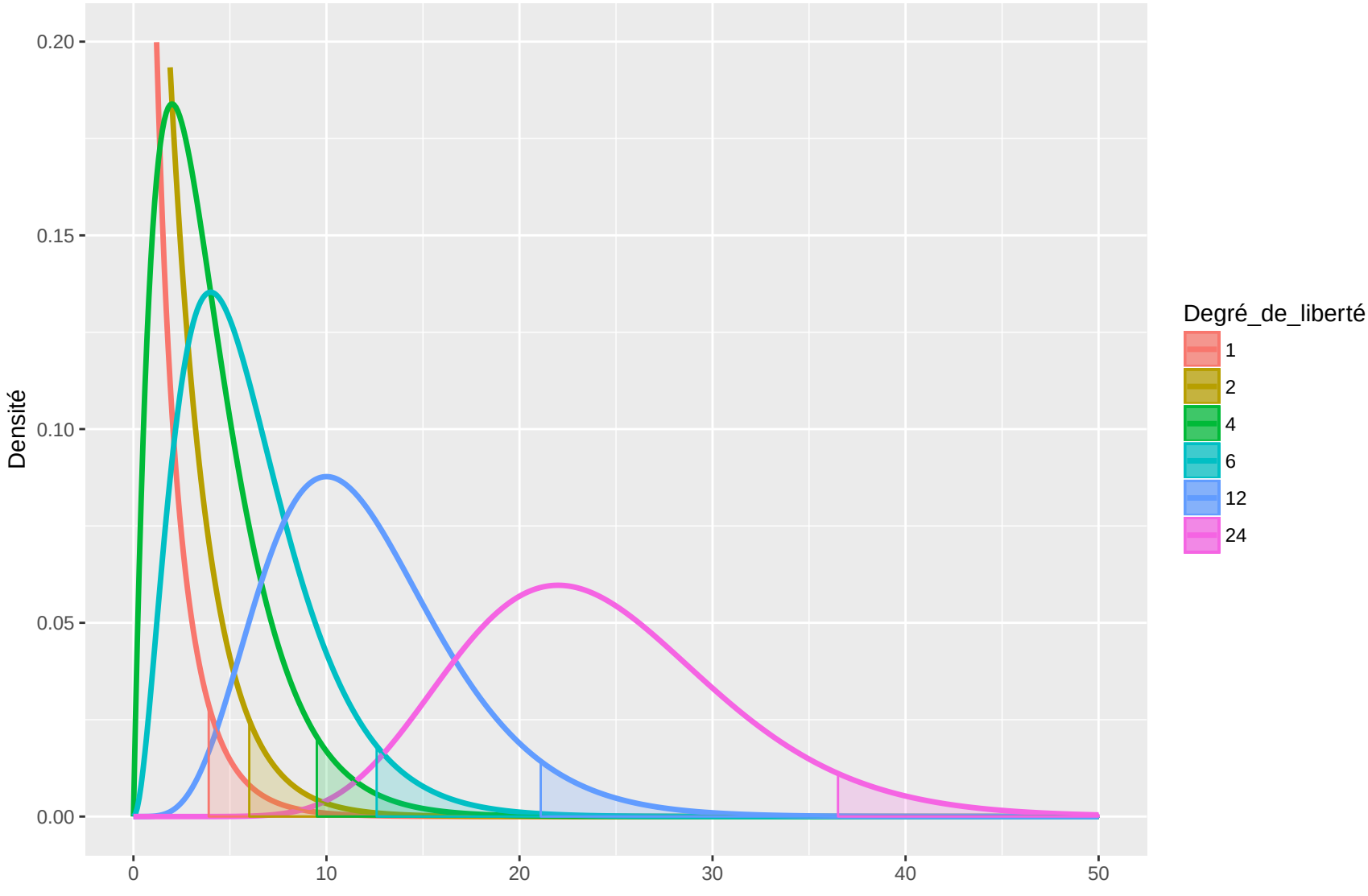
Déterminations d'un seuil critique pour k = 1



Détermination d'un seuil critique pour k = 1



Généralisation : seuil critique par ddl



Plus belle la vie : Une écoute genrée ?

Écarts et χ^2 par case

		Ne cite pas *PBLV*	Cite *PBLV*	Total
Homme	Observé	2185	209	2394
	Théorique	2114	280	
	$O-T$	71	-71	
	χ^2	2,4	18,2	
Femme	Observé	2233	377	2610
	Théorique	2304	306	
	$O-T$	-71	71	
	χ^2	2,2	16,7	
Total		4418	586	5004

$$\chi^2_{ij} = \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \qquad \chi^2 = \sum_{i,j} \chi^2_{ij} = 39,4$$

$\alpha = 0.05$
$\chi^2 > 3{,}9$

Conclusion

$$39{,}4 > 3{,}9$$

On sait donc que probabilité de trouver une distance du χ^2 telle que celle que nous avons mesurée pour ce tableau est inférieure à 0.05. Les logiciels de statistiques permettent de calculer précisément cette probabilité. Elle est de 0.

On peut donc **rejeter** l'hypothèse nulle. Les variables sexe et écoute de plus belle la vie ne sont pas indépendantes.

Précautions d'interprétation : la significativité

Le test d'hypothèse mesure la probabilité d'observer la situation observée si les deux variables sont indépendantes. Lorsque cette probabilité est très faible, on peut *penser que* les deux variables ne sont pas indépendantes.

Le test de χ^2 ne démontre pas la dépendance, mais seulement **la faible probabilité de l'indépendance**.

À l'inverse, **l'absence de significativité n'est pas la significativité de l'absence** : si deux variables n'apparaissent pas liées entre elles par un test de χ^2, cela ne signifie pas qu'elles soient entièrement indépendantes. Le résultat du test est affecté par le nombre de modalités, par l'effectif total, etc.

Précautions d'interprétation : la force

Le test du χ^2 ne renseigne que sur la significativité de la liaison, et non sur sa force (son intensité), ou le sens de la corrélation. Un test positif permet de penser que le lien existe, mais ce lien peut être très ténu.

Le test de χ^2 doit donc être accompagné d'autres tests, comme le V de cramer, ainsi que de la description précise du tableau de contingence.

La significativité statistique n'est pas toujours la significativité scientifique.

Précautions d'usage : effectifs et seuil

Le test de χ^2 est sensible aux effectifs théoriques (la division tend vers l'infini quand le dénominateur tend vers 0). On considère conventionnellement que le test ne peut être fait si plus de 20% des cases du tableau de contingence ont un effectif théorique inférieur à 5. Pour s'assurer que ce soit le cas, on peut recoder les variables en agrégeant les modalités rares, ou employer d'autres formes de test.

Ce seuil de 5 est cependant arbitraire. Il a été fixé par l'inventeur du test, Karl Pearson. De la même manière, le seuil de significativité α à 0.05 (5%) est un seuil arbitraire.

Précautions d'usage : chasser la p-value

Comme tous les tests d'hypothèses, le test du χ^2 est sensible à la répétition : plus on fait de tests et plus le risque de commettre une erreur de type 1 s'accroît.

Si on prend un seuil α de 5%, cela signifie que, une fois sur 20, on se trompera en interprétant un test positif. En répétant un test de χ^2 vingt fois de suite sur un même échantillon, on est à peu près sûr d'avoir des erreurs.

La publication d'articles scientifiques dans les disciplines reposant fortement sur les statistiques est en général conditionnée à la présentation de tests dont la p-value est faible. Cela conduit à favoriser la multiplication abusive de tels tests : on croise alors toutes les variables entre elles jusqu'à trouver un résultat "significatif", sans penser que cette "significativité" est sans doute du au hasard. C'est un comportement qu'il ne faut surtout pas adopter.

Comparaison de moyennes : le test de Student

Principe du test de student

Le test de Student est un *test d'hypothèse*. Il est également appelé *test t*.

Comme tous les tests d'hypothèses, le test de Student permet de calculer la distribution de données sous l'hypothèse d'indépendance.

Le test de Student est employé pour comparer la moyenne d'une variable aléatoire (c'est-à-dire d'une variable normalement distribuée) à une autre valeur, habituellement la moyenne d'une autre variable aléatoire.

Situations d'usages

Il existe donc plusieurs usages du test de Student :

- Comparer la moyenne d'une variable entre deux sous-échantillons (exemple : la taille moyenne des filles est-elle significativement différente de la taille moyenne des garçons). De ce point de vue, le test de Student permet de tester *la corrélation entre une variable quantitative (normalement distribuée) et une variable catégorielle dichotomique (qui n'a que deux modalités)*.
- Comparer deux variables quantitatives (normalement distribuées) similaires dans une même population, comme par exemple deux variables appariées (une même mesure répétée par deux fois).
- Comparer une variable quantitative (normalement distribuée) à une valeur connue (utilisé en particulier dans les sciences expérimentales, comme la médecine, pour comparer un échantillon à un échantillon-témoin).

Étapes d'un test

On suit pour le test de Student les mêmes étapes que pour le χ^2.

- Déterminer l'hypothèse nulle et l'hypothèse alternative
- Fixer un seuil α
- Calculer la probabilité p de la distribution observée sous l'hypothèse d'indépendance
- Comparer p et α et rejeter l'hypothèse d'indépendance si $p \leqslant \alpha$

Hypothèse nulle

L'hypothèse nulle, H_0, est toujours l'hypothèse d'indépendance :

- si l'on étudie le lien entre une variable quantitative et une variable catégorielle dichotomique : *la moyenne de la variable quantitative est égale dans les deux catégories*.
- si l'on étudie le lien entre deux variables quantitatives : *les moyennes des deux variables sont égales*.
- si l'on compare la moyenne d'une variable quantitative à une valeur fixée : *la moyenne est égale à la valeur*.

Hypothèse alternative

L'hypothèse alternative H_1 est que les deux valeurs ne sont pas égales. Plus précisément, on peut faire deux hypothèses alternatives :

- *Test bilatéral* : les deux valeurs ne sont pas égales
- *Test unilatéral* : les deux valeurs ne sont pas égales et l'une est plus grande que l'autre.

Calcul de la distance : t

Comme dans le cas du test du χ^2, on calcule la distance entre la distribution observée et la distribution sous l'hypothèse d'indépendance. Cette distance se calcule par la statistique t.

Le calcul précis de cette statistique varie selon le type de test que l'on réalise. Il dépend dans tous les cas de la taille de l'échantillon (ou de la taille de chacun des groupes étudiés dans le cas d'une comparaison) et de la dispersion de la ou des variables concernées (mesurée soit par la variance, soit par l'écart-type).

Calcul de t : cas de la comparaison d'une moyenne à une valeur connue

Soit X une variable quantitative (normalement distribuée) de moyenne $\bar{x}$ et d'écart-type σ, n la taille de l'échantillon, et μ_0 la valeur à laquelle on souhaite comparer la moyenne. Alors :

$$t = (\bar{x} - \mu_0) * \frac{\sqrt{n}}{\sigma}$$

Calcul de t : cas de la comparaison de deux moyennes (variances égales)

Soit X et Y deux variables quantitatives (normalement distribuées) de moyenne $\bar{x}$ et $\bar{y}$, d'effectif n_x et n_y et *de même écart-type* σ. Alors,

$$t = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Par conséquent, si l'on souhaite tester la différence de moyenne d'une même variable entre deux sous-groupes, on applique le même formule. Soit X une variable quantitative, $\bar{x}_a$ sa moyenne dans le premier sous-groupe, $\bar{x}_b$ sa moyenne dans le second sous-groupe, n_a la taille du premier sous-groupe et n_b la taille du second sous-groupe

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sigma\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

Calcul de t : cas de la comparaison de deux moyennes (variances inégales : test de Welch)

Soit X et Y deux variables quantitatives (normalement distribuées) de moyenne $\bar{x}$ et $\bar{y}$, d'effectif n_x et n_y et d'écart-type σ_x et σ_y. Alors,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x}{n_x} + \frac{\sigma_y}{n_y}}}$$

Entre deux sous-groupes : Soit X une variable quantitative, $\bar{x}_a$ sa moyenne dans le premier sous-groupe, $\bar{x}_b$ sa moyenne dans le second sous-groupe, n_a la taille du premier sous-groupe et n_b la taille du second sous-groupe, σ_a et σ_b les écart-types du premier et du second sous-groupe :

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{\sigma_a}{n_a} + \frac{\sigma_b}{n_b}}}$$

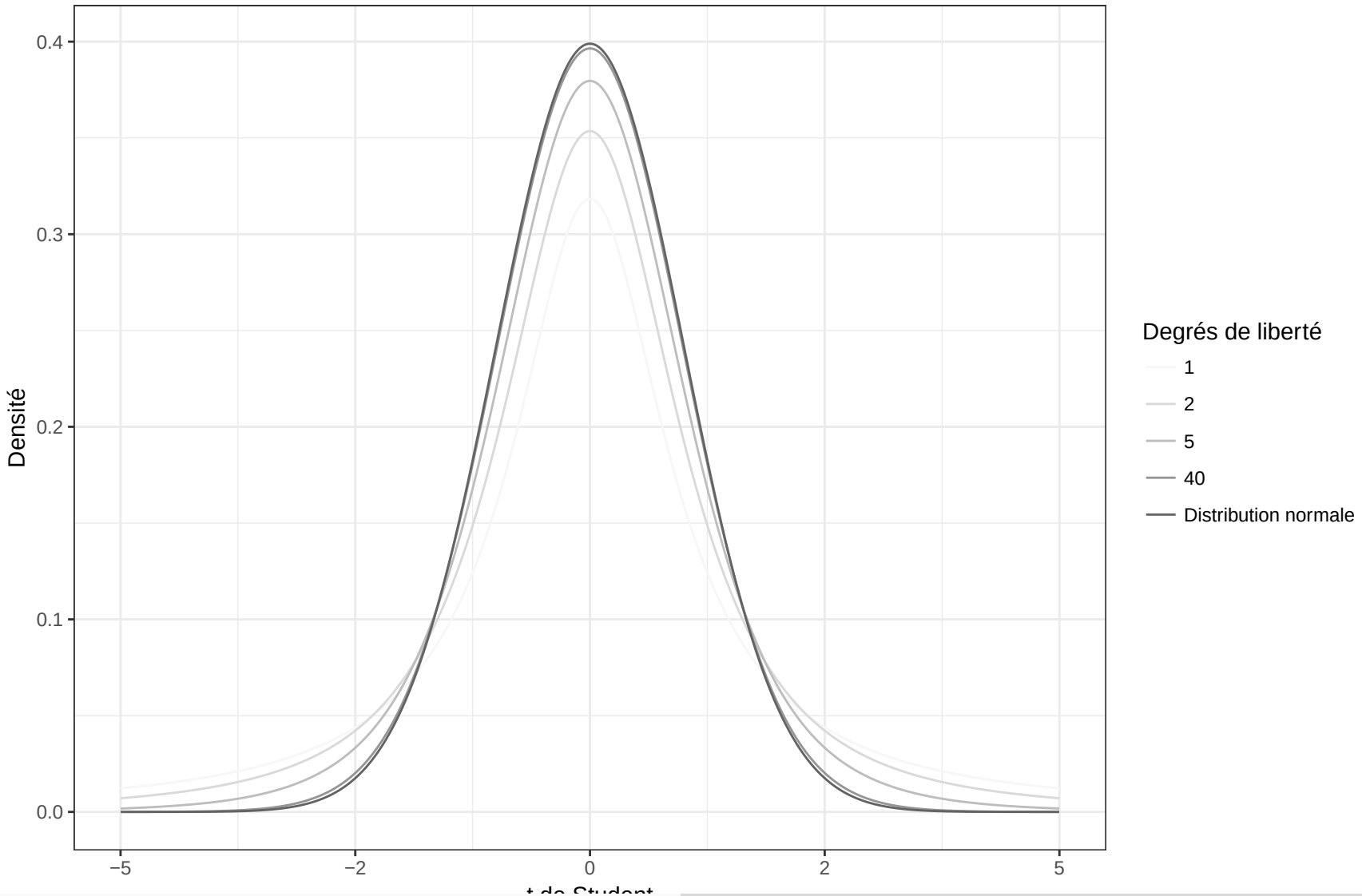
Calcul de p : la loi de Student

La statistique t suit une loi de Student à n-1 degrés de liberté :

$$St_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

On remarque que cette loi converge vers une loi normale lorsque n tend vers l'infini.

Distribution de la loi de Student



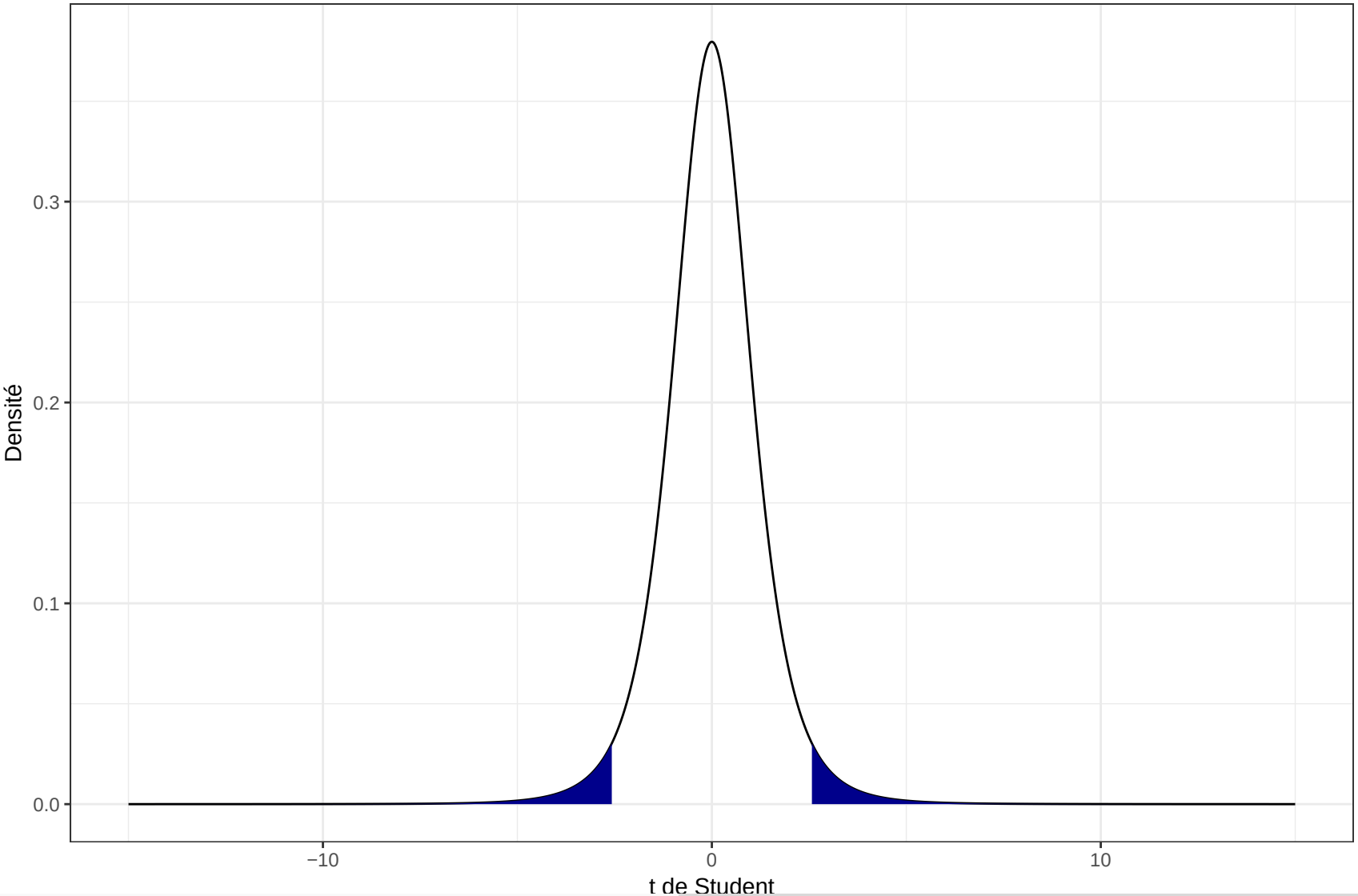
Détermination de la valeur-test

À α égal (traditionnellement fixé à 0.05 ou 5%), la détermination des aires critiques (et donc, le calcul de la valeur p) diffère selon selon que le test est unilatéral ou bilatéral.

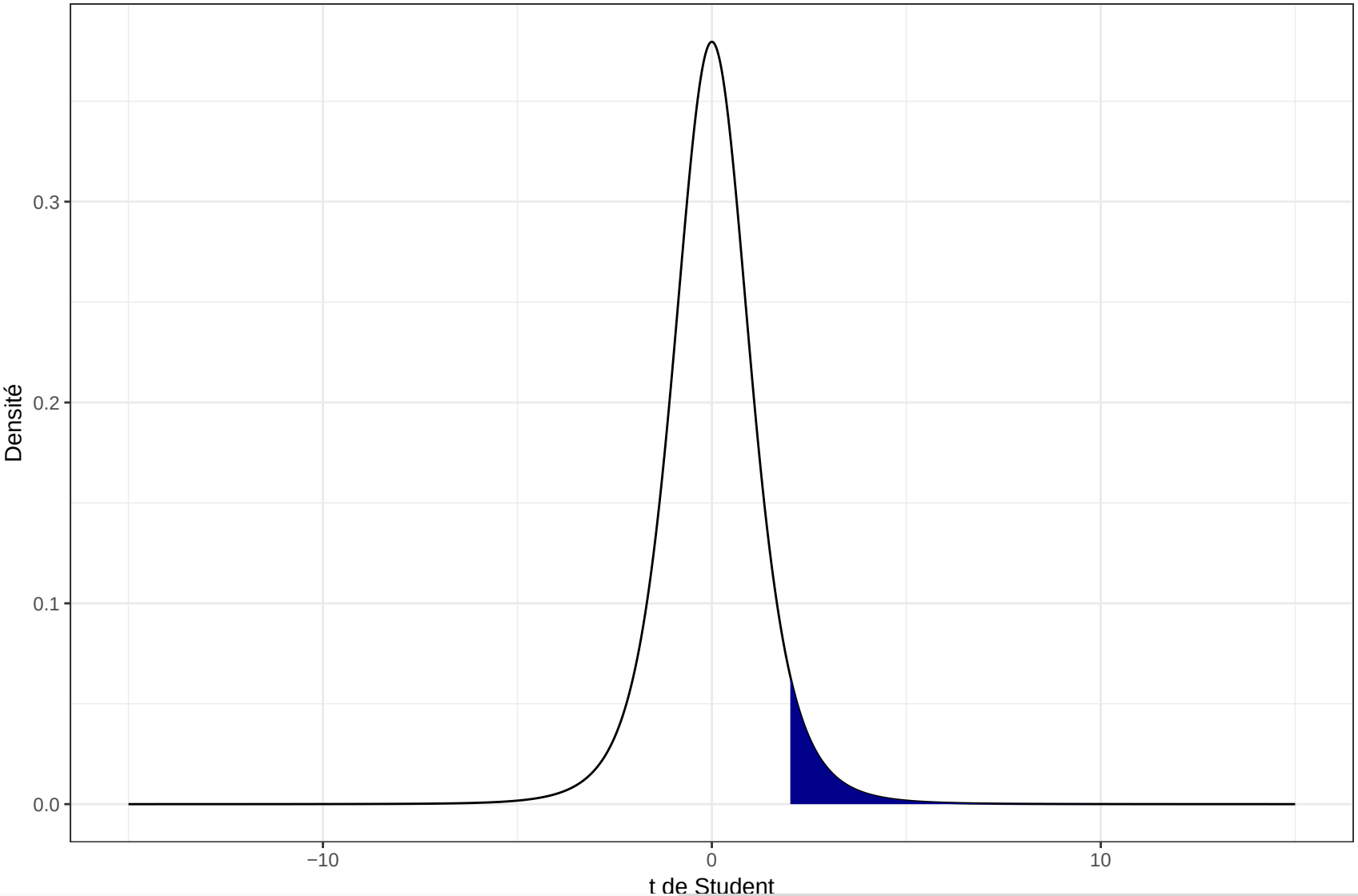
- *Test bilatéral* : H_0 peut être rejeté si t se situe dans l'une des deux aires critiques [voir slide suivante].
- *Test unilatéral* : H_0 peut être rejeté si t est dans l'aire critique, à gauche ou à droite selon le test [voir slide suivante].

Cette différence est illustrée dans les deux graphiques suivants. L'aire colorée sous la courbe représente les valeurs de t correspondant à $p \leqslant \alpha$ selon le type de test.

Valeur-test pour 5 degrés de liberté : test bilatéral



Valeur-test pour 5 degrés de liberté : test unilatéral à droite



Conclusion du test d'hypothèse

Comme dans le cas du test de χ^2, on rejette l'hypothèse nulle si $p \leqslant \alpha$ et on la conserve sinon. Le graphique, ainsi que les tables t donnent la valeur seuil au-delà (ou en-deça) de laquelle $p \leqslant \alpha$. Les logiciels de statistique permettent eux de calculer exactement la valeur p.

Le plus souvent, le test est présenté de la façon suivante : donne le degré de liberté, la valeur t, le type de test employé, et la p-value associée. Par exemple, le résultat d'un test donné par le logiciel R :

estimate	statistic	p.value	parameter	method	alternative
2.401	3.917	0.000165	99	One Sample t-test	two.sided

Table de Student (test bilatéral)

ddl	.10	.05	.01	.001
1	6.314	12.71	63.66	636.6
2	2.92	4.303	9.925	31.6
3	2.353	3.182	5.841	12.92
4	2.132	2.776	4.604	8.61
5	2.015	2.571	4.032	6.869
10	1.812	2.228	3.169	4.587
20	1.725	2.086	2.845	3.85
30	1.697	2.042	2.75	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
100	1.66	1.984	2.626	3.39
1000	1.646	1.962	2.581	3.3

Table de Student (test unilatéral)

ddl	.10	.05	.01	.001
1	3.078	6.314	31.82	318.3
2	1.886	2.92	6.965	22.33
3	1.638	2.353	4.541	10.21
4	1.533	2.132	3.747	7.173
5	1.476	2.015	3.365	5.893
10	1.372	1.812	2.764	4.144
20	1.325	1.725	2.528	3.552
30	1.31	1.697	2.457	3.385
40	1.303	1.684	2.423	3.307
50	1.299	1.676	2.403	3.261
100	1.29	1.66	2.364	3.174
1000	1.282	1.646	2.33	3.098

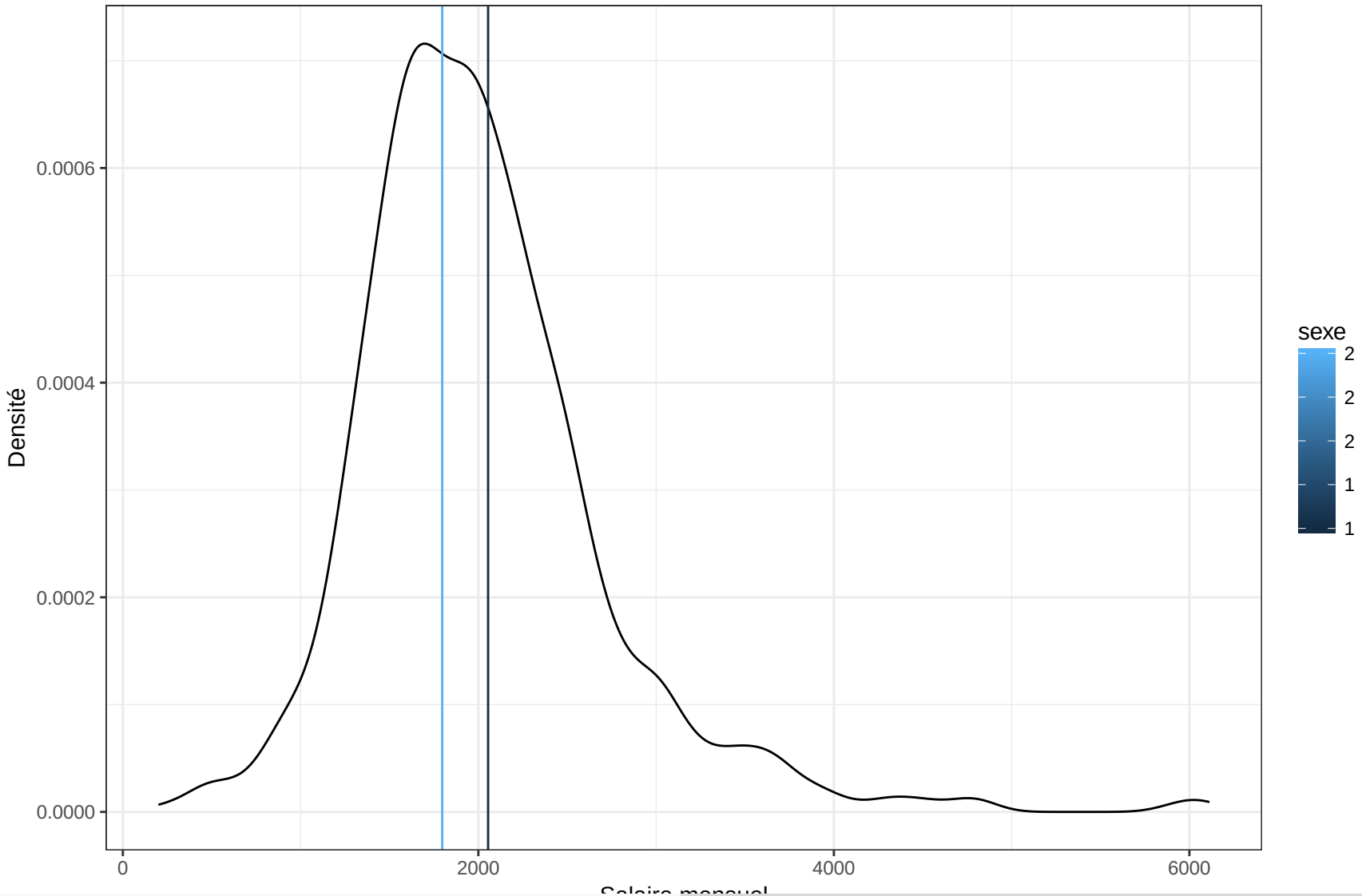
Exemple

On analyse ici les données issues de l'enquêtes emploi 2013. On s'intéresse aux revenus salariés de la catégorie sociale des techniciens (sous-catégorie des professions intermédiaires).

On cherche à savoir si le salaire mensuel moyen des 0 techniciens de l'échantillon (NaN, écart-type : NA) est significativement *supérieur* au salaire mensuel moyen des 0 techniciennes de l'échantillon (NaN, écart-type : NA).

- H_0 : la différence entre le salaire mensuel moyen des hommes et des femmes techniciens est inférieure ou égale à 0 ;
- H_0 : la différence entre le salaire mensuel moyen des hommes et des femmes techniciens est supérieure à 0.

Revenus salariés mensuels des techniciens : par sexe



Calcul de t

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sigma\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

$$t = \frac{NaN - NaN}{NA\sqrt{\frac{1}{0} + \frac{1}{0}}}$$

$$t = 4.7$$

$$p = 0.00000174$$

Estimation d'un paramètre et intervalles de confiance

Motivation

Si nous cherchons à connaître le revenu moyen d'un technicien, nous pouvons utiliser l'enquête emploi 2013 et calculer le revenu moyen des techniciens *dans cet échantillon*. Or le but final de notre étude, c'est de connaître le revenu moyen *dans la population*.

Dans certaines circonstances, il est possible de **calculer un paramètre statistique d'un échantillon** puis **d'estimer sa valeur dans la population d'origine**. C'est à cela que servent les intervalles de confiance, en fournissant une idée du degré de précision du paramètre obtenu avec l'échantillon.

Interprétation et intuition

Le revenu moyen d'un technicien dans l'échantillon Enquête Emploi est de 2017 €. Pour un degré de confiance à 95%, l'intervalle de confiance associé est $[1978; 2055]$.

Cela signifie que si la "vraie" moyenne de la population n'est pas comprise dans cet intervalle, alors la probabilité d'obtenir une moyenne de 2017 dans un échantillon aléatoire de même taille est inférieure à 5%.

Une autre façon de comprendre le degré de confiance 95% est la suivante : en prenant 100 échantillons de la même population et en calculant un intervalle de confiance pour chacun, 95 contiendraient la "vraie" moyenne.

Une erreur fréquente d'interprétation

Rappel : le revenu moyen d'un technicien est de 2017 €. L'intervalle de confiance à 95% est $[1978; 2055]$.

Il ne faut surtout pas interpréter un intervalle de confiance comme suit :

> *Le revenu moyen de tous les techniciens a 95% de chances de se situer dans cet intervalle.*

En effet, le revenu moyen de tous les techniciens n'est pas une variable aléatoire : il se situe soit dans l'intervalle, soit en dehors de l'intervalle. C'est le paramètre (dans ce cas, le revenu moyen des techniciens dans l'échantillon) qui est une variable aléatoire !

Cette interprétation n'a donc aucun sens dans le paradigme fréquentiste des intervalles de confiance.

Cas général

Soit une population P, dont nous cherchons à estimer un paramètre p et un échantillon E dont le paramètre estimé est $\bar{p}$.

Sous certaines conditions, on peut montrer que le paramètre estimé $\bar{p}$ est une variable aléatoire, qui suit une certaine loi de probabilité associée à la population.

Un intervalle de confiance autour de la valeur estimée $\bar{p}$ peut alors être construit, avec un certain degré de confiance.

Le degré de confiance ne s'applique pas à un intervalle particulier ou à un paramètre, mais à la procédure en général.

Le calcul d'un IC dépend du paramètre estimé : nous verrons le cas d'une moyenne et d'une proportion.

Calcul d'un IC pour une moyenne

Soit une population P dont la variable X suit une distribution normale et soit E un échantillon aléatoire de P de taille n, avec $\bar{x}$ la moyenne de X dans E. On cherche à estimer la moyenne $\bar{X}$.

Si l'on connaît l'écart-type $\sigma(X)$, alors l'IC I à 95% est :

$$I = [\bar{x} - \frac{2\sigma(X)}{\sqrt{n}}; \bar{x} + \frac{2\sigma(X)}{\sqrt{n}}]$$

Remarques :

- En pratique, $\sigma(X)$ est souvent remplacé par l'écart-type dans E
- Le facteur 2 change selon le degré de confiance souhaité (1 pour 68%, 3 pour 99,7%, etc.)

Exemple

Calcul de l'intervalle de confiance à 99,7% pour le revenu moyen d'un technicien :

$$n = 1246$$

$$\sigma(x) = 699.3$$

$$\bar{x} = 2016.6$$

Solution et question

$$I_{99,7\%} = [2016.6 - 3\frac{699.3}{\sqrt{1246}}; 2016.6 + 3\frac{699.3}{\sqrt{1246}}]$$

$$I_{99,7\%} = [1918; 2019]$$

Or $I_{95\%} = [1935; 2001]$

- Quel intervalle est plus large ? Pourquoi ?
- Faut-il préférer un petit ou un grand intervalle ? Un petit ou un grand degré de confiance ?
- Si nous voulions obtenir un intervalle plus petit que $I_{95\%}$ tout en maintenant le degré de confiance à 95%, que devrions-nous faire ? Pourquoi ?

Calcul d'un IC pour une proportion

Le calcul d'un IC pour une proportion peut être effectué selon plusieurs méthodes. Nous voyons ici la plus simple (méthode de Wald) :

Soit une population P dont on cherche à estimer une proportion π, et E un échantillon aléatoire de taille n dont la proportion estimée est p. Alors l'intervalle de confiance I avec un degré de confiance $1 - \alpha$ ($\alpha \in [0; 1]$) :

$$I = p \pm z\sqrt{\frac{p(1-p)}{n}}$$

où z est la valeur pour laquelle la fonction de répartition de la loi normale centrée réduite est égale à $1 - \alpha/2$.

En pratique, $z = 1,96$ pour un degré de confiance de 95% ; 1,64 pour 90% ; 2,58 pour 99%.

Exemple

Sur nos 1246 techniciens, 85% sont des hommes.

$I_{95\%}$ de masculinisation $=$ [0.8 ; 0.9].

Les limites du NHST

Tester des hypothèses, tester des données

On aimerait souvent tester directement des hypothèses : la statistique inférentielle devrait être capable de mesurer la probabilité d'une hypothèse d'intérêt pour le chercheur. La statistique inférentielle fréquentiste ne fait pas cela. Elle mesure en réalité la probabilité d'obtenir des données au moins aussi extrême sous l'hypothèse nulle, et par conséquent, le risque d'erreur en cas de rejet de cette hypothèse nulle.

Cette distinction est fondamentale : *il est abusif d'interpréter le résultat d'un NHST comme une mesure de vraisemblance de l'hypothèse alternative*.

Il y a en fait deux problèmes distincts :

- Les NHST mesurent la vraisemblance *des données* et pas la vraisemblance *des hypothèses*.
- Les NHST renseignent sur *l'hypothèse nulle* et non sur *l'hypothèse alternative*. Le NHST ne mesure jamais la probabilité d'une hypothèse alternative (or, c'est ce que l'on souhaite mesurer).

Exemple

Je mesure l'écart entre salaire moyen des hommes et des femmes $\omega_H - \omega_F$ dans mon échantillon et je souhaite inférer ce résultats à la population. Je réalise pour cela un test de student :

- Hypothèse nulle : $\omega_H - \omega_F \leq 0$ (les salaires des hommes et des femmes ne sont pas différents (0) ou les salaires des femmes sont supérieurs à ceux des hommes ($<$0)).
- Hypothèse alternative : $\omega_H - \omega_F > 0$ (les salaires des hommes sont supérieurs à ceux des femmes)

Je fais mon test, qui me fournit un résultat, une *p-value*, faible, disons 0.01 (1%).

Que signifie ce chiffre ? *Si les salaires des hommes et des femmes étaient égaux ou les salaires des hommes inférieurs à ceux des femmes (c'est-à-dire sous l'hypothèses nulle), alors la probabilité de mesurer des données aussi extrêmes que celles dont nous disposons serait de 1%. Donc, nous pouvons raisonnablement penser que ça n'est pas le cas.*

Significativité et force

Les NHST mesurent des significativités, entendu comme la vraisemblance des données sous une hypothèse nulle. Le plus souvent, cependant, cette hypothèse nulle est une pure fiction. Par exemple, de nombreux tests d'hypothèse cherchent à déterminer si une valeur d'un paramètre aléatoire est strictement égal à 0 (hypothèse nulle). Or, la probabilité d'une égalité stricte à zero est infime.

Cela signifie que, à mesure que la taille d'un échantillon augmente, tous les NHST que l'on peut faire sur cet échantillon tendent à devenir "positifs" ($p \leqslant \alpha$) ; autrement dit, comme les hypothèses nulles sont une fiction, toutes vont finir un jour ou l'autre par être rejetées.

Mais la significativité n'est pas la force. Un paramètre peut être différent de zero tout en étant très proche. Il est souvent plus informatif de savoir si une valeur est *grande* que de savoir si elle est *significative*. La *force* d'une corrélation est plus importante que sa *significativité*.

Significativité et force

Par exemple, le fait de savoir simplement si les salaires des hommes et celui des femmes est différent est peu intéressant : il est quasiment impossible que, dans la population totale, les salaires moyens des hommes et ceux des femmes soient exactement égaux, *même s'il n'y avait aucune inégalité de genre* (car il y a un aléa de la répartition $=>$ en cas d'égalité stricte, on aurait toujours un petit écart des moyennes).

Par contre, la taille de l'écart est intéressant : une société égalitaire serait une dans laquelle la différence entre les deux moyennes est négligeable (n'a pas de conséquence sur l'autonomie financière, la qualité de vie, etc.) ; une société est d'autant plus inégalitaire qu'elle est importante (une société ou l'écart est de 10% est moins inégalitaire qu'une ou il est de 15%).

La conclusion pour l'usage des statistique est la suivante : *il ne faut pas fonder ses conclusions scientifique sur le seul examen de la significativité*.

Corrélation et causalité

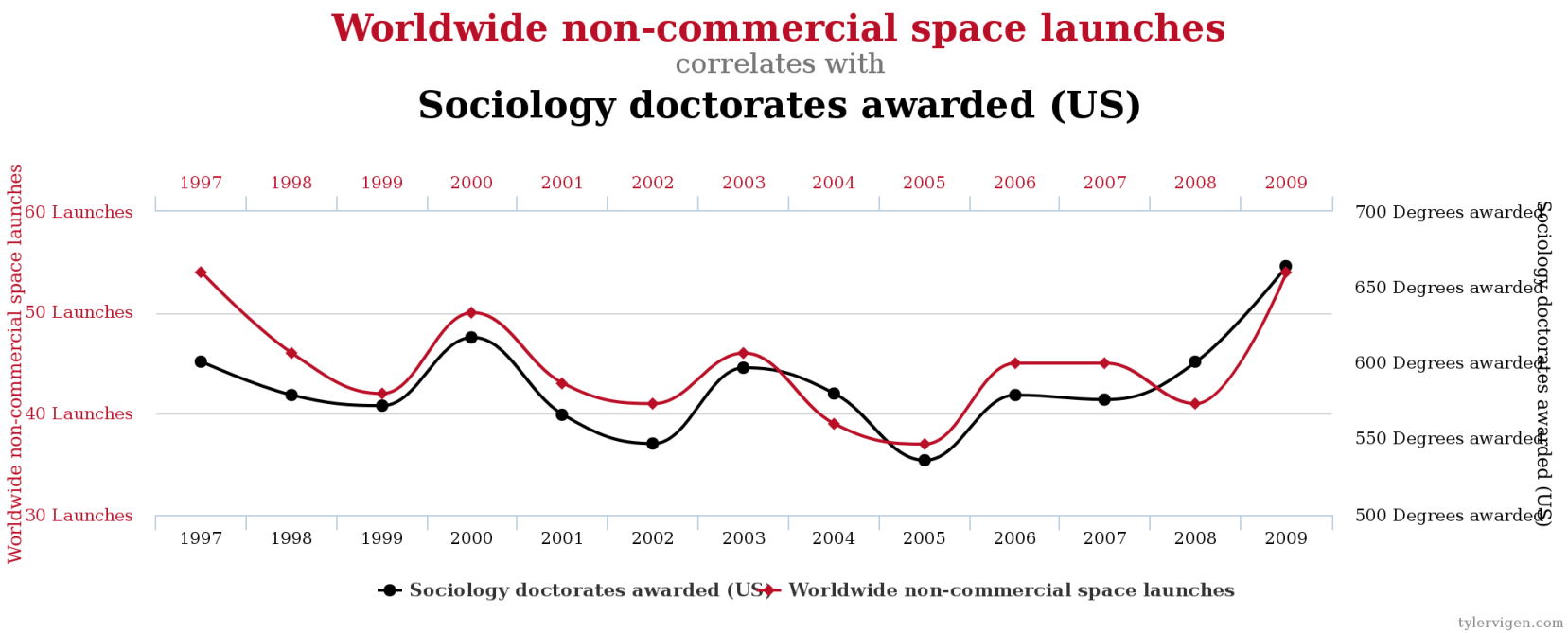
La causalité se référe à la *nécessité* d'une *loi* (logique, physique, éventuellement sociale, économique, etc.).

Définir cause et effet n'est pas dans les ambitions de ce cours. Notons simplement que les définitions sont multiples. En particulier, une cause peut être **déterministe** (A cause *toujours* B) ou **probabiliste** (A augmente la probabilité de B).

En utilisant des données d'enquête, *nous ne pouvons qu'observer des corrélations*. *Une corrélation est une variation concomittante de deux phénomènes* selon une relation statistique, le plus souvent linéaire.

Ce sont les méthodes statistiques qui peuvent *éventuellement* appuyer des explications causales.

Corrélation n'implique pas causalité (1/2)



Source : http://www.tylervigen.com/spurious-correlations

Corrélation n'implique pas causalité (2/2)

Une corrélation peut être dûe :

- à des phénomènes extérieurs à notre enquête (**variable exogène**)
- à d'autres phénomènes présents dans notre échantillon (**variable cachée**)
- à des variations liées au hasard, aux conditions de l'enquête
- à une causalité entre les deux phénomènes

De même, l'absence de corrélation n'implique pas absence de causalité :

- Un effet causal ne se décrit pas toujours selon un modèle simple et linéaire.
- Les données peuvent contenir un biais de sélection, et donc ne pas présenter suffisamment de variations pour observer une corrélation. Exemple : étudier les inégalités de revenu dans une profession aux revenus homogènes.