C5. S’assurer des différences. Intervalles de confiance et tests

Retour sur l'article de François Héran

- Dans l'enquête de François Héran, les chefs de ménage hommes possèdent en moyenne 0,56 chiens. Les chefs de ménages de femmes possèdent 0,27 chiens.
  - Est-ce différent ?
  - Est-on prêt à croire que, dans la population globale, il existe une différence du nombre de chiens possédés en fonction du sexe du chef de ménage ?
- Les hommes possèdent 0,39 chats contre 0,25 chats pour les femmes.
  - Femmes et hommes se différencient-ils plus par le nombre de chiens possédés que par le nombre de chats ?
- 53% des sans diplôme possèdent un animal domestique contre 44% des diplômés supérieurs au bac.
  - Que peut-on dire de cette différence ?

L'approche par des intervalles de confiance

- Rappel. Expérimentation vignette
- Que signifie le petit crochet autour du bâton ?
- Que peut-on conclure ?

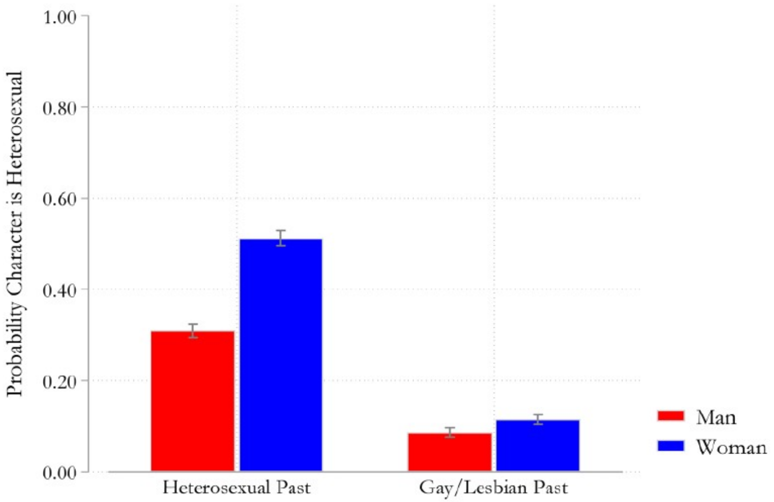


Figure 2. Probability Rating That Vignette Character Is Heterosexual, Study 1

Un exemple de test de comparaison de moyennes

TABLEAU 4
AGE MOYEN D'ACCÈS À LA CLASSE 2 DU RANG A (PROFESSEURS ET DIRECTEURS DE RECHERCHE) DES MATHÉMATICIENS ET SCIENTIFIQUES CONNEXES*
VARIATIONS SELON LA FILIÈRE DE FORMATION ET LE SEXE

		Mathéma-ticiens	Scientifiques connexes
Ensemble	moyenne	**38,5**	**40,3**
	écart type	**5,4**	**5,7**
Hommes	moyenne	38	41,4
	écart type	(5,3)	(5,5)
	effectif	123	64
Femmes	moyenne	40,9	41,9
	écart type	(5,4)	(7,7)
	effectif	18	16
	seuil de significativité du F test de différence entre moyennes	.029	ns

Un exemple de test d'un tableau croisé

TABLEAU I. – *Valorisation des différentes vertus des mathématiques*

Il fallait attribuer un rang de 1 à 4 aux quatre vertus ainsi classées de la moins à la plus valorisée, le même rang pouvant être attribué à deux vertus distinctes, le rang 0 attribué à une vertu absolument non valorisée. On compare les vertus deux à deux : V > W signifie que V est plus valorisée que W. La notation a/p est relative au test de comparaison des mathématiciens appliqués aux mathématiciens purs ; on comprendra de même les notations c/p et c/a où c désigne les scientifiques connexes.

	Utilité > beauté	Utilité = beauté	Utilité < beauté	**Total**	**Seuil X^2** a/p	**Seuil X^2** c/p et c/a
Mathématiciens purs (432 observations)	19	24	57	100	.0001	.0001
appliqués (224 observations)	52	23	25	100		ns
scientifiques connexes (357 observations)	55	24	21	100		

Outil pour dire si une différence est significative

- Connaître la dispersion autour de la moyenne
- Établir un intervalle de confiance de la moyenne
- Ou faire un test
- Le test : une manière (imparfaite) de mesurer le risque que l'on prend lorsque l'on fait confiance dans une différence statistique apparente

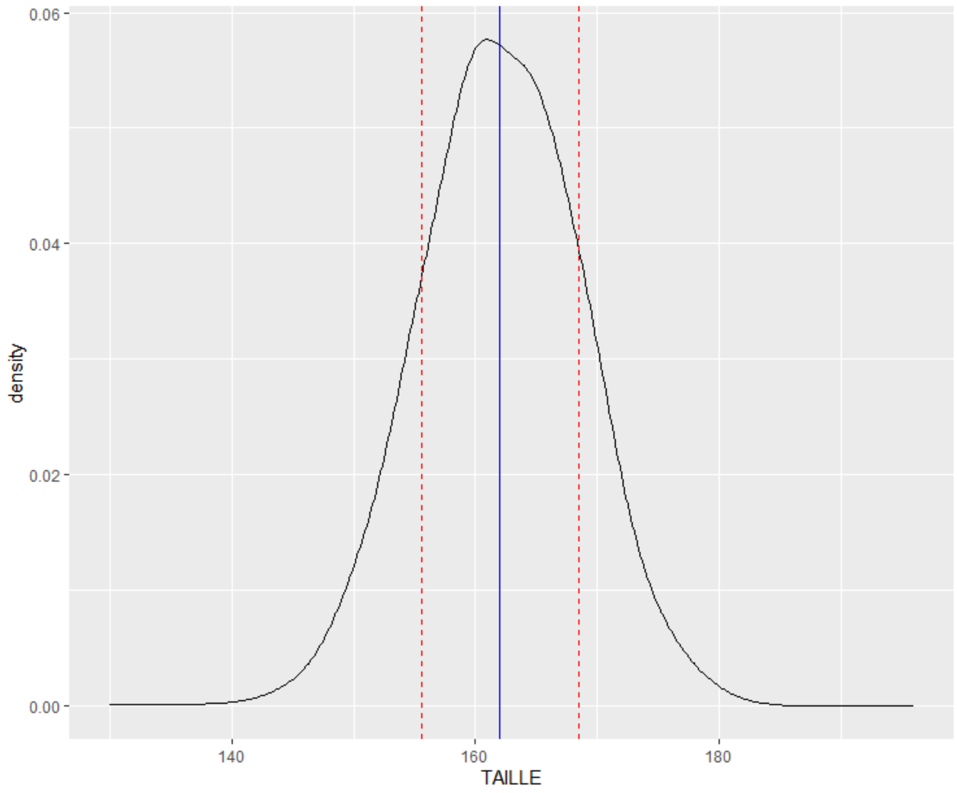
L'intervalle de confiance

- Les propriétés de la loi normale
- Pourquoi la moyenne a une distribution normale ?
- Comment calculer l'intervalle de confiance

La loi normale

- La loi normale : une loi de probabilité
  - 2 paramètres : moyenne μ, variance σ²
- Formule compliquée
  - Densité $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - Probabilité continue
  - Fonction de répartition

$$P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2}} \, dt$$



De nombreux phénomènes suivent (à peu près) une loi normale
Ici la taille des femmes : m=162, s=6.5

Une forme normale indique la distribution des observations

- 50 % des observations

  [m – 0.675*s ; m + 0.675*s]

  avec m : moyenne et s : écart-type

- 68,27 % des observations dans l'intervalle

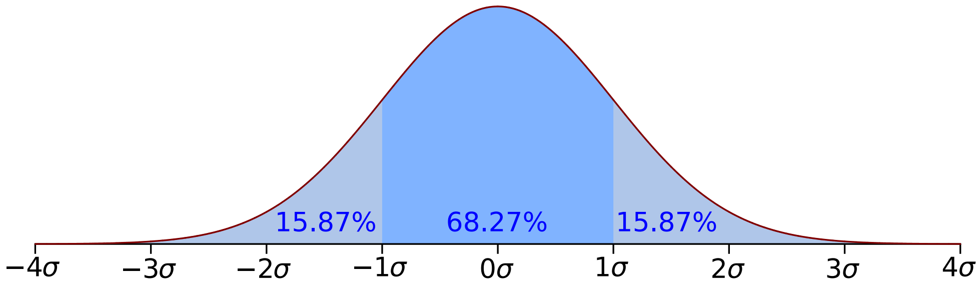
  [m – s ; m + s]

- 95 % des observations

  [m – 1.96*s ; m+ 1.96*s]

- 95,45 % des observations se trouvent

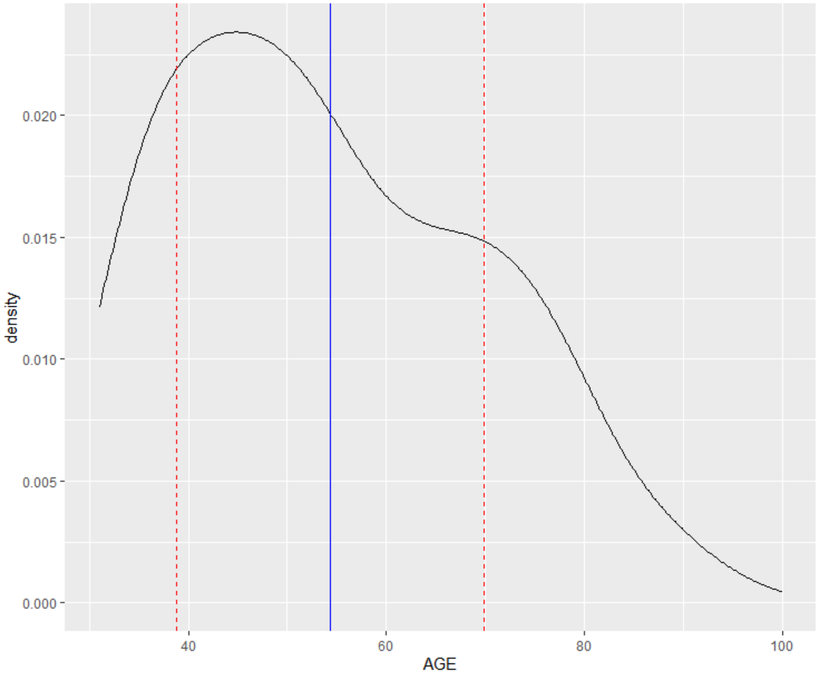
  [m – 2*s ; m+ 2 *s]

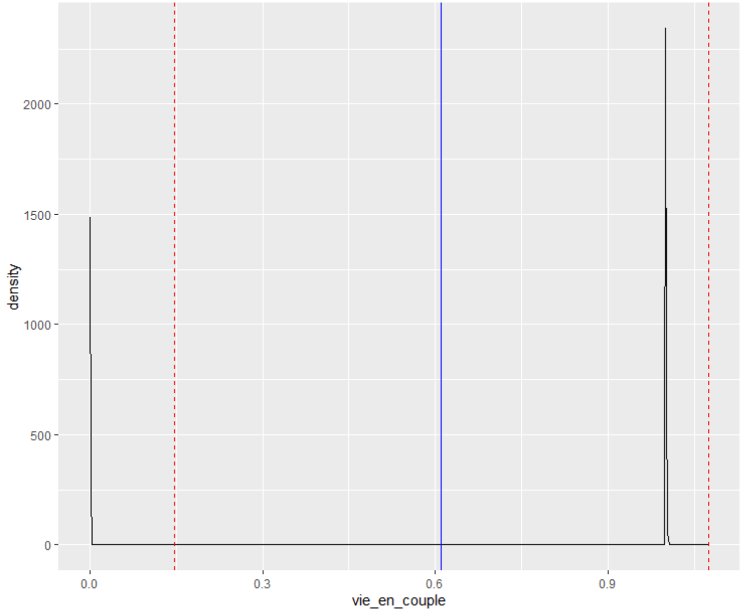


Exemple la taille des femmes (m=162, s=6.5, n=2278)

- Entre m-s et m+s : 68.65 %
- Entre m-2*s et m+2*s : 93.3 %

Tous les phénomènes ne suivent pas une loi normale





Ex l'âge des femmes (>30 ans), enquête EPCV 2001
61 % dans [m-s , m+s]

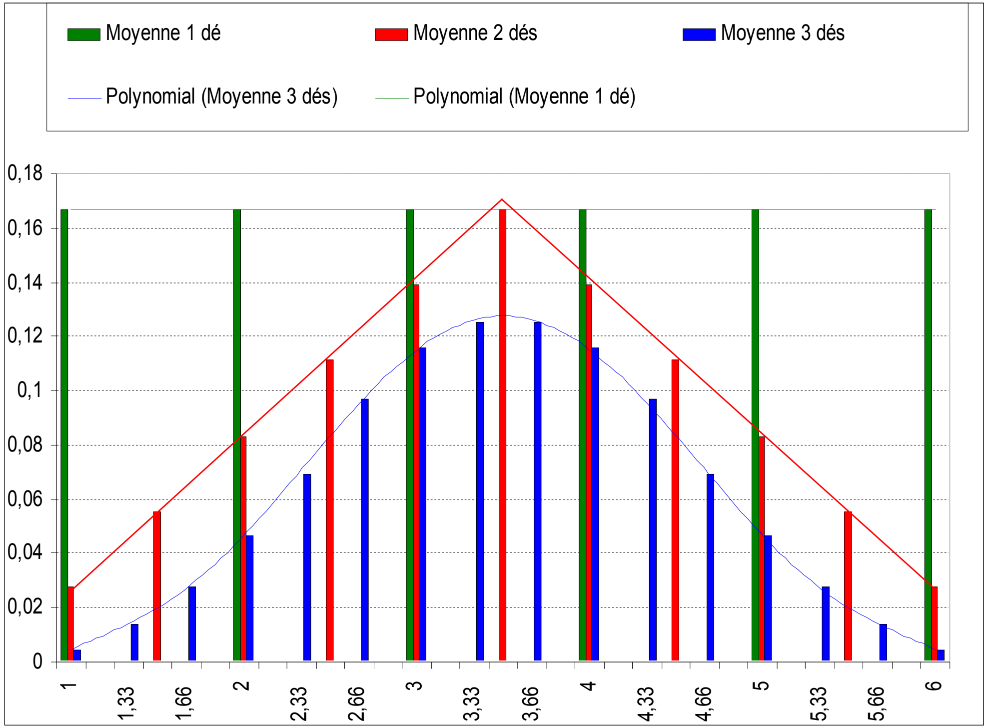
Ex La vie en couple (0=non, 1=oui), enquête EPCV 2001
61 % dans [m-s , m+s]

De l’observation à la moyenne

- Si je tire au sort une observation
  - Dans une loi normale
  - 68 % de chance d’être [moyenne – 1 écart-type ; moyenne + 1 écart-type]
  - Mais pas si le phénomène ne suit pas une loi normale
- La moyenne a une propriété particulière
- Sa distribution tend vers une loi normale
- Même pour une variable qui n’est pas normale

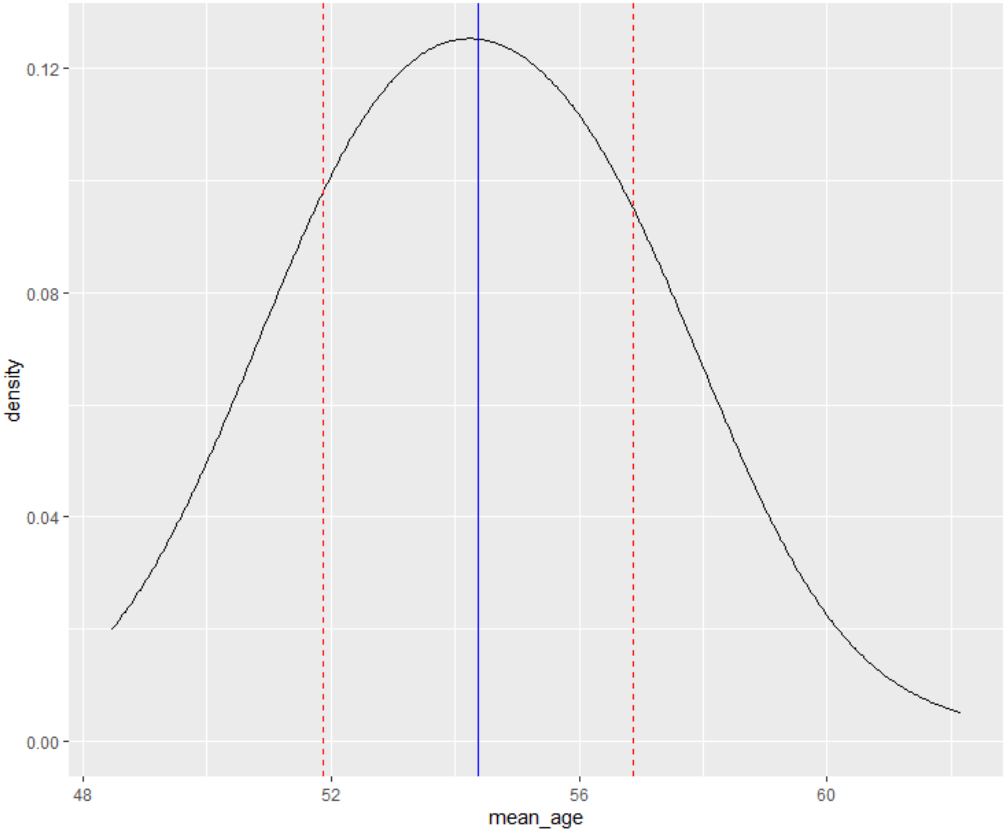
Théorème central limite

- Dé à 6 faces
  - Tirage 1 dé. Loi uniforme discrète
  - Moyenne tirage x dés → tend vers la loi normale
- Distribution d'échantillonage = La distribution des moyennes des échantillons de taille N
- Théorème Central Limite (TCL) : La distribution d'échantillonage se rapproche de la loi normale au fur et à mesure que N augmente (et ce quelle que soit la forme de la distribution de départ).



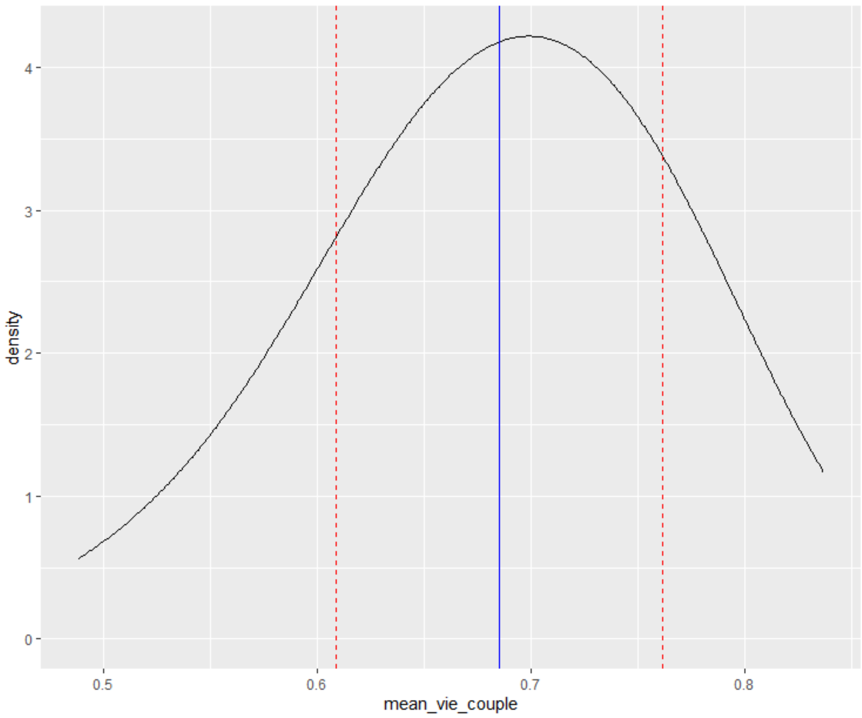
Exemple : moyenne d'âge

- Tire 100 échantillons de 43 personnes dans une population de 4296
- Calcule la moyenne sur chaque échantillon
- Distribution des moyennes d'âge (« distribution d'échantillonage »)
- La distribution des moyennes d'âge est bien normale alors que la variable âge ne l'était pas



Exemple vie en couple

- Tire 100 échantillons de 43 personnes dans une population de 4296
- Calcule la moyenne sur chaque échantillon
- Distribution des moyennes (distribution d'échantillonage)
- La distribution des moyennes de vie en couple est bien normale alors que la variable vie en couple ne l'était vraiment pas



Rappels

- Moyenne : Somme des valeurs des observations divisée par l'effectif

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

- <u>Calcul sur une population de la variance et de l'écart- type :</u>
  - Variance est un indicateur de dispersion. C'est la moyenne des carrés des écarts à la moyenne.

$$\sigma^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n} \qquad s^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}$$

  - L'écart-type est aussi un indicateur de dispersion plus pratique que la variance car plus en rapport avec la moyenne. C'est la racine carrée de la variance

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}} \qquad s = \sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}}$$

La notion d'erreur-type

- Une erreur typique : confondre erreur-type et écart-type
- Erreur-type = écart-type/racine(effectif)
- L'erreur type, c'est l'écart-type de la moyenne
- Elle correspond à l'écart type d'une distribution des moyennes des échantillons de même taille tirée au hasard dans la population totale

L’intervalle de confiance de la moyenne

- Pour un effectif > 120
- Intervalle de confiance à 95 % d’une moyenne donnée

  [moyenne-1.96*erreur-type ; moyenne+1.96*erreur-type]
- Intervalle de confiance à 90 % d’une moyenne donnée

  [moyenne-1.645*erreur-type ; moyenne+1.645*erreur-type]

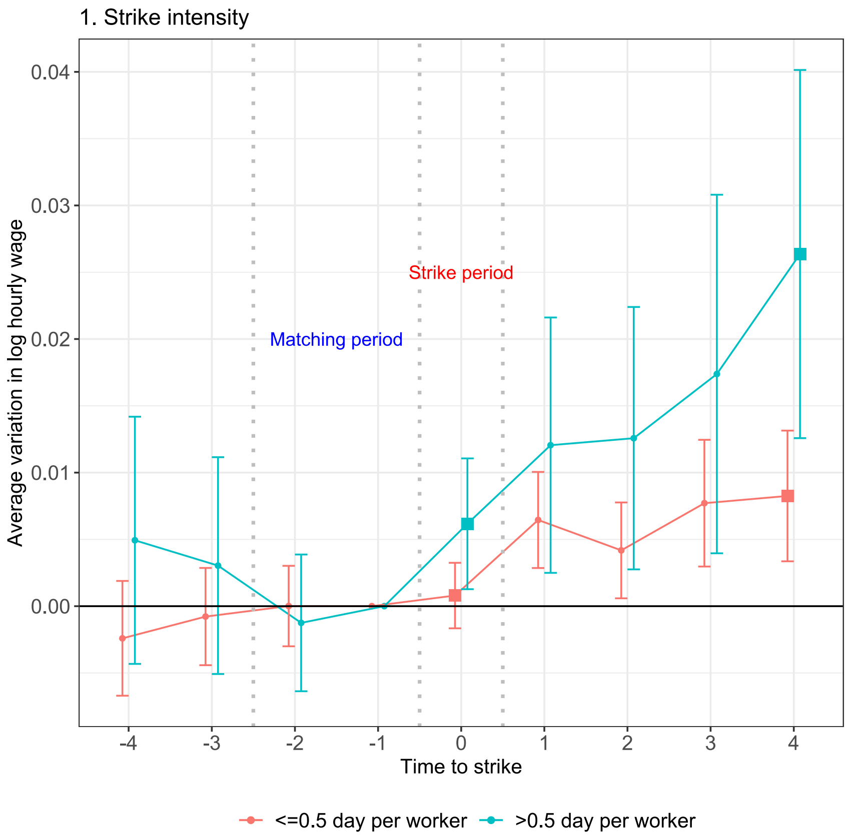
- Pour un effectif < 120
- Prendre la valeur dans le tableau suivant
- Degré de liberté =(Effectif-1)
- Exemple pour un effectif de 10, DDL=9, valeur seuil 2.262 pour IC à 95

  IC_{95}= [m-2.262*s ; m+2.262*s]

Intervalle de confiance à	50%	60%	70%	80%	90%	**95%**	98%	99.0%	99.5%	99.8%	99.9%
P(de rejeter H0 à tort)	50%	40%	30%	20%	10%	**5%**	2%	1.0%	0.5%	0.2%	0.1%
Degré de liberté											
1	1	1.376	1.963	3.078	6.314	**12.71**	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.92	**4.303**	6.965	9.925	14.09	22.33	31.6
3	0.765	0.978	1.25	1.638	2.353	**3.182**	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.19	1.533	2.132	**2.776**	3.747	4.604	5.598	7.173	8.61
5	0.727	0.92	1.156	1.476	2.015	**2.571**	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.44	1.943	**2.447**	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	**2.365**	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.86	**2.306**	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.1	1.383	1.833	**2.262**	2.821	3.25	3.69	4.297	4.781
10	0.7	0.879	1.093	1.372	1.812	**2.228**	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	**2.201**	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	**2.179**	2.681	3.055	3.428	3.93	4.318
13	0.694	0.87	1.079	1.35	1.771	**2.16**	2.65	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	**2.145**	2.624	2.977	3.326	3.787	4.14
15	0.691	0.866	1.074	1.341	1.753	**2.131**	2.602	2.947	3.286	3.733	4.073
16	0.69	0.865	1.071	1.337	1.746	**2.12**	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.74	**2.11**	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.33	1.734	**2.101**	2.552	2.878	3.197	3.61	3.922
19	0.688	0.861	1.066	1.328	1.729	**2.093**	2.539	2.861	3.174	3.579	3.883
20	0.687	0.86	1.064	1.325	1.725	**2.086**	2.528	2.845	3.153	3.552	3.85
21	0.686	0.859	1.063	1.323	1.721	**2.08**	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	**2.074**	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.06	1.319	1.714	**2.069**	2.5	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	**2.064**	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	**2.06**	2.485	2.787	3.078	3.45	3.725
26	0.684	0.856	1.058	1.315	1.706	**2.056**	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	**2.052**	2.473	2.771	3.057	3.421	3.69
28	0.683	0.855	1.056	1.313	1.701	**2.048**	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	**2.045**	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.31	1.697	**2.042**	2.457	2.75	3.03	3.385	3.646
40	0.681	0.851	1.05	1.303	1.684	**2.021**	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	**2.009**	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	**2**	2.39	2.66	2.915	3.232	3.46
80	0.678	0.846	1.043	1.292	1.664	**1.99**	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.29	1.66	**1.984**	2.364	2.626	2.871	3.174	3.39
120	0.677	0.845	1.041	1.289	1.658	**1.98**	2.358	2.617	2.86	3.16	3.373
∞	0.674	0.842	1.036	1.282	1.645	**1.96**	2.326	2.576	2.807	3.09	3.291

Comparaison visuelle des intervalles de confiance

- Si les deux intervalles de confiance sont disjoints
  - La différence entre les deux moyennes est significative
- Si les deux moyennes sont dans les intervalles de confiance croisés
  - La différence entre les deux moyennes n'est pas significative
- Autre cas
  - Il faut faire un test



Exemple Excel et intervalle de confiance

	A	B	C	D	E	F	G
1							
2							
3	**Étiquettes de lignes**	**Moyenne de nb_chien**	**Écartype de nb_chien**	**Nombre de nb_chien**			
4	femme	0.268	0.629	1303			
5	homme	0.563	0.885	4579			
6	**Total général**	**0.497**	**0.844**	**5882**			
7							
8							
9							
10	**Étiquettes de lignes**	**Moyenne de nb_chien**	**Écartype de nb_chien**	**Nombre de nb_chien**	**Erreur-type**	**Borne inférieure**	**Borne supérieure**
11	femme	0.268	0.629	1303	0.01743	0.234	0.302
12	homme	0.563	0.885	4579	0.01307	0.537	0.588
13	Total général	0.497	0.844	5882	0.01100	0.476	0.519
14							
15				Formules ligne 1	=C11/RACINE(D11)	=B11-1.96*E11	=B11+1.96*E11
16							
17							

Qu'est-ce qu'un test ?

- Tester une proposition en statistique, c'est :

« Montrer que les contradicteurs ont peu de chance d'avoir raison »

ou bien

« Calculer la probabilité que le contraire de ce que l'on pense soit vrai ! »

- Et montrer que cette probabilité est très petite.
- Une espèce de raisonnement par l'absurde : un raisonnement par double négation

Un test de student : le nombre moyen de chiens possédés diffère-t-il en fonction du sexe du chef de ménage?

Statistiques							
Variable	**CSEX**	**Nb**	**Moyenne**	**Écart-type**	**Erreur Std**	**Minimum**	**Maximum**
chienn	**1**	4579	0,5628	0,8846	0,0131	0	11
chienn	**2**	1303	0,2678	0,6292	0,0174	0	6
chienn	**Diff (1-2)**		0,2949	0,8349	0,0262		

Tests de Student					
Variable	**Méthode**	**Variances**	**DF**	**Valeur du test t**	**Pr > |t|**
chienn	**Pooled**	Equal	5880	11,25	<,0001
chienn	**Satterthwaite**	Unequal	2916	13,54	<,0001

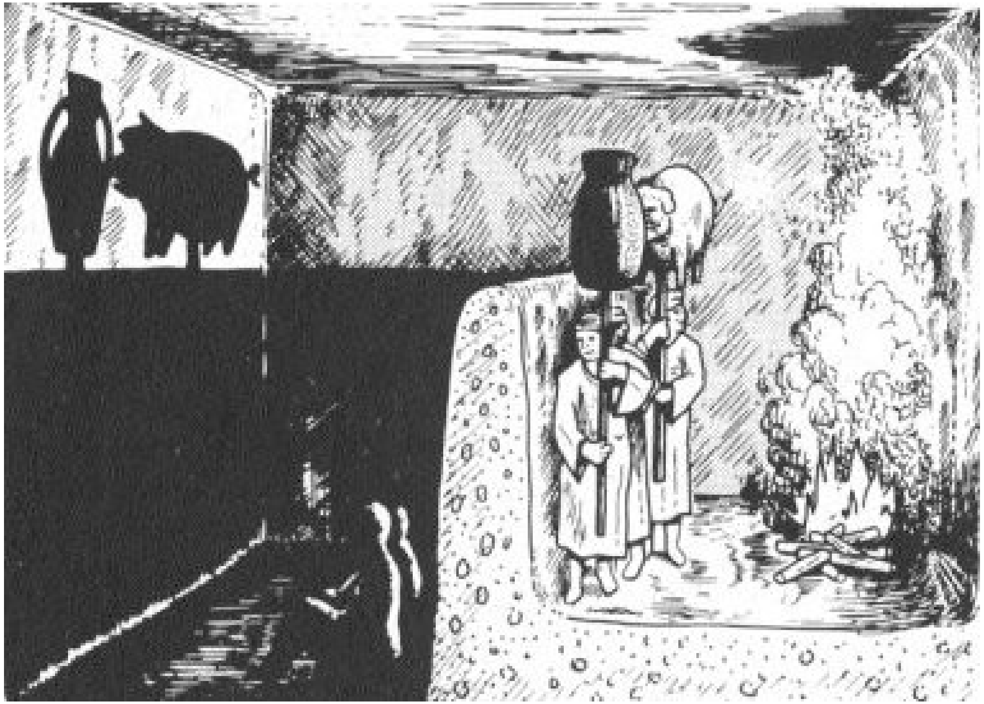
- La probabilité de se tromper en rejetant l'égalité du nombre moyen de chiens possédés quel que soit le sexe du chef de ménage, est inférieure à 1 pour 10 000

La logique Platonnicienne du test

- Échantillon (le phénomène, l'empirique / ce dont on dispose) versus population (le noumène, l'idéal, le théorique, le benchmark)
- Échantillon empirique vu comme une réalisation empirique d'un échantillon tiré dans une population théorique
- On se dote d'une loi de probabilité de distribution des grandeurs dans des échantillons théoriques tirés dans une population théorique
- Permet de dire si on prend des risques à affirmer des choses sur la base d'un échantillon empirique

Dans la caverne de Platon…

- Eh Socrate! Cette amphore semble plus grande que ce cochon!
- Mais est-ce vrai ? Nous n’avons accès qu’au monde imparfait des réalités empiriques [échantillons] et pas au vrais idées [population totale]. Faisons un test statistique !
- C’est à dire?



Dans la caverne de Platon 2

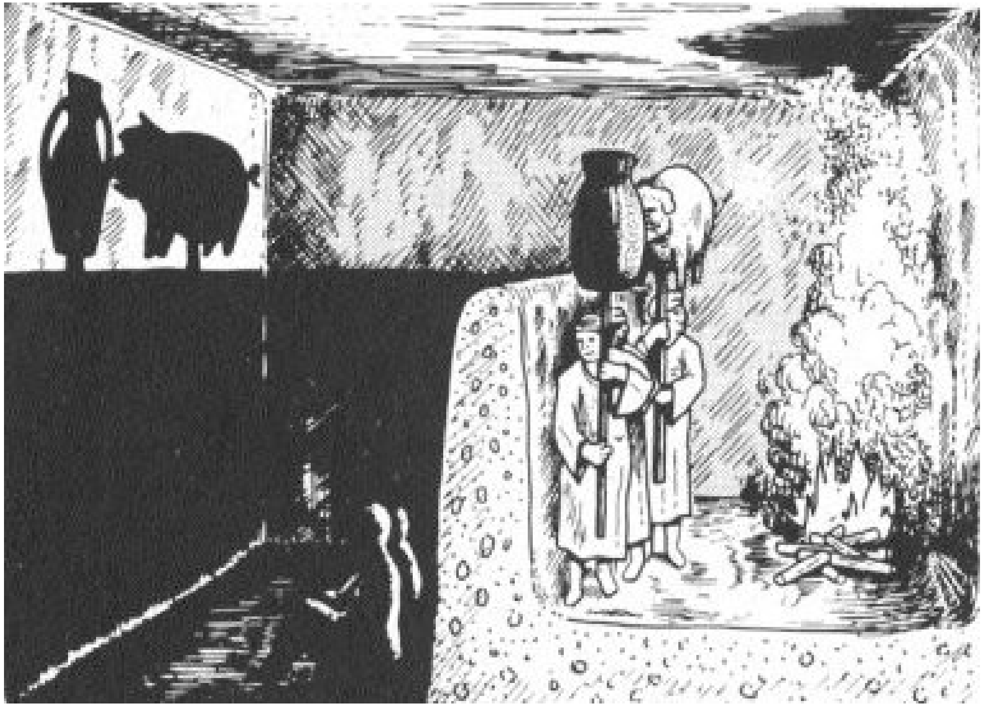
— Commençons par une hypothèse contrariante. Imaginons un monde où l'idée de cochon est aussi grande que l'idée d'une amphore.

— Oui, Socrate, je peux!

— Imaginons que la distribution de probabilité de représentation de ces idées (i.e. la taille de ombres)

— Alors?

— Alors on peut calculer la probabilité de générer des représentations avec des différences aussi grandes que celle qu'on voit sur le mur.

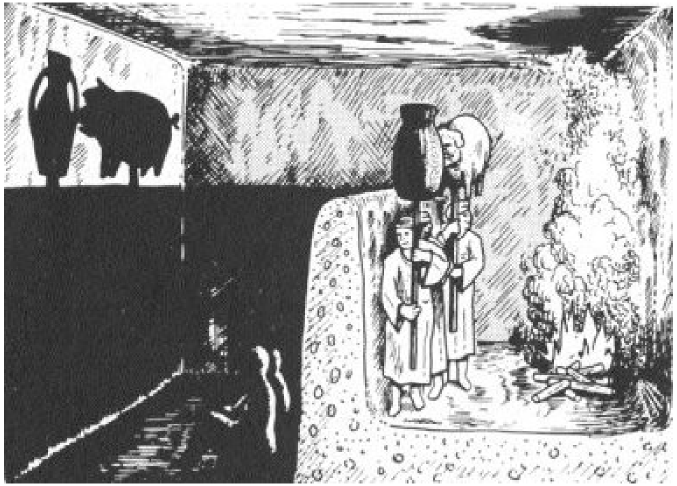


Dans la caverne de Platon 3…

… Et si la proba est faible (<5%). Les différences que nous voyons sont très peu compatibles avec cette hypothèse. Alors il doit y avoir une erreur. On peut alors considérer notre hypothèse contrariante comme erronée Et on peut la rejeter!

— Super! Cool! … Mais on n'a pas confirmé mon jugement.

— C'est vrai! Mais on a rejeté l'argument contraire.



— Et si la probabilité est élevée?

— Eh bien on peut pas la rejeter!

— Peut-on alors l'accepter?

— Non, plus!

— Ah Zut!

Pourquoi tester ?

- En statistiques, en général on travaille sur des échantillons et non sur la population complète.
- Même quand on travaille sur la population complète, on considère parfois que la population complète n'est qu'une réalisation de l'univers infini des possibles.
- Par conséquent les grandeurs observées sur l'échantillon (moyenne, écart-type, etc.) représentent plus ou moins bien ces mêmes grandeurs à l'échelle théorique, i.e. à l'échelle de la population [ou de l'univers des possibles].
- On fait alors des hypothèses sur ces grandeurs au niveau théorique et on mesure la compatibilité des réalisations empiriques avec les hypothèses théoriques.

Le contraire de ce que l'on pense

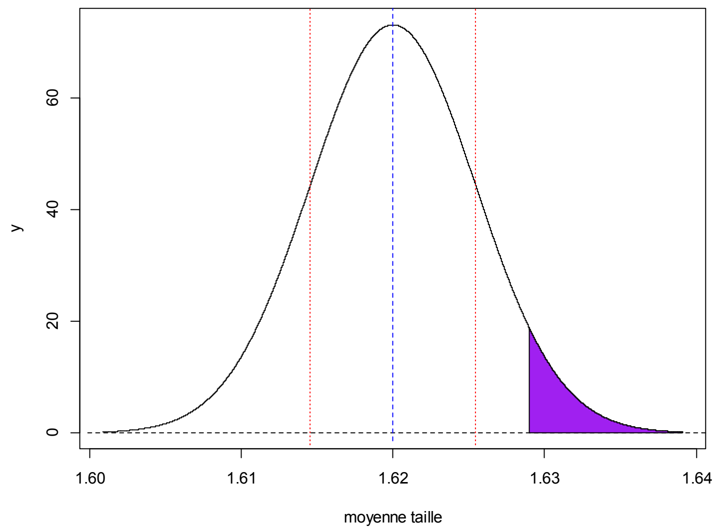
- On établit une hypothèse sur le paramètre au niveau théorique : On l'appelle H1, ou hypothèse alternative.
  - En général une différence (**≠)** ou une inégalité (< ou >)
    - Exemple1. H1: $m1_{théo} \neq m2_{théo}$
    - Exemple2. H1: $m1_{théo} > m2_{théo}$
- Le contraire de ce que l'on pense : c'est H0, l'hypothèse nulle, celle que l'on veut "nullifier" [to nullify] !
  - C'est une égalité (**=)** ou une inégalité (< ou >)
    - Exemple1. H0: $m1_{théo} = m2_{théo}$
    - Exemple2. H0: $m_{théo}1 < m2_{théo}$

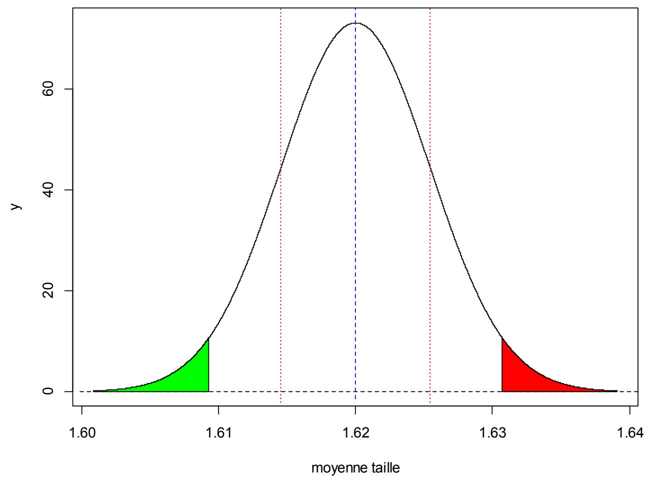
La probabilité que le contraire soit vrai

- Il faut faire des hypothèses sur la distribution théorique (notamment ses paramètres, moyenne et écart-type) dans le cadre de H0. Par exemple, la moyenne théorique est nulle (et le cas échéant l'écart-type théorique est de tant…)
- On se dote d'une loi de probabilité ou d'une table de probabilité mesurant la répartition des écarts autour des paramètres de H0.
- Ensuite on calcule, sous les hypothèses de H0, la probabilité lorsque l'on génère des échantillons (de même taille que celui de notre échantillon empirique) d'obtenir des écarts à l'hypothèse nulle <u>au moins aussi grands</u> que l'écart entre les données empiriques et les paramètres de H0.
- Si cette probabilité est faible (<10%). On dit qu' « on rejette H0 au seuil de x% ». On a alors x% de chance de se tromper en acceptant H1.
- On dit alors généralement que la différence est significative.
- Si cette probabilité est importante (>10%), « on ne peut pas rejeter H0 ». Peut-on alors accepter H0 ? C'est une question discutée.... (Fisher vs Neyman Pearson).

Test unilatéral – test bilatéral

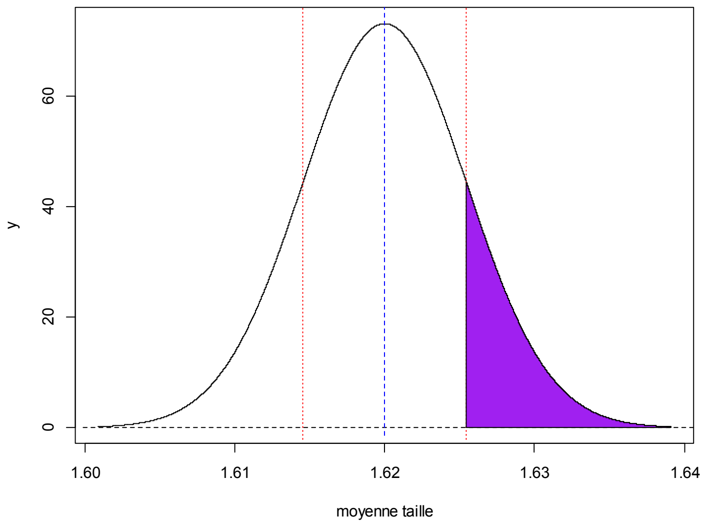
- Les scandinaves sont plus grandes que la moyenne des françaises dans le monde. Échantillon : 121 femmes scandinaves tirées au hasard
- On imagine une population théorique où les femmes scandinaves et françaises ont la même taille et suivent la distribution de taille de la population française
  - Moyenne : 1.62, Ecart-type : 0.06
  - Erreur-type pour 4 personnes = 0.06/Racine(121)=0.0055
- En haut on pose H0: m <1,62 et à droite H0: m =1,62.
  - Les valeurs empiriques permettant de rejeter H0 au seuil de 5% sont marquées par les zones en violet.
  - En haut, une moyenne empirique sur un échantillon de 100 personnes supérieure à 1,629 permet de rejeter H0, en revanche des valeurs très faibles ne le permettent pas.
- En bas on rejettera H0 pour une moyenne empirique soit supérieure à 1,631 soit inférieure 1,609. D’un côté donné de la courbe, on n’est plus exigeant pour rejeter H0.

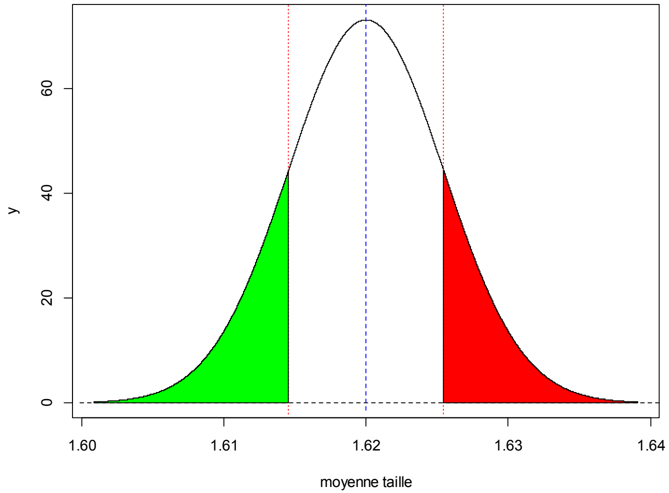




Test unilatéral – test bilatéral

- Plutôt que de regarder un seuil donné comme 5%, on peut se calculer la probabilité exacte de se tromper en rejetant H0 pour un méch. empir.
  - Exemple un méch. empir. =1.6255
- Test unilatéral : H0: m <1,62
  - On calcule la probabilité qu'un échantillon sous H0 ait une moyenne au moins supérieure à la valeur empirique, 1.6255.
  - On calcule la surface représentée par la partie violette qui commence au point 1.6255.
- Test Bilatéral : H0: m. =1,62.
  - On calcule la probabilité qu' un échantillon sous H0 ait un écart absolu à la moyenne supérieur à 1.6255-1.62=0.0055.
  - La surface rouge ne représente alors que la moitié de la probabilité de se tromper et on « ajoute » la partie verte qui est en miroir de l'autre côté de la courbe.





Test bilatéral ou test unilatéral ?

- Le test bilatéral plus exigeant pour rejeter H0.
  - En général :

Probabilité de se tromper en rejetant H0 test unilatéral = 2*Probabilité de se tromper en rejetant H0 test bilatéral

Si on peut rejeter H0 au seuil de 5% pour un test unilatéral, on ne peut rejeter qu'au seuil de 10% dans le cadre d'un test bilatéral.

  - La pratique conduit généralement à favoriser l'hypothèse nulle d'égalité plutôt que celle d'inégalité (même si on a une inégalité en tête).
  - Raison exigence d'une part + détecter toutes sortes d'écarts au-delà de ceux préconçus.
- NB : Le test unilatéral plus exigeant pour « accepter » H0.

Test de Student

- Developpé par William Sealy Gosset dans *The application of the law of error to the work of the Brewery* (1904, *Guinness* internal note)
- Autorisé par Guinness de publier sous le pseudonyme : "Student"
- Test pour tester les égalités et les inégalités de moyennes
- Très utile à connaître. C'est le même test qui sert pour tester la nullité des paramètres dans une régression

Le test de student (suite)

- On fait des hypothèses théoriques pour H0 : sous H0, la moyenne pour la population est μ et l'écart-type σ.
- Le TCL nous dit que la distribution des moyennes d'échantillons d'une taille N se rapproche de la loi normale
- On se dote d'une loi proche de la loi normale : la loi de Student
- En tirant des échantillons de même taille que l'échantillon empirique dans une loi de student de paramètres H0, on regarde la probabilité d'engendrer des écarts à H0 au moins aussi grands que ceux observés empiriquement entre les paramètres empiriques et ceux de l'hypothèse H0.

Les éléments du test de student quand on compare un échantillon et des paramètres théoriques d'une population (μ,σ)

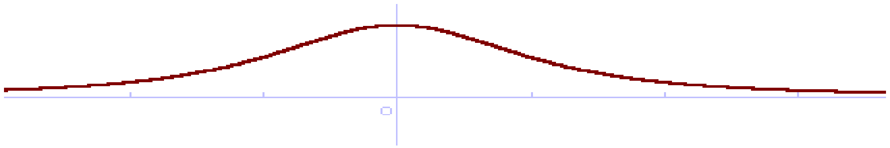
Une mesure particulière des écarts: la statistique de student *t*

$$t = \frac{\overline{X} - \mu}{\sigma\sqrt{\frac{1}{N}}}$$

Le degré de liberté :

DL=*N-1*

La loi de student : Graphe plus aplati que la loi normale, s'en rapprochant quand le DL augmente. Formule de densité particulièrement barbare.



Densité de la loi de Student à deux degrés de liberté

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}},$$

Un calcul de probabilité

- $P_{Student\ (N-1)}(X<-t \text{ et } X>t) = 2\text{-}2*P_{Student\ (N-1)}(X<=t)$
  $=2\text{-}2*T_{N-1}(t)$

- Problème 1 : on ne connaît pas l'écart-type de la population σ. On remplace alors par l'écart type empirique de l'échantillon s.
- Problème 2 : on ne s'intéresse généralement pas à l'écart à une moyenne de la population qu'on ne connaît pas mais à la différence de moyenne de deux échantillons.

Comparaison de deux moyennes

- Hypothèses :
  - Normalité de la distribution d'échantillonnage
  - Égalité des variances. (En principe il faudrait le tester...)
- Mise en oeuvre
  - Variance commune : $$s_{com}^2 = \frac{s_1^2 \times (N_1 - 1) + s_2^2 \times (N_2 - 1)}{N_1 + N_2 - 2}$$
  - Erreur-type commune $$= \sqrt{s_{com}^2 \times \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$
  - Statistique t : $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_{com}^2 \times \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$
  - Le degré de liberté : DL=N_1+N_2-2

Exemple : le nombre de chiens possédés par les ménages hommes et les ménages femmes

CSEX	Nb	Moyenne	Écart-type
Homme	4579	0,5628	0,8846
Femme	1303	0,2678	0,6292

Différence de moyenne : 0,2949

Variance commune :

$$Vcom=[(4579-1)*0,8846^2+(1303-1)*0,6292^2]/[4579+1303-2] = 0,6969$$

Écart-type commun : $\sqrt{Vcom}=0,8349$

Erreur-type commune : $0,8349* \sqrt{[(1/4597)+(1/1303)]}=0,0262$

$t = DifMoy/ErreurType = 11,25$

DL : 4579+1303-2=5880

Proba = $2-2*T_{5880}(11,25)=4,58*10^{-29}$

Sous Excel

	A	B	C	D	E
1					
2					
3	**Étiquettes de lignes**	**Moyenne de nb_chien**	**Var de nb_chien**	**Écartype de nb_chien**	**Nombre de nb_chien**
4	femme	0.267843438	0.395946728	0.62924298	1303
5	homme	0.562786635	0.782590398	0.884641395	4579
6	**Total général**	**0.497449847**	**0.711862225**	**0.843719281**	**5882**
7					
8					
9	Étiquettes de lignes	Moyenne de nb_chien	Var de nb_chien	Écartype de nb_chien	Nombre de nb_chien
10	femme	0.268	0.396	0.629	1303
11	homme	0.563	0.783	0.885	4579
12	Total général	0.497	0.712	0.844	5882
13					
14	Calcul		Formules		
15	Différence homme-femme	0.295	=B11-B10		
16	Variance commune	0.697	=(C10*(E10-1)+C11*(E11-1))/(E10+E11-2)		
17	Statistique T	11.252	=ABS(C14)/RACINE(C15*(1/E10+1/E11))		
18	DDL	5880	=E10+E11-2		
19	Test de student	4.487E-29	=LOI.STUDENT(C16;C17;2)		
20					
21					

Que se passe-t-il si les variances ne sont pas « égales » ?

- Test de l'égalité des variances :
  - Test de Fisher :
    - Rapport entre plus grande variance et plus petite.
    - On regarde si ce rapport est différent de 1.
- On ne prend plus la variance commune :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}}$$

- Diminution du degré de liberté selon la formule horrible de Welch Sattherwaite. Le degré de liberté est alors v (arrondi).

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

Le test du chi-deux

- Usage :
  - Test de Student : bon pour les moyennes.
  - On peut utiliser le test de student pour comparer deux proportions…
  - Mais il existe aussi un autre outil pour les proportions (surtout quand il y en a plus de deux) et en particulier les tableaux croisés : Le Chi-deux
- Qu'est-ce qu'on teste dans un tableau croisé ?
  - On veut montrer généralement que certaines catégories sont plus ceci que telles autres (en gros que les pratiques ne sont pas les mêmes).
  - H0 : On ne teste pas à proprement parler une égalité ou une inégalité comme dans le test de Student mais le problème complexe suivant :
    - la répartition observée des réponses a la même répartition que celle d'un tableau fictif où les deux variables croisées seraient indépendantes.

Un test du chi-deux. La possession de chien(s) dépend-elle de l'habitat ?

Fréquence Pourcentage Pourct. en ligne Pourct. en col.	chien	pas de chien	Total
Immeuble	458 7.80 17.51 22.16	2157 36.74 82.49 56.70	2615 44.54
Maison	1609 27.41 49.42 77.84	1647 28.05 50.58 43.30	3256 55.46
Total	2067 35.21	3804 64.79	5871 100.00

Fréquence manquante = 11

Statistique	DF	Valeur	Proba.
Khi-2	1	647.0328	<.0001

- Lecture littérale du test : *On a moins d'une chance sur 10 000 de se tromper en rejetant l'idée que la possession ou non de chien(s) ne dépend pas du type d'habitat.*
- Lecture moins littérale : La situation est significativement différente (au seuil 1/10 000) de l'indépendance.
- Ou encore : Le taux de possession de chien est significativement supérieur dans une maison que dans un immeuble

Fondement du test de chi-deux

- Loi multinomiale (loi des proportions) : tirage de n boules avec remise dans une urne avec p1 boules de couleur 1, p2 boules de couleur 2, p3 de couleur 3… pm de couleur m. Probabilité d'avoir k1 boules de couleur 1, k2 de couleur 2, … km de couleur m.
- Le carré (pondéré) de l'écart entre la proportion des boules tirées d'une couleur donnée dans le tirage et leur proportion globale dans l'urne tend vers une loi de chi-deux lorsque n tend vers l'infini
- H0 pour le chi-deux : la répartition des proportions observées suit une répartition théorique abstraite donnée.
  - H0 la plus commune pour un chi-deux dans un tableau croisé : la répartition des proportions observées suit la répartition des proportions d'un tableau (abstrait) où les deux phénomènes croisés sont indépendants.
- On calcule la probabilité qu'un échantillon de taille n tiré dans une urne ait un écart au moins aussi grand à la répartition théorique que celui mesuré entre l'échantillon empirique et la distribution théorique.

Tableau empirique observé et tableau à l'indépendance

- Voici notre tableau observé.

	chien	pasdechien	Total
Immeuble	458	2157	2615
Maison	1609	1647	3256
Total	2067	3804	5871

- Construisons le tableau fictif à l'indépendance

	chien	pasdechien	Total
Immeuble	35%	65%	100%
Maison	35%	65%	100%
Total	**35%**	**65%**	**100%**

	chien	pasdechien	**Total**
Immeuble	45%	45%	**45%**
Maison	55%	55%	**55%**
Total	100%	100%	**100%**

- Taux de possession de chien et habitat sont ici indépendants si le taux de possession de chien ne dépend pas de l'habitat.
- C'est le donc tableau où sur chaque ligne on retrouve le même pourcentage en ligne que le pourcentage en ligne total ou de manière équivalente, celui où on trouve sur chaque colonne le % en colonne total.

	chien	pasdechien	Total
Immeuble	=2615*35%	=45%*3804	2615
Maison	=3256*2067/5871	=3804*3256/5817	3256
Total	2067	3804	5871

- En appliquant aux effectifs marginaux (produit du % en ligne total avec l'effectif total en ligne ou de manière équivalente du % en colonne total à l'effectif total en colonne), on trouve que l'indépendance est bien le produit des marges divisé par l'effectif total.

Formule pour des tableaux croisés

- Statistique (ou distance) du chi-deux :
  - Somme des carrés des écarts entre l'effectif observé et l'effectif théorique, divisés par l'effectif théorique.

$$D^2 = \sum_{i=1}^{p} \sum_{i=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} = n \sum_{i=1}^{p} \sum_{i=1}^{q} \frac{\left(p_{ij} - p_{i.} p_{.j} \right)^2}{p_{i.} p_{.j}}$$

- Degré de liberté
  - DL (ou DF) = (p-1)*(q-1) où p nombre de catégories en ligne et q nombre de catégories en colonne (sans compter les lignes et les colonnes de total ou d'intitulés !)

- Le test du chi-deux est un test unilatéral (et non généralement bilatéral comme le test de student)

- PChi-deux[(p-1)*(q-1)] $(X>D^2)=1-\chi^2(p-1)(q-1)[D]$

Mise en oeuvre simple sous Excel

- 1. Tableau des effectifs observés
- 2. Tableau des effectifs théoriques à l'indépendance : produit des marges divisé par l'effectif
- 3. Tableau des écarts entre effectifs observés et théoriques
- 4 . Tableau des écarts au carré divisés par l'effectif théorique
- 5. Statistique du chi-deux : D^2= Somme du tableau 4
- 6. Degré de liberté: dl = (p-1)*(q-1)
- 7. On calcule grâce à la loi de probabilité du chi-deux, la probabilité d'être supérieur à la statistique D^2 avec un tel degré de liberté. Dans la cellule :
  - =Loi.Khideux (D^2; dl) ou =1 – Loi.Khideux.N(D^2; dl;1)
- C'est la probabilité de se tromper si on rejette l'hypothèse H0 d'indépendance. Si cette probabilité est petite <10%. On peut prendre ce risque.

Excel

	A	B	C	D	E	F	G	H
1								
2	**Nombre de a_chien**	**Étiquettes de colonnes**						
3	**Étiquettes de lignes**	**1. Oui**	**2. Non**	**Total général**				
4	a.Maison	1609	1647	3256				
5	b.Immeuble	458	2157	2615				
6	**Total général**	**2067**	**3804**	**5871**				
7								
8	Tableau empirique							
9	Nombre de a_chien	Étiquettes de colonnes						
10	Étiquettes de lignes	1. Oui	2. Non	Total général				
11	a.Maison	1609	1647	3256				
12	b.Immeuble	458	2157	2615				
13	Total général	2067	3804	5871				
14								
15	Tableau indépendance (effectif théorique)							
16	Nombre de a_chien	Étiquettes de colonnes						
17	Étiquettes de lignes	1. Oui	2. Non	Total général		Formule coin haut gauche		
18	a.Maison	1146.338273	2109.7	3256		=B$14*$D12/D14		
19	b.Immeuble	920.6617271	1694.3	2615				
20	Total général	2067	3804	5871				
21								
22								
23	Tableau des écarts							
24	Nombre de a_chien	Étiquettes de colonnes						
25	Étiquettes de lignes	1. Oui	2. Non	Total général		Formule coin haut gauche		
26	a.Maison	-462.6617271	462.66	0		=B18-B11		
27	b.Immeuble	462.6617271	-462.7	0				
28	Total général	0	0	0				
29								
30								
31	Tableau carré des écart pondérés par l'effectif théorique							
32	Nombre de a_chien	Étiquettes de colonnes						
33	Étiquettes de lignes	1. Oui	2. Non	Total général		Formule coin haut gauche		
34	a.Maison	186.7301117	101.46	0		=B26^2/B11		
35	b.Immeuble	232.5021965	126.34	0				
36	Total général	0	0	0				
37								
38	Chi2	647.0328289				=SOMME(B34:D36)		
39								
40	DDL	1						
41								
42	Test chi deux	0.00000000000000000000				=LOI.KHIDEUX(B38;B40)		
43								

Conclusion

- Les tests reposent fortement sur l'effectif.
  - Plus l'effectif est important, plus on a confiance dans nos données
  - Quand l'effectif est très important (millions), toute différence, même infime est significative…
  - Biais de publication
  - Biais d'estimation

Savoir juger des tests et des critères de significativité avec doigté en tenant compte de ce fait !!!

Résumé de l'approche

- Hypothèses théoriques
  - H_1 (ce qu'on pense) : $m_1 \neq m_2$ ou dépendance entre deux variables catégorielles
  - H_0 (ce qu'on va réfuter) : $m_1 = m_2$ ou indépendance de deux variables catégorielles
- Données empiriques
  - Échantillon empirique de taille N
  - La *statistique* empirique (*t* de student, D Chi2) : mesure de l'écart dans l'échantillon empirique entre H_0 et les valeurs empiriques.
- Projections théoriques
  - Population abstraite suivant les paramètres de H_0.
  - Probabilité que des échantillons de taille N de la population abstraite aient des statistiques d'écart à H_0 supérieures à la *statistique* empirique (calcul avec Loi de student, loi de Chi2)
- Rejeter H0 si $P<0{,}05$ (ou éventuellement $P<0{,}1$)

Appendices R

Le test de Student sous R

- `t.test(he$chienn~he$CSEX)`

```
Welch Two Sample t-test
data:  he$chienn by he$CSEX
t = 13.536, df = 2916.093, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2522189 0.3376675
sample estimates:
mean in group 1 mean in group 2
      0.5627866       0.2678434
```

  - Par défaut l'égalité des variances n'est pas acceptée
- `t.test(he$chienn~he$CSEX,var.equal=TRUE)`
- Test de la variance :
  - `var.test(he$chien~he$CSEX)`

Khi-deux sous R

- t<-table(he$CSEX,he$achat)
- chisq.test(t)

Pearson's Chi-squared test with Yates' continuity correction

data: t

X-squared = 29.7695, df = 1, p-value = 4.866e-08

  - Attention chisq.test utilise la correction de continuité de Yates pour les tableaux 2*2, qui est discutée.

- Sans correction :

chisq.test(t, correct=FALSE)