Statistiques fondamentales

M. Clertant [1]

Statistiques biomédicales

1. clertant@math.univ-paris13.fr, Lab. Analyse, Géométrie et Applications

Plan

Echantillon, statistique et estimation :

(i) Un n-échantillon de X est un vecteur $(X_1, X_2, \ldots, X_n) = X_1^n$ de n variables aléatoires indépendantes et de même loi que X.

(ii) Une réalisation de l'échantillon est un vecteur $(x_1, x_2, \ldots, x_n) = x_1^n$ de valeurs observées.

(iii) Une statistique de l'échantillon est une variable aléatoire $\phi(X_1, X_2, \ldots, X_n)$, où ϕ est une application de $\mathbb{R}^n$ dans $\mathbb{R}^p$.

(iv) Un estimateur $\hat{\theta}_n$ de θ est une statistique. Une estimation est une réalisation d'un estimateur.

Expérience statistique :

Un modèle statistique est une famille de loi de probabilité $\{\mathbb{P}_\theta, \theta \in \mathbb{R}\}$ dont on suppose qu'elle puisse contenir la vraie loi générant le phénomène aléatoire étudié, X.

- Si $X \sim \mathbb{P}_\theta$ alors le n-échantillon X_1^n suit la loi produit $\mathbb{P}_\theta^{\otimes n}$
- Un modèle est identifiable si : $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2$ ($\theta \mapsto \mathbb{P}_\theta$ est injective).

Exemples de modèles : Loi normale $\mathcal{N}(\mu, \sigma^2)$, loi de Bernoulli $\mathrm{Be}(\theta)$, loi exponentielle $\mathcal{E}(\theta)$

Biais et Risque quadratique

- Le biais de l'estimateur $\hat{\theta}_n$ de θ, noté $b_\theta(\hat{\theta}_n)$, est :

$$b_\theta(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta.$$

Un estimateur sans biais a un biais nul pour tout $n \in \mathbb{N}^\star$ et un estimateur asymptotiquement sans biais vérifie : $\lim_{n \to +\infty} E[\hat{\theta}_n] = \theta$.

- Le risque quadratique de l'estimateur $\hat{\theta}_n$ de θ est :

$$R(\theta, \hat{\theta}_n) = \mathbb{E}\left((\hat{\theta}_n - \theta)^2\right) = b_\theta(\hat{\theta}_n) + \text{Var}(\hat{\theta}_n).$$

On dit que $\hat{\theta}_n$ est un meilleur estimateur de θ que $\tilde{\theta}_n$ si :

1. pour tout θ, on a : $R(\theta, \hat{\theta}_n) \leq R(\theta, \tilde{\theta}_n)$,
2. il existe au moins un θ tel que : $R(\theta, \hat{\theta}_n) < R(\theta, \tilde{\theta}_n)$.

Remarque : Si $\hat{\theta}_n$ et $\tilde{\theta}_n$ sont deux estimateurs sans biais de θ. Il suffit de comparer les variances $\text{Var}(\hat{\theta}_n)$ et $\text{Var}(\tilde{\theta}_n)$. Le terme "meilleur" est alors remplacé par le terme "plus efficace".

Convergence d'estimateur :

(i) Un estimateur $\hat{\theta}$ est dit convergent si la suite $(\hat{\theta}_n)_{n\in\mathbb{N}}$ converge en probabilité vers θ_T, ce qui signifie que, pour tout $\epsilon > 0$, on a :

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta_T| > \epsilon\right) = 0.$$

(ii) On parle d'estimateur fortement convergent lorsqu'on a convergence presque sûre vers θ_T, ce qui signifie que :

$$\mathbb{P}\left(\lim_{n\to+\infty} \hat{\theta}_n = \theta_T\right) = 1.$$

On dit aussi que l'estimateur converge presque sûrement.

Convergence en loi :

On dit que $(X_n)_n$ converge en loi vers X, noté $X_n \xrightarrow{\mathcal{L}} X$, si pour toute fonction f continue bornée, on a : $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$.

Estimateur ponctuels classiques

Soit X_1^n un n-échantillon dont la loi a une espérance m et une variance σ^2.

La moyenne empirique, construit à partir de la loi empirique, est :

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

$\overline{X}_n$ est sans biais et fortement convergent.

L'estimateur de la variance empirique, construit à partir de la loi empirique, est :

$$\overline{S_n^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2.$$

$\overline{S_n^2}$ est biaisé, asymptotiquement sans biais et fortement convergent.

L'estimateur de la variance empirique corrigée est :

$$\widehat{S_n^2} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2.$$

$\widehat{S_n^2}$ est fortement convergent et sans biais.

Théorèmes classiques

Soit X une v.a.r et X_1^n un n-échantillon de X.

La loi des grands nombres (forte)
Si, X admet une espérance, $\mathbb{E}(X) = m$, alors on a : $\lim_{n\to\infty} \overline{X}_n = m,$ p.s.

Si X admet une espérance $\mathbb{E}(X) = m$ et une variance $\mathrm{Var}(X) = \sigma^2$,
alors on a (1) et (2) :

(1) L'inégalité de Bienaymé-Tchebychev : $\mathbb{P}(|X - m| \geq a) \leq \dfrac{\sigma^2}{a^2}$.

Conséquence :Si $\hat{\theta}_n$ est un estimateur de θ sans biais tel que : $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0,$
alors $\hat{\theta}_n$ est convergent.

(2) Le théorème centrale limite (normalité asympto.) : $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$

Le théorème de Slutsky : Si X_n converge en loi vers X et si Y_n converge en probabilité vers une constante c alors : $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, c).$

Conséquence : Sous les hypothèses du TCL, on a : $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\sqrt{\widehat{S_n^2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$

Remarque : Autre inégalités de concentration : Markov, Chernoff, Hoeffding, Bernstein ... ; autre théorèmes utiles : Cochran, loi du logarithme itérée, Donsker ...

Plan

Intervalle de confiance

Une région de confiance de θ de niveau $1-\alpha$ est une région $C(X_1^n) \subset \Theta$ (dépendant du n-échantillon X_1^n) telle que, pour tout $\theta \in \Theta$:

$$\mathbb{P}(\theta \in C(X_1^n)) \geq 1-\alpha.$$

On parle ici de région par excès. Dans le cas d'une égalité, il s'agit d'une région de confiance exacte.

En pratique, quand θ est dans $\mathbb{R}$, on s'intéresse le plus souvent à une région de confiance non-trouée c'est à dire un intervalle bilatéral ou unilatéral.

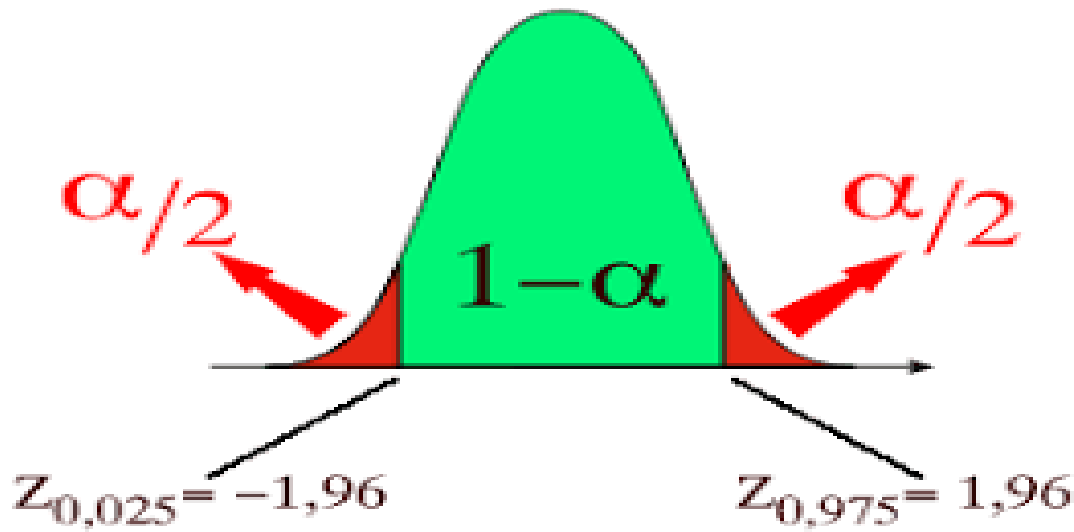


Figure – Intervalle de confiance de niveau 95% pour une statistique suivant une loi normale $\mathcal{N}(0,1)$.

Exemple pour une proportion

On estime une proportion p par sa moyenne empirique $\overline{X}_n$.

Un intervalle de confiance de niveau $1-\alpha$ de p est un intervalle $[\overline{X}_n - a, \overline{X}_n + b]$ tel que :

$$\mathbb{P}(\overline{X}_n - a \leq p \leq \overline{X}_n + b) = 1 - \alpha.$$

Or,

$$\mathbb{P}(\overline{X}_n - a \leq p \leq \overline{X}_n + b) = \mathbb{P}\left(-b\frac{\sqrt{n}}{\sqrt{p(1-p)}} \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \leq a\frac{\sqrt{n}}{\sqrt{p(1-p)}} \right).$$

D'après le TCL, on obtient un intervalle de confiance symétrique en choisissant $a = -b = q_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$, où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$.

Comme on ne connaît pas p dans cet expression, on peut le remplacer par son estimation ponctuelle $\overline{X}_n$ (Th. de Slutsky). Un intervalle de confiance approché de p au niveau $1-\alpha$ est :

$$\left[\overline{X}_n - q_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + q_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \right],$$

Moyen alternatif : le bootstrap

Lorsque l'on n'a pas accès à la loi de l'estimateur $\hat{\theta}_n$, il est possible d'utiliser la méthode du bootstrap (*"Pull yourself up by your own bootstraps."*).

Bootstrap : On se base uniquement sur l'échantillon X_1^n et sur sa loi empirique $e(X_1^n) = \sum_{i=1}^n \delta_{X_i}$. On génère un grand nombre de jeux de données de taille n à partir de cet échantillon (tirage avec remise (iid)).
Pour chacun des m jeux de données, on évalue l'estimateur :
$\hat{\theta}_{n,1}^m = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,m})$. Cela nous donne accès à une estimation de la distribution de l'estimateur -> intervalle de confiance.

```
> library(boot)
> x <- c(52, 10, 40, 104, 50, 27, 146, 31, 46)
> mean(x)
 56.22222
> bb <- boot(data = x, statistic = function(x, index)
mean(x[index]), R = 1000)
Bootstrap Statistics :
original     bias     std. error
 56.22222  -0.7888889   13.31232
```

Idée : La fonction de répartition empirique de l'échantillon X_1^n converge vers la vraie fonction de répartition (Th. de Glivenko-Cantelli) et donc la fonction de répartition empirique de l'échantillon bootstrap de l'estimateur, $\hat{\theta}_{n,1}^m$, devrait converger (quand n et m converge vers l'infini) vers la fonction de répartition de $\hat{\theta}_n$.

Les tests, analogie avec un procés

Sur la base des observations, il s'agit de répondre à des questions dans lesquelles il est possible d'identifier clairement des alternatives. Les réponses ne sont pas "oui" ou "non", mais dépendent d'un seuil de signification α.

- L'homéopathie est-elle plus efficace qu'un placebo au seuil de signification α?
- Cette échantillon peut il être considéré de loi normale au seuil α ?
- La personne est-elle coupable au seuil α?

- H_0 : "la personne jugée est innocente."
- H_1 : "la personne jugée est coupable."

		Décision	
		Acquittement	Condamnation
Réalité	Innocent (H_0)	Bonne décision	Erreur de 1ère espèce
	Coupable (H_1)	Erreur de 2nd espèce	Bonne décision

Les magistrats/statisticiens espèrent montrer la culpabilité. Il y a présomption d'innocence. Y a-t-il assez d'évidence de la culpabilité ?

- Si oui, condamnation ... on rejette H_0.
- Sinon, acquittement ... on ne rejette pas H_0.

"Combien d'innocents courraient encore, si il n'y avait pas d'erreur judiciaire" de première espèce. (Desproges)

Les tests statistiques

- H_0 : l'hypothèse nulle.
- H_1 : l'hypothèse antagoniste.

		Décision	
		Non-rejet de H_0	Rejet de H_0
Réalité	H_0 vraie	Bonne décision	Erreur de 1ère espèce
	H_1 vraie	Erreur de 2nd espèce	Bonne décision

Les statisticiens regardent si il est possible de rejeter H_0 sur la base de l'échantillon au seuil de signification α.

"On souhaite que la probabilité de l'erreur de première espèce soit inférieure à α."

Un peu d'histoire :
- R. Fisher est le premier statisticien à introduire les tests. Il n'utilise qu'une seule hypothèse H_0. Il se pose la question suivante : Sur la base de l'échantillon, peut-on rejeter l'hypothése H_0 au seuil significatif α ?

- J. Neyman et E. Pearson ne sont pas d'accords avec cette formulation, ils préfèrent définir deux hypothèse antagonistes H_0 et H_1.

L'histoire a tranchée : Il existe des zones d'acceptation de H_0 et H_1 (comme Neyman et Pearson), mais on accepte jamais H_0 ! On ne rejette pas H_0 (comme Fisher).

Définition d'un test statistique

Soit $X_1^n = (X_1, \ldots, X_n)$ un n-échantillon de X, une v.a.r de loi $\mathcal{L}$.
Soient $\mathcal{L}_0$ et $\mathcal{L}_1$, deux ensembles disjoints de lois de probabilité.
On considère les hypothèses antagonistes :

$$H_0 = \{\mathcal{L} \in \mathcal{L}_0\} \text{ et } H_1 = \{\mathcal{L} \in \mathcal{L}_1\}.$$

Un test de l'alternative H_0 contre H_1 est une statistique du n-échantillon, $T(X_1^n)$, qui ne prend que les valeurs 0 et 1 :

$$T(X_1^n) = \begin{cases} 1 & \text{et la décision est " } H_1 \text{ est vraie" (Rejet de } H_0 \text{),} \\ 0 & \text{et la décision est " } H_0 \text{ est vraie" (Non-rejet de } H_0 \text{).} \end{cases}$$

Le risque de première espèce du test T associée à la loi $\mathcal{L} \in \mathcal{L}_0$ est :

$$\alpha_{\mathcal{L}} = \mathbb{P}_{\mathcal{L}}(T(X_1^n) = 1) = \mathbb{P}(\text{"Rejet de } H_0\text{"} \mid X \sim \mathcal{L}).$$

Le seuil de signification du test, noté α, est le plus petit nombre tel que, pour tout $\mathcal{L} \in \mathcal{L}_0$:

$$\alpha \geq \alpha_{\mathcal{L}}.$$

Cela correspond au plus grand risque de se tromper lorsque l'hypothèse H_0 est vraie.

Test, intervalle de fluctuation sous H_0 (1) et risque de 2nd espèce (2)

1) Soit un test T basé sur une statistique $S(X_1^n)$ dont on connait la loi sous H_0. Le test est une interprétation de la région de fluctuation (le plus souvent, un intervalle) de $S(X_1^n)$ sous H_0 :

• Les régions d'acceptation de H_0 et de H_1, notées respectivement A_0 et A_1, sont des ensembles complémentaires dans l'ensemble image de S tel que :

$$S(X_1^n) \in A_0 \Leftrightarrow T(X_1^n) = 0 \quad \text{et} \quad S(X_1^n) \in A_1 \Leftrightarrow T(X_1^n) = 1$$

• La p-valeur est la probabilité d'observer, en supposant que "H_0 est vraie", des réalisations de $S(X_1^n)$ au moins aussi extrêmes que celles que nous avons observées.

2) Le risque de seconde espèce du test T associée à la loi $\mathcal{L} \in \mathcal{L}_1$ est :

$$\begin{aligned} \beta_{\mathcal{L}} = & \quad 1 - \mathbb{P}_{\mathcal{L}}(T(X_1^n) = 1) = \mathbb{P}_{\mathcal{L}}(T(X_1^n) = 0) \\ = & \quad \mathbb{P}(\text{"Non-rejet de } H_0\text{"} \mid X \sim \mathcal{L}). \end{aligned}$$

Le risque β, est le plus petit nombre tel que, pour tout $\mathcal{L} \in \mathcal{L}_1$:

$$\beta \geq \beta_{\mathcal{L}}.$$

Note : Le risque β est rarement calculable. Cependant ce risque est intéressant car il apporte des informations sur la qualité du test.

Exercice : Dans la population standard, le taux d'une enzyme dans le sang est modélisée par une variable aléatoire X de loi normale $\mathcal{N}(m = 8, \sigma^2 = 4)$ (en mg/L). On cherche à savoir si un traitement augmente la présence de cette enzyme en moyenne. On considèrera la variance fixée et connue : $\sigma^2 = 4$. On dispose d'un 16-échantillon X_1^{16} relevé après traitement. La moyenne $\overline{X}_{16}$ est 8.89 mg/L.

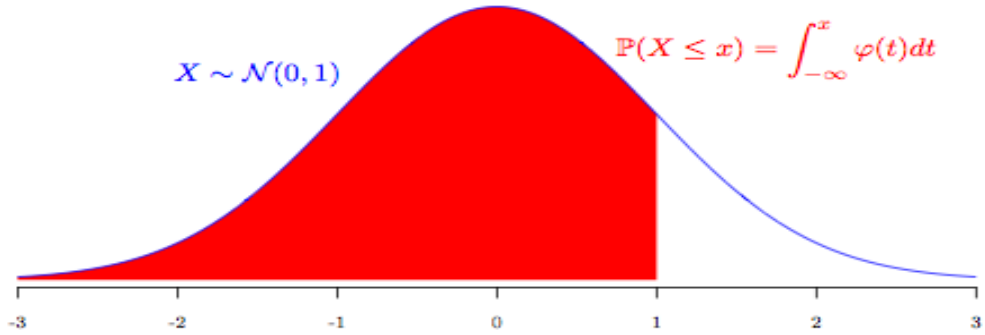
1.Prenant en considération que le développement du traitement est cher, quelles sont les hypothèses à poser ?

2. Sur quelle statistique va être basé le test ? Déterminer la loi de cette statistique.

3. Donner la forme de la région d'acceptation de H_0.

4. Déterminer les régions d'acceptation et la p-valeur pour un test au seuil de signification de 5%.

5. En supposant que les taux d'enzyme après traitement suivent une loi normale $\mathcal{N}(m, 4)$. Déterminer un intervalle de confiance au niveau 95%.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

Un peu de Méthode

Méthode IC :

1. Déterminer un estimateur ponctuel de θ : $\hat{\theta}_n = S(X_1, \ldots, X_n)$.
2. L'estimateur $\hat{\theta}_n$ est une variable aléatoire. Déterminer où approcher sa loi $\mathbb{P}_\theta$. Elle dépend de θ, inconnu.
3. Sous la loi $\mathbb{P}_\theta$, exclure les valeurs les plus extrêmes qu'aurait pu prendre $\hat{\theta}_n$, pour ne garder que les valeurs regroupant une probabilité de $1-\alpha$. On obtient : $\hat{\theta}_n \in C(\theta, \alpha)$, avec probabilité $1-\alpha$ (Intervalle de fluctuation théorique : θ inconnu).
4. Inverser le rôle de $\hat{\theta}_n$ et θ, ce qui donne : $\theta \in C(\hat{\theta}_n, \alpha)$, avec probabilité $1-\alpha$.

Méthode tests :

1. Déterminer deux hypothèses antagonistes : H_0 et H_1. Le niveau α correspond à la probabilité de l'erreur de 1ère espèce (rejeter H_0 à tort).
2. Déterminer une statistique $S(X_1^n)$ dont on connaît la loi sous l'hypothèse H_0. Cette loi est notée $\mathbb{P}_0$.
3. Calculer la probabilité d'obtenir une valeur aussi "extrême" que $S(X_1^n)$ sous H_0. La probabilité obtenue est la p-valeur, noté p_0. Alternativement, il est possible d'étudier les régions d'acceptation (intervalle de fluctuation sous H_0, la loi $\mathbb{P}_0$ est connue).
4. La statistique de test T est : $T(X_1^n) = \mathbb{1}_{\{p_0 \leq \alpha\}}$. Quand T vaut 1, on rejette H_0 ("acceptation" de H_1). Quand T vaut 0, on ne rejette pas H_0.

Dans la suite de ce cours, on adopte la présentation des cours avancés ou articles de statistiques et on ne présente que les points éventuellement communs aux IC et aux tests : "Quelle statistique, quelle loi ?"

Lois asymptotiques pour un échantillon

Les propriétés suivantes permettent l'approximation des lois lorsque n est suffisamment grand. Elles utilisent le TCL et le théorème de Slutsky.

(i) Soit $\overline{X}_n$ la moyenne empirique d'un n-échantillon de loi de Bernoulli $\mathcal{B}(p)$. On a :

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} N(0,1)$$

(ii) Soit $(X_1, \ldots, X_n)$ un n-échantillon de X possédant une espérance et une variance σ^2. On a :

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\widehat{S_n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

où $\widehat{S_n}$ désigne la racine de la variance empirique corrigée : $\widehat{S_n} = \sqrt{\widehat{S_n^2}}$.

Loi du χ^2 pour échantillon de loi normale

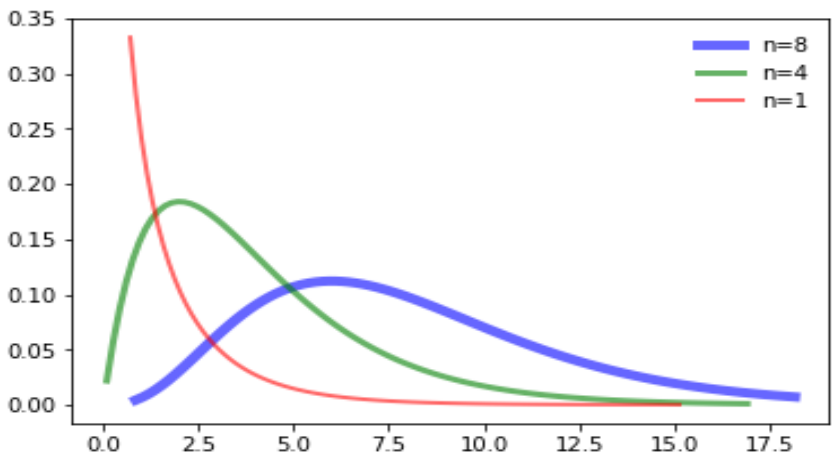
Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi normale $\mathcal{N}(0,1)$. La loi du khi-2 à n degrés de liberté, noté $\chi^2(n)$, et la loi d'une variable aléatoire Z définie comme la somme du carré de n v.a. indépendantes de loi normale $\mathcal{N}(0,1)$:

$$Z = \sum_{k=1}^{n} X_k^2.$$

Une loi du khi-2 a n degrés de liberté a une espérance égale à n et une variance égale à $2n$.

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi normale $\mathcal{N}(0,1)$. Le théorème de Cochran nous donne :

$$\sum_{k=1}^{n} (X_k - \overline{X}_n)^2 \sim \chi^2(n-1) \quad \text{et} \quad \sum_{k=1}^{n} (X_k - \overline{X}_n)^2 \perp\!\!\!\perp \overline{X}_n.$$



Loi de Student pour échantillon de loi normale

Soient X et Y deux variables aléatoires indépendantes suivant respectivement la loi normale centrée réduite et la loi du khi-2 à n degrés de liberté. La variable aléatoire

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

suit la loi de Student à n degrés de liberté, notée $\mathcal{S}(n)$.
La loi de Student n'a pas d'espérance pour $n = 1$ et pas de variance pour $n < 3$.
Dans les autres cas, on a $\mathbb{E}(T) = 0$ et $\mathrm{Var}(T) = n/(n-2)$.

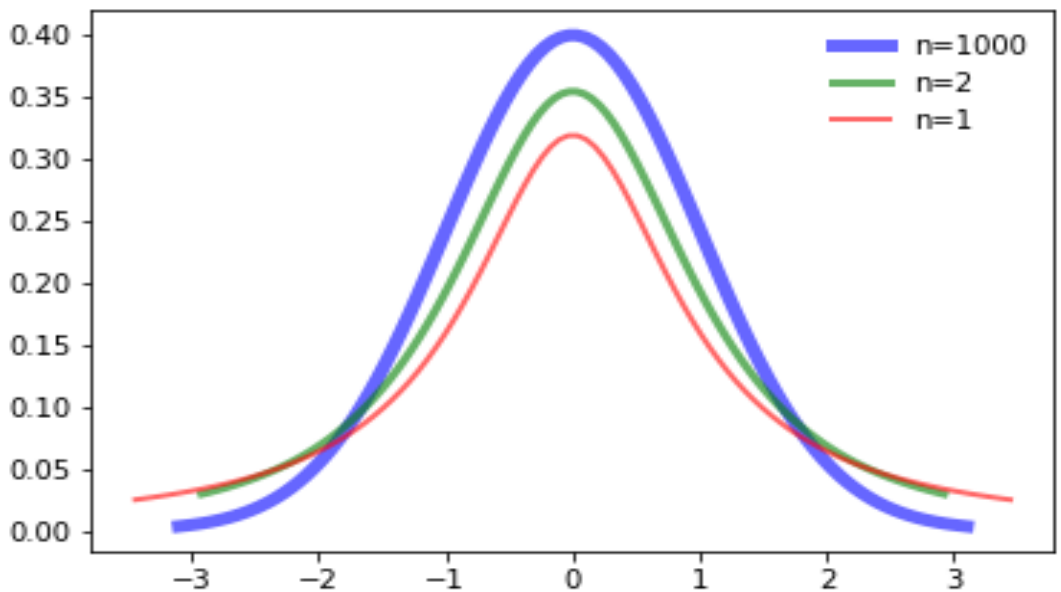


Figure – Densité de la loi de Student pour différents degrés de liberté n.

Lois exactes des statistiques classiques

Soit $(X_1, \ldots, X_n)$ un n-échantillon de $X \sim \mathcal{N}(m, \sigma^2)$. Les propriétés suivantes sont vraies dès lors que $n \in \mathbb{N}^\star$ et l'indice $n-1$ est positif si il se trouve dans l'expression.

(i) Lorsque l'on s'intéresse au paramètre m, on a :

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \sim \mathcal{N}(0,1) \quad \text{et} \quad \sqrt{n}\frac{\overline{X}_n - m}{\widehat{S_n}} \sim \mathcal{S}(n-1)(2),$$

où $\widehat{S_n}$ désigne la racine de la variance empirique corrigée : $\widehat{S_n} = \sqrt{\widehat{S_n^2}}$.

(ii) Lorsque l'on s'intéresse au paramètre σ, on a :

$$n \times \frac{S_n^2}{\sigma^2} \sim \chi^2(n) \quad \text{et} \quad (n-1) \times \frac{\widehat{S_n^2}}{\sigma^2} \sim \chi^2(n-1)(1).$$

avec $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$.

Démontrer (1) et (2).

Comparaison d'échantillons appariés

Les échantillons sont appariés si il est naturel de grouper les observations par paires inter-groupes (une observation dans chaque groupe). Par exemple :

- On dispose des ventes de fromages pour 10 marques sur l'année 2000 et sur l'année 2020. Les français mangent-ils plus de fromages qu'il y a 20 ans ?
- On dispose de la glycémie pour 100 patients avant et après une opération bariatrique. L'opération a-t-elle fait baisser la glycémie ?
- Laquelle des deux crèmes fonctionnent mieux pour traiter l'eczéma ? Protocole de test : tester les deux crèmes sur des endroits symétriques pour un échantillon de patients (une crème sur chaque coude).

C'est le cas le plus favorable. Cela réduit voir supprime l'influence de facteurs exogènes : résultats plus fiables.

	E_1	E_2	E
	x_{1i}	x_{2i}	$d_i=x_{1i}-x_{2i}$
1	x_{11}	x_{21}	$d_1=x_{11}-x_{21}$
2	x_{12}	x_{22}	$d_2=x_{12}-x_{22}$
	...	...	...
n	x_{1n}	x_{2n}	$d_n=x_{1n}-x_{2n}$

Et c'est aussi le cas le plus simple. On ne compare pas les moyennes de E_1 et E_2, on étudie la moyenne de l'échantillon E, voir on la compare à 0 pour tester l'égalité. On utilise donc les résultats connus pour un unique échantillon.

Mais tout n'est pas toujours idéal, et souvent les échantillons sont indépendants (deux groupes de patients différents, des marques de fromage différentes, etc ...)

Comparaison d'échantillons indépendants : contexte

Soit deux échantillons indépendants :

- $X_{n_1}^{(1)} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ de $X^{(1)}$, de moyenne m_1 et de variance σ_1^2,
- $X_{n_2}^{(2)} = (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ de $X^{(2)}$, de moyenne m_2 et de variance σ_2^2.

On considérera que ces échantillons suivent soit des lois quelconque, soit des lois de Bernoulli (résultats approchés), soit des lois normales (résultats exacts).

Pour la différence des moyennes, on utilise la statistique D_n :

$$D_n = \overline{X}_n^{(1)} - \overline{X}_n^{(2)},$$

Pour le rapport des variances, on utilise la statistique R_n :

$$R_n = \frac{\widehat{S_{n_1}^2}^{(1)}}{\widehat{S_{n_2}^2}^{(2)}}.$$

Ces estimateurs sont sans biais et fortement convergents.

Différence de moyenne et variance pour des lois normales

1) Soient deux échantillons indépendants $X_{1,n_1}^{(1)}$ et $X_{1,n_2}^{(2)}$ dont les éléments suivent les lois normales $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$.
Les variances σ_1^2 et σ_2^2 sont connues.

Lorsque l'on s'intéresse à la différence des moyennes, on a :

$$\frac{D_n - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

2) Soient deux échantillons indépendants $X_{1,n_1}^{(1)}$ et $X_{1,n_2}^{(2)}$ dont les éléments suivent les lois normales $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$.

Lorsque l'on s'intéresse au rapport des variances, on a :

$$\frac{\sigma_2^2}{\sigma_1^2} \times R_n \sim \mathcal{F}_{n_1-1,n_2-1}.$$

Exercice : 1. Démontrer ces résultats.
2. Construire le test sur la différence des moyennes pour les hypothèses unilatérales suivantes :

$$H_0 : m_1 \geq m_2 \quad \text{et} \quad H_1 : m_1 < m_2.$$

Différence de moyenne pour des lois normales de variances égales et inconnues

Contexte :

Soit deux échantillons indépendants $X_{1,n_1}^{(1)}$ et $X_{1,n_2}^{(2)}$ dont les éléments suivent les lois normales $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$.
Les variance sont égales mais inconnues :$\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Avant d'utiliser le résultat suivant, il parait très raisonnable de tester l'hypothèse d'égalité des variances (test bilatéral, voir diapo précédente).

Si l'hypothèse d'égalité des variances est (ou paraît suffisamment) vérifiée , on a :

$$\frac{D_n - (m_1 - m_2)}{\sqrt{\frac{\widehat{S^2_{n_1+n_2}}^{(1,2)}}{n_1} + \frac{\widehat{S^2_{n_1+n_2}}^{(1,2)}}{n_2}}} \sim \mathcal{S}(n_1 + n_2 - 2),$$

avec

$$\widehat{S^2_{n_1+n_2}}^{(1,2)} = \frac{(n_1 - 1)\widehat{S^2_{n_1}}^{(1)} + (n_2 - 1)\widehat{S^2_{n_2}}^{(2)}}{n_1 + n_2 - 2}.$$

($\widehat{S^2_{n_1}}^{(1)}$ et $\widehat{S^2_{n_2}}^{(2)}$ désignent les variances empiriques des deux échantillons.)

Démontrer ce résultat.

Lois approchées des statistiques de deux échantillons indépendants

Il s'agit de TCL dans des cas particuliers. Les variances peuvent être remplacés par leurs estimateurs (Théorème de Slutsky).

(i) Lorsque $X^{(1)}$ et $X^{(2)}$ sont des lois de Bernoulli, on a :

$$\frac{D_n - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

(ii) Lorsque $X^{(1)}$ et $X^{(2)}$ sont des lois quelconques de moyenne m_1 et m_2, on a :

$$\frac{D_n - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Plan

Quel est le but ?

Il s'agit de tester l'adéquation d'un n-échantillon $X_1, \ldots, X_n$ à une loi de référence, qui est discrète et à support fini, ou bien à une famille de loi discrète :

1. Cet échantillon a-t-il été généré par une loi uniforme sur $\{1, \ldots, 6\}$?
2. Cet échantillon a-t-il été généré par une loi de Poisson ?

Dans le premier cas, on connaît déjà entièrement la distribution ($P(X = i) = 1/6$ pour $i = 1, \ldots 6$). Dans le second cas, il faut d'abord estimer un meilleur candidat pour la distribution.

Dans les deux cas, on finira par comparer nos observations avec une distribution $(p_1, p_2, \ldots, p_k)$ où p_i désigne la probabilité d'une classe i.

L'hypothèse nulle est donc :

H_0 : "l'échantillon a été généré par la distribution $(p_1, p_2, \ldots, p_k)$"

La statistique de test est :

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

avec

- n_i, le nombre d'observation de la classe i, l'effectif réel.
- np_i, la probabilité de la classe i multipliée par la taille de l'échantillon, l'effectif théorique.

Presque une somme de loi normale au carré ?

Supposons que les k classes soient les valeurs de 1 à k.
On considère les estimateurs empiriques de la probabilité de chacune des classes :

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_i=i\}}.$$

On a alors :

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(n\hat{p}_i - np_i)^2}{np_i} = \sum_{i=1}^{k} n\frac{(\hat{p}_i - p_i)^2}{p_i}$$

Or, pour $i = 1, \ldots, k$, et si H_0 est vrai, on a :

$$\sqrt{n}\frac{\left(\hat{p}_i - p_i\right)}{\sqrt{p_i(1-p_i)}} \xrightarrow[n\longrightarrow+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

À l'aide du théorème de Cochran, on démontre que :

$$T \xrightarrow[n\longrightarrow+\infty]{\mathcal{L}} \chi^2(k-1-r)$$

avec k, le nombre de classes et r, le nombre de paramètres à estimer dans la famille de lois. Si il y a une seule loi de référence, le degré de liberté du χ^2 est $k-1$.

Exemple 1 : le dé est-il équilibré ?

On a effectué 100 tirages d'un dé dont les résultats sont présentés dans le tableau le tableau ci-dessous :

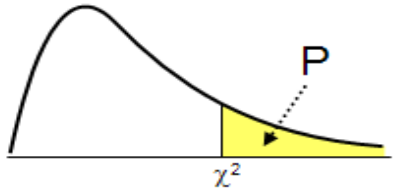
Classes	1	2	3	4	5	6
Effectif réels, n_i	7	18	26	15	18	6
Effectif théorique, np_i	16.67	16.67	16.67	16.67	16.67	16.67

1. Comment obtient-on la valeur de la statistique $T = 11.24$?

2. Quelle est l'hypothèse H_0?

3. Quelle est la loi asymptotique de T sous H_0?

4. Peut-on rejeter l'hypothèse nulle au seuil significatif de 5% ? Même question au seuil significatif de 1%.



	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797
22	8.643	10.982	27.301	30.813	33.924	36.781	37.659	40.289	42.796	45.962	48.268
23	9.260	11.689	28.429	32.007	35.172	38.076	38.968	41.638	44.181	47.391	49.728
24	9.886	12.401	29.553	33.196	36.415	39.364	40.270	42.980	45.559	48.812	51.179
25	10.520	13.120	30.675	34.382	37.652	40.646	41.566	44.314	46.928	50.223	52.620
26	11.160	13.844	31.795	35.563	38.885	41.923	42.856	45.642	48.290	51.627	54.052
27	11.808	14.573	32.912	36.741	40.113	43.195	44.140	46.963	49.645	53.023	55.476
28	12.461	15.308	34.027	37.916	41.337	44.461	45.419	48.278	50.993	54.411	56.892
29	13.121	16.047	35.139	39.087	42.557	45.722	46.693	49.588	52.336	55.792	58.301
30	13.787	16.791	36.250	40.256	43.773	46.979	47.962	50.892	53.672	57.167	59.703
31	14.458	17.539	37.359	41.422	44.985	48.232	49.226	52.191	55.003	58.536	61.098

Exemple 2 : cet échantillon suit-il une loi de Poisson ?

Dans le cadre d'une étude météo, on cherche à modéliser la survenue de dégats liés à la foudre dans des communes rurales de montagne. On dispose du relevé du nombre de frappes de foudre en 1 an dans deux cent communes.

Nbre de frappes	≤ 1	2	3	4	5	6	7	≥ 8
Effectif réels, n_i	17	31	37	41	30	23	13	8

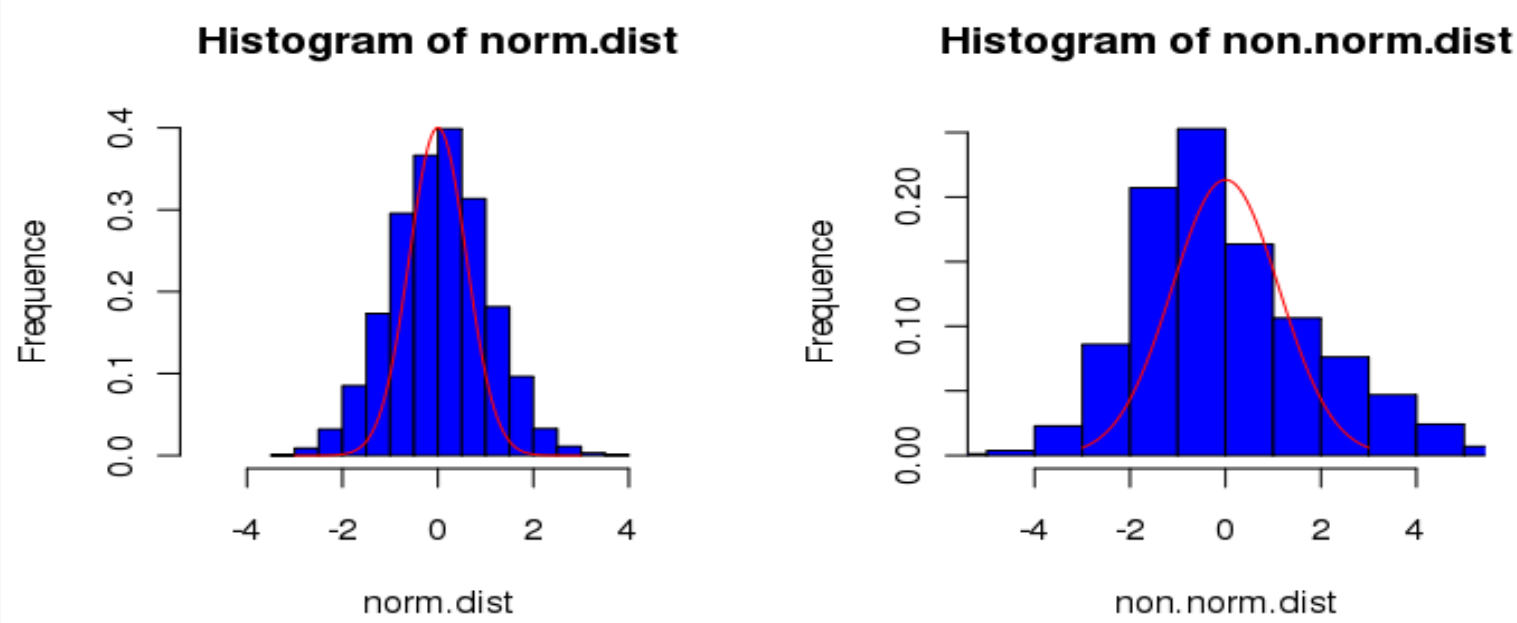
1. Quel estimateur peut-on choisir pour le paramètre de la loi de poisson ? 1. Comment obtient-on la valeur de la statistique $T = 1.3$?

2. Quelle est l'hypothèse H_0?

3. Quelle est la loi asymptotique de T sous H_0?

4. Peut-on rejeter l'hypothèse nulle au seuil significatif de 5% ? Même question au seuil significatif de 1%.

Un test de normalité ?



Oui, le test du χ^2 peut servir à tester la normalité d'un n-échantillon. La méthode :

- découper l'espace des observations en classes,
- assimiler chacune des valeurs de l'échantillon à la classe à laquelle elle appartient (un histogramme),
- calculer les paramètres empiriques de la loi normales : $\overline{X}_n$ et $\widehat{S_n^2}$.
- calculer la statistique-distance du χ^2, T, entre les effectifs réels des classes et les effectifs théorique des classes d'après la loi normale estimée. Celle ci suit une loi du χ^2 à $n-3$ degré de liberté.

Cependant le test de Kolmogorov-Smirnov est plus puissant. Il se sert de toute l'information de l'échantillon, contrairement au test du χ^2 qui en perd en assimilant les données à leur classe.

Controverse : homogénéité vs indépendance

Le professeur Labarbe : - Je vais vous présenter deux importants problèmes dont on étudiera les liens potentiels.

- **Test d'homogénéité du χ^2** : On observe deux échantillons X_1^n et Y_1^n prenant un nombre fini de valeurs. Ces deux échantillons ont-ils pu être générés par la même loi de probabilité ?
- **Test d'indépendance du χ^2** : Dans une population, on observe deux caractères X et Y prenant un nombre fini de valeurs. Peut-on considérer que X est indépendant de Y ?

Le professeur Moustache : C'est faux, il s'agit du même problème !

Qq arguments

Le professeur Moustache :
Exemple 1 : Dans une population, chaque personne est, d'une part, soit un homme, soit une femme et d'autre part, soit mathématicien, soit non-mathématicien. Se demander si les mathématiciens/non-mathématicien sont répartis de la même manière chez les hommes que chez les femmes revient à savoir si le fait d'être mathématicien est indépendant du sexe de la personne.

Exemple 2 : Pour deux échantillons de patients provenant de deux hôpitaux A et B, on relève le fait qu'ils soient bien soignés ou mal-soignés. Étudiez si la proportion des bien soignés/mal-soignés est la même dans les 2 échantillons revient à savoir si la qualité des soins est indépendante de l'hôpital choisi.

Le professeur Kératine Labarbe : N'importe quoi. J'ai deux dés et je fais des tirages pour savoir si les deux dés ont la même loi de probabilité ... Il n'y a pas de question d'indépendance la-dedans.

Le professeur Moustache : Et pourtant si, il y a une question d'indépendance. C'est simple ...

Terminez la phrase ci-dessus.

Exemple 1

Soit X qui vaut 1 quand la personne est une femme, 0 sinon. Soit Y qui vaut 1 quand la personne est mathématicienne, 0 sinon.

X, Y		$X = 0$	$X = 1$
		$N_{0\bullet} = 100$	$N_{1\bullet} = 100$
$Y = 0$	$N_{\bullet 0} = 20$	$O_{00} = 8$	$O_{10} = 12$
$Y = 1$	$N_{\bullet 1} = 180$	$O_{01} = 92$	$O_{11} = 88$

Table – Table des effectifs réels

X, Y		$X = 0$	$X = 1$
		$N_{0\bullet} = 100$	$N_{1\bullet} = 100$
$Y = 0$	$N_{\bullet 0} = 20$	$N_{00} = 10$	$N_{10} = 10$
$Y = 1$	$N_{\bullet 1} = 180$	$N_{01} = 90$	$N_{11} = 90$

Table – Table des effectifs théoriques

Il ne reste plus qu'à effectuer un test d'adéquation du χ_2 entre les deux tables $\rightarrow$

Suite de l'exemple 1

L'hypothèse nulle, H_0, est au choix :

1. il y a la même répartition de mathématiciens/non mathématicien chez les hommes que chez les femmes.
2. la qualité mathématicien/non-mathématicien est indépendante du sexe de la personne.

On se concentre sur les 4 cases principales des deux tableaux (pas les marginales). La statisque T est :

$$T = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(88-90)^2}{90} + \frac{(92-90)^2}{90}$$

Sous l'hypothèse nulle la statistique T suit la loi du $\chi^2(k-1-r)$, avec k le nombre de cases du tableau, et $r = 2$ le nombre de paramètre estimé pour connaître la répartition théorique (1 case à connaître dans chaque marginale).

Cas général

Soient $X \in \{1, \ldots, I\}$ et $Y \in \{1, \ldots, J\}$.
On observe un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$.
L'hypothèse nulle :

$$H_0 : X \text{ et } Y \text{ sont indépendantes}$$
ou
$$H_0 : X|Y = j \text{ et } X|Y = j' \text{ ont la même loi, pour tout } j \text{ et } j'.$$

Soient $O_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=i, Y_k=j\}}$, les effectifs observés pour les $I \times J$ possibilités, et $N_{i\bullet} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=i\}}$ et $N_{\bullet j} = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k=j\}}$, les effectifs marginales.
Les effectifs théoriques sous l'hypothèse nulles sont : $N_{ij} = N_{i\bullet} N_{\bullet j}$.
Sous l'hypothèse nulle, on a :

$$T = \sum_{i=1}^{I} \sum_{i=1}^{I} \frac{(O_{ij} - N_{ij})^2}{N_{ij}} \sim \chi^2 \left[(I-1)(J-1)\right].$$

Remarque : $(I-1)(J-1)$ est égale au nombre de classes $I \times J$ moins 1, moins le nombre de paramètres à estimer dans chacune des marginales pour déterminer les effectifs théoriques, respectivement $I-1$ et $J-1$.

Exercice : test du χ^2 pour des échantillons appariés (test de Mc Nemar)

Un traitement préventif est donné à une population de malades. On souhaite déterminer si ce traitement permet de diminuer la présence d'un symptôme. Les données de présence du symptôme ont été récoltées avant la prise du traitement : $X_0^i = 1$ si l'individu i à le symptôme avant traitement, et 0 sinon. On a traité les patients et on a effectué le même relevé post-traitement : $X_1^i = 1$ si l'individu i à le symptôme après traitement, et 0 sinon. La table des effectifs avant et après traitement est la suivante.

X_0, X_1		$X_1 = 0$	$X_1 = 1$
		$N_{0\bullet} = 115$	$N_{1\bullet} = 108$
$X_0 = 0$	$N_{\bullet 0} = 95$	$O_{00} = 45$	$O_{10} = 50$
$X_0 = 1$	$N_{\bullet 1} = 128$	$O_{01} = 70$	$O_{11} = 58$

On cherche à tester l'efficacité du traitement. On se place sous l'hypothèse H_0 qu'il n'en a pas :

$$H_0 : N_{0\bullet} = N_{\bullet 0} \quad (\text{ou } N_{1\bullet} = N_{\bullet 1}).$$

Cela revient à tester l'égalité de $O_{01} = O_{10}$. Un tel test peut être effectué en supposant que l'on dispose de patients dont la présence des symptômes change entre la première et la seconde mesure (effectif : $O_{01} + O_{10}$) et que ces patients se répartissent dans les classe "0,1" et "1,0" avec une probabilité 1/2. Effectuer un test du χ^2 dans cette circonstance. En particulier, on montrera que la statistique d'intérêt est :

$$\chi = \frac{(O_{01} - O_{10})^2}{O_{01} + O_{10}},$$

et on donnera sa loi.

Plan

ANalysis Of VAriance, ANOVA.

C'est une méthode de construction de test pour comparer des moyennes et plus exactement pour effectuer un test d'égalité des moyennes entre plusieurs groupes.

L'idée principale est basée sur la loi de la variance totale, une propriété probabiliste :
Supposons que l'on s'intéresse à une variable X, et que l'on ait accès à une information Y sur cette variable. L'information Y peut être une appartenance à un groupe ; par exemple X est le quotient intellectuel des étudiants qui se répartissent en deux groupes : ceux qui aiment le café $Y = 1$ et ceux qui n'aiment pas $Y = 0$.

Evidemment, les aficionados du café sont plus futés ... mais peut-être pas ?

La variance de X peut être décomposé comme suit :

$$\mathrm{Var}(X) = \mathrm{Var}[\mathbb{E}(X|Y)] + \mathbb{E}[\mathrm{Var}(X|Y)].$$

Le premier des deux termes est la variance expliqué par Y : on calcule l'espérance pour chaque groupe, puis considérant que la répartition des groupes est aléatoire, on calcule la variance de ces espérances.

Le second terme est la variance non expliqué par Y : on calcule la variance dans chaque groupe, puis considérant que la répartition des groupes est aléatoire, on calcule la moyenne de ces variances.

En pratique, on dispose d'un échantillon réparti en p groupes, ou dit autrement de p échantillons $x_1, \ldots, x_p$. L'échantillon x_i comprend n_i variables aléatoires de lois normales : $x_i = (x_{i,1}, \ldots, x_{i,n_i})$.
Et on calcule des sommes d'erreurs élevées au carré.
Sum Squared Between groups, SSB :

$$SSB = \sum_{i=1}^{p} n_i (\overline{x_i} - \overline{x})^2,$$

Sum Squared Within groups, SSW :

$$SSW = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x_i})^2,$$

et bien sûr, Sum Squared Total, SST :

$$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2 = SSB + SSW,$$

Si, on divise ces termes par $n = n_1 + \ldots + n_p$, on obtient des variances empiriques (mais c'est inutile en pratique).

ANOVA, un test de Fisher

On suppose que les X_i sont des n_i-échantillons de lois normales et de même variance σ^2. Ces échantillons forment p groupes et i est le label du groupe.
L'espérance associée à l'échantillon X_i est notée m_i, on a :

$$\forall (i,j),\ X_{i,j} \sim \mathcal{N}(m_i, \sigma^2)$$

L'hypothèse nulle est : $H_0 : m_1 = m_2 = \ldots = m_p$.
L'hypothèse nulle devrait être rejetée si une ou plusieurs des moyennes diffèrent de la valeur commune.
Dans cette circonstance, on a :

$$SSB \sim \chi^2(DDL_B), \quad \text{avec} \quad DDL_B = p - 1$$

et $SSW \sim \chi^2(DDL_W)$ avec

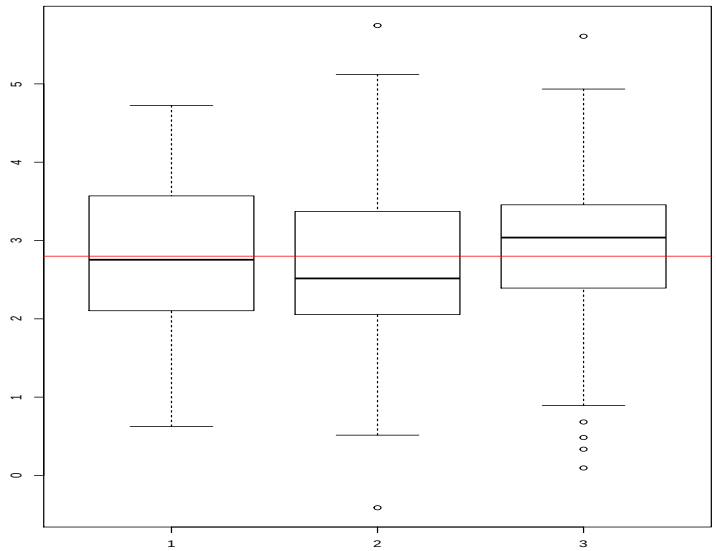
$$DDL_W = (n_1 - 1) + (n_2 - 1) + \ldots + (n_p - 1) = n - p$$

Ainsi,

$$F_{stat} = \frac{SSB/DDL_B}{SSW/DDL_W} \sim \mathcal{F}(p-1, n-p).$$

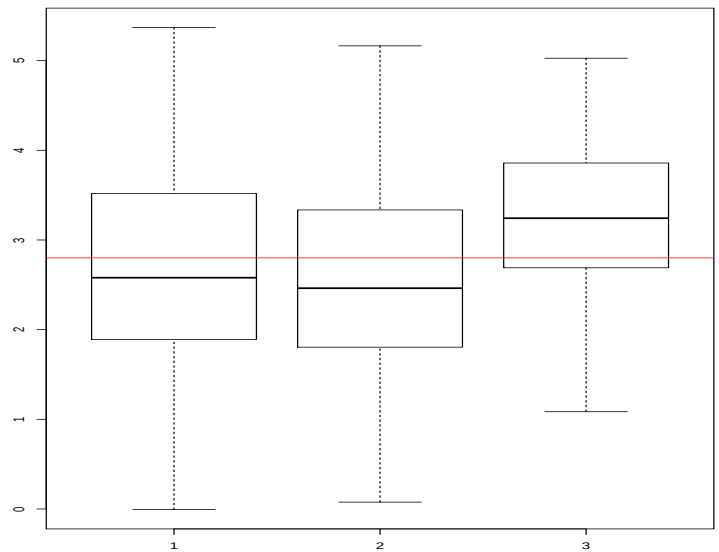
On ne rejette qu'à droite (grande valeur de F_{stat}) ... les valeurs à gauches (SSB petit par rapport à SSW) signifient que la différence entre groupes est encore plus petite qu'elle ne devrait être au vu de la variabilité totale (donc pas de rejet de H_0) ... cependant, les valeurs les plus communes sous H_0 sont proches de 1.

Deux exemples : l'argent dépensé à midi par les étudiants



Contexte : trois groupes d'étudiants, les L1, les L2 et les L3.
En réalité les trois échantillons de tailles 100 sont générés par par une loi normales $\mathcal{N}(2.8, 1)$
$F_{stat} = 1.34$, $p_{val} = 0.26$.
Conclusion : On ne rejette pas H_0. Cette répartition en groupe ne semble pas expliqué grand chose.

Contexte : trois groupes de 100 étudiants, les Licences ($\mathcal{N}(2.6, 1)$), les masters $\mathcal{N}(2.6, 1)$ et les thésards ($\mathcal{N}(3.2, 1)$).
$F_{stat} = 14.57$, $p_{val} = 10^{-7}$
Conclusion : On rejette H_0. Cette répartition en groupe permet d'observer des différences.



Quels sont les paramètres de la loi de Fisher ?

Plan

Tests de Kolmogorov-Smirnov (1/3)

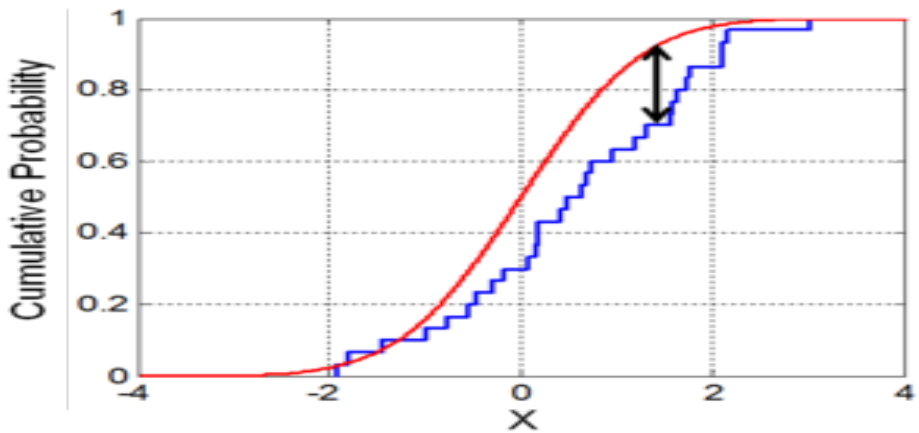
Le test de Kolmogorov-Smirnov permet de se poser les questions suivantes :

1. Un échantillon de v.a. continue suit-il une loi de référence ?
2. Deux échantillons suivent-ils la même loi ?

Pour répondre à cela, on utilise la fonction de répartition empirique d'un n-échantillon.

$$F_n(x) = \frac{\text{Nombre de points } \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq n\}}$$

Dans la première question, si la loi de référence a pour fonction de répartition F, La statistique de test est : $D_n = \sup_x |F_n(x) - F(x)|$.



Tests de Kolmogorov-Smirnov (2/3)

Sous l'hypothèse nulle, la distribution asymptotique de D_n ne dépend pas de la loi de référence, F (à démontrer). Et il est possible d'approximer cette loi (voir table à la diapositive suivante). On peut donc construire un test qui marche pour toute les lois à partir de la statistique D_n.

Remarques :

- On peut modifier le test pour comparer deux échantillons : $D_n = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$.
- On peut tester la normalité d'un échantillon. Il faut alors le standardiser et comparer sa fonction de répartition empirique avec celle de la loi normale $\mathcal{N}(0,1)$. Cela revient à estimer les paramètres de la loi normale est D_n n'a plus exactement la même loi. La modification du test a été proposée par Lilliefors, d'où le nom test de Lilliefors.
- Le test de Kolmogorov-Smirnov peut servir à tester si un générateur aléatoire fonctionne bien.

Avec R :

```
> X <- c(8, 9, 9, 10, 10, 10, 11, 13, 14, 14)
> lillie.test(X)
Lilliefors (Kolmogorov-Smirnov) normality test
data: X
D = 0.2451, p-value = 0.0903
```

L'hypothèse de normalité de l'échantillon X peut-elle être rejetée au seuil de 5% ?

Tests de Kolmogorov-Smirnov (3/3)

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
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897
36	.17418	.19910	.22119	.24732	.26532
37	.17188	.19646	.21826	.24404	.26180
38	.16966	.19392	.21544	.24089	.25843
39	.16753	.19148	.21273	.23786	.25518
40	.16547	.18913	.21012	.23494	.25205
41	.16349	.18687	.20760	.23213	.24904
42	.16158	.18468	.20517	.22941	.24613
43	.15974	.18257	.20283	.22679	.24332
44	.15796	.18053	.20056	.22426	.24060
45	.15623	.17856	.19837	.22181	.23798
46	.15457	.17665	.19625	.21944	.23544
47	.15295	.17481	.19420	.21715	.23298
48	.15139	.17302	.19221	.21493	.23059
49	.14987	.17128	.19028	.21277	.22828
50	.14840	.16959	.18841	.21068	.22604
51	.14697	.16796	.18659	.20864	.22386
52	.14558	.16637	.18482	.20667	.22174
53	.14423	.16483	.18311	.20475	.21968
54	.14292	.16332	.18144	.20289	.21768
55	.14164	.16186	.17981	.20107	.21574
56	.14040	.16044	.17823	.19930	.21384
57	.13919	.15906	.17669	.19758	.21199
58	.13801	.15771	.17519	.19590	.21019
59	.13686	.15639	.17373	.19427	.20844
60	.13573	.15511	.17231	.19267	.20673
61	.13464	.15385	.17091	.19112	.20506
62	.13357	.15263	.16956	.18960	.20343
63	.13253	.15144	.16823	.18812	.20184
64	.13151	.15027	.16693	.18667	.20029
65	.13052	.14913	.16567	.18525	.19877
66	.12954	.14802	.16443	.18387	.19729
67	.12859	.14693	.16322	.18252	.19584
68	.12766	.14587	.16204	.18119	.19442
69	.12675	.14483	.16088	.17990	.19303
70	.12586	.14381	.15975	.17863	.19167
71	.12499	.14281	.15864	.17739	.19034
72	.12413	.14183	.15755	.17618	.18903
73	.12329	.14087	.15649	.17498	.18776
74	.12247	.13993	.15544	.17382	.18650
75	.12167	.13901	.15442	.17268	.18528
76	.12088	.13811	.15342	.17155	.18408
77	.12011	.13723	.15244	.17045	.18290
78	.11935	.13636	.15147	.16938	.18174
79	.11860	.13551	.15052	.16832	.18060
80	.11787	.13467	.14960	.16728	.17949
81	.11716	.13385	.14868	.16626	.17840
82	.11645	.13305	.14779	.16526	.17732
83	.11576	.13226	.14691	.16428	.17627
84	.11508	.13148	.14605	.16331	.17523
85	.11442	.13072	.14520	.16236	.17421
86	.11376	.12997	.14437	.16143	.17321
87	.11311	.12923	.14355	.16051	.17223
88	.11248	.12850	.14274	.15961	.17126
89	.11186	.12779	.14195	.15873	.17031
90	.11125	.12709	.14117	.15786	.16938
91	.11064	.12640	.14040	.15700	.16846
92	.11005	.12572	.13965	.15616	.16755
93	.10947	.12506	.13891	.15533	.16666
94	.10889	.12440	.13818	.15451	.16579
95	.10833	.12375	.13746	.15371	.16493
96	.10777	.12312	.13675	.15291	.16408
97	.10722	.12249	.13606	.15214	.16324
98	.10668	.12187	.13537	.15137	.16242
99	.10615	.12126	.13469	.15061	.16161
100	.10563	.12067	.13403	.14987	.16081
$n > 100$	$1.073/\sqrt{n}$	$1.223/\sqrt{n}$	$1.358/\sqrt{n}$	$1.518/\sqrt{n}$	$1.629/\sqrt{n}$

Test de Shapiro-Wilk

Soit $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ un n-échantillon ordonné
(on parle de statistique d'ordre).

Le test de Shapiro-Wilk teste la normalité (H_0 : l'échantillon est normal) à l'aide de la statistique suivante :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

où les coefficients a_i sont calculés sur la base des espérances et variances, covariances des statistique d'ordre d'une loi normale standardisé.

La statistique W n'a pas de nom, mais sous l'hypothèse nulle elle suit une unique loi dont la table est approchée par simulation de Monte-Carlo.

"Des simulations ont montré que le test de Shapiro-Wilks a une plus grande puissance que le test de Lilliefors "(wikipedia en anglais).

Plan

Contexte et notion de résultats potentiels

On souhaite tester l'efficacité d'un traitement (ou les effets d'une mesure sociales). Le contexte privilégié est celui d'un essai clinique randomisé.

La population contient N patients répartis en deux groupes : $T_i = 1$ si le patient i est traité, et $T_i = 0$ si le patient i appartient au groupe contrôle.

Le patient i a pour covariable X_i, et le résultat de l'expérience pour ce patient est $Y_i(T_i)$, une v.a. dépendant de l'assignation à l'un des deux groupes.

Exemple : Si Y est le niveau sonore des ronflements pendant la nuit (en db), alors $Y_4(1) = 56$ signifie que le quatrième patient a ronflé à un niveau sonore de 56db durant la nuit post-traitement.

Idées principales : Chacun des patients est affecté à un seul des deux groupes. On cherche à comparer les données du groupe traité et du groupe contrôle.
Traditionnellement, la statistique ne s'occupe que de corrélation. Un test classique étudiera la corrélation entre Y et T.
Cependant, on peut se ramener à de la causalité en considérant les résultats potentiels des patients selon la valeur de T, on étudie :

$$Y(1) - Y(0), \text{ pour un certain patient.}$$

Hypothèse : Les résultats potentiels d'un patients ne varie pas en fonction du traitement assignés aux autre patients, et il n'existe pas différentes forme ou version associés au deux groupe de traitement qui aurait amené à d'autres résultats potentiels.

Expérience randomisée

La variable $\mathbf{T} = (T_1, \ldots, T_N) \in \{0,1\}^N$ mais elle n'est pas forcément uniforme dans cette ensemble.

Expérience de Bernoulli : Les variables d'assignation des patients, T_i, sont mutuellement indépendantes.

$$\mathbb{P}(\mathbf{T} \mid X) = \prod_{i=1}^{N} p(X_i)^{T_i} (1 - p(X_i))^{1-T_i}.$$

Expérience complètement randomisée : Un nombre N_t de patients appartenant au groupe traité est fixé. Et le tirage s'effectue indépendamment des covariables. Si $N_t = \sum_{i=1}^{N} T_i$, on a :

$$\mathbb{P}(\mathbf{T} \mid X) = \binom{N}{N_t}^{-1}.$$

Expérience randomisée par strates : On considére J blocs et $B_i \in \{1, \ldots, J\}$ indique le bloc du patient i. Chaque bloc contient $N(j)$ patients et on souhaite que $N_t(j)$ patients du blocs j appartiennent au groupe traité. Sous ces conditions,

$$\mathbb{P}(\mathbf{T} \mid X) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1}.$$

Si chaque blocs contient deux patients (1 pour chaque groupe), on parle d' experience randomisée appariée.

Exemple : traitement des ronflements (1/2)

On se place dans le cadre d'un essai clinique complètement randomisé. Un nouveau traitement permettant de diminuer le niveau sonore des ronflements est testé.

- La variable T correspond à l'assignation au groupe traitement ou au groupe placebo.
- La covariable X est le niveau sonore moyen des ronflements lors de la semaine précédent le traitement (ou le placebo).
- La variable $Y(0)$ est la réponse potentielle du patient soumis au placebo.
- La variable $Y(0)$ est la réponse potentielle du patient soumis au traitement.
- Seule la variable $Y(T)$ est observée pour chaque patient.

Patient	groupe T	cov. X	rés. $Y(0)$	rés $Y(1)$
1	1	59	?	58
2	1	89	?	75
3	0	72	71	?
4	0	50	53	?
5	1	76	?	65
6	0	72	69	?

Les moyennes d'observation pour les deux groupes sont :

$$\overline{Y}_t^{obs} = \frac{1}{N_t}\sum_{i=1}^{N} Y_i(1)\mathbb{1}_{\{T_i=1\}} \quad \text{et} \quad \overline{Y}_c^{obs} = \frac{1}{N_c}\sum_{i=1}^{N} Y_i(0)\mathbb{1}_{\{T_i=0\}}$$

Exemple : traitement des ronflements (2/2)

On s'intéresse à la différence absolue entre les moyennes des réponses pour chaque groupes :

$$S(\mathbf{T}, \mathbf{Y}^{obs}) = |\overline{Y}_t^{obs} - \overline{Y}_c^{obs}|$$

On cherche à tester l'hypothèse H_0 que le traitement a le même effet que le placebo. Sous H_0, on connaît les réponses potentielles des patients qui nous étaient jusqu'alors inaccessibles :

Patient	groupe T	cov. X	rés. $Y(0)$	rés $Y(1)$
1	1	59	(58)	58
2	1	89	(75)	75
3	0	72	71	(71)
4	0	50	53	(53)
5	1	76	(65)	65
6	0	72	69	(69)

Dans cette circonstance, $S(\mathbf{T}, \mathbf{Y}^{obs}) = 2.66$.

Test exact de Fisher : Si on change le vecteur $\mathbf{T}$ pour $\tilde{T}$, sous H_0, on ne change pas les observations mais on change la statistique $S(\tilde{\mathbf{T}}, \mathbf{Y}^{obs})$. Il est alors possible de calculer la p-valeur du test exact de Fisher en tirant aléatoirement $\tilde{\mathbf{T}}$ selon la méthode d'assignation utilisée. Si l'expérience est complètement randomisée, il y a $\binom{6}{3}$ possibilités. La p-valeur pour l'hypothèse d'égalité des effets est la proportion des statistiques $S(\tilde{\mathbf{T}}, \mathbf{Y}^{obs})$ plus extrêmes que $S(\mathbf{T}, \mathbf{Y}^{obs})$.

Différentes statistiques possibles

Median statistics :

$$S_{med} = \left| \text{med}(\mathbf{Y}_t^{obs}) - \text{med}(\mathbf{Y}_c^{obs}) \right| .$$

T-test statistics :

$$S_{test} = \frac{\left| \overline{Y}_t^{obs} - \overline{Y}_c^{obs} \right|}{\sqrt{\frac{\hat{S}_t^2}{N_t} + \frac{\hat{S}_c^2}{N_c}}}$$

Kolmogorov statistics :

$$S_{ks} = \max_{i=1,\ldots,N} \left| \hat{F}_{t,N_t}^{obs}(Y_i^{obs}) - \hat{F}_{t,N_t}^{obs}(Y_i^{obs}) \right|$$

On note que ces statistiques ne sont pas comparées à leur lois usuelles. La p-valeur exacte est évaluée par randomisation. Elle est éventuellement approchée par simulation de Monte Carlo si l'échantillon est grand.

Dans le cadre de l'exemple au slide précédent, proposez une statistique qui inclut la covariable.

Panorama

Tests d'adéquation

- Test de student pour l'adéquation de l'espérance d'une loi normale
- Test de wilcoxon non-paramétrique pour la médiane
- Test du χ^2 pour l'adéquation à une loi en découpant l'espace probabilisé en k classes.
- Test de kolmogorov-Smirnov pour l'adéquation à une loi continues (distance max. entre les fonctions de répartition théorique et empirique)
- Test de Shapiro-Wilk ou de Lilliefors pour l'adéquation à la famille des lois normales.

Tests de comparaisons

- Test de Student pour la comparaison de deux moyennes de lois normales
- Test de Fisher pour la comparaison de deux variances de lois normales
- ANOVA, comparaison de la moyenne de plus de deux échantillons (hypothèse : normalité et homoscédasticité)

Tests d'indépendance

- Test du χ^2
- Tests des coefficients de corrélation de spearman, de Pearson ou de Kendall.