Tests statistiques

1 Échantillonage

Pour étudier certains caractères d'une population, on peut avoir recours à la statistique descriptive : on étudie le caractère voulu sur la totalité des individus, mais c'est souvent impossible en pratique. On se limite donc à l'étude sur un échantillon, c'est à dire un sous-ensemble de la population. On extrapole les résultats obtenus sur l'échantillon à l'ensemble de la population. La détermination d'un échantillon permettant une bonne extrapolation est un problème complexe.

1.1 Méthodes d'échantillonnage

Échantillon aléatoire : un tel échantillon –de taille n– s'obtient en effectuant n tirages indépendants, avec remise, au sein de la population. C'est également assez difficile à réaliser en pratique, mais c'est la seule façon d'obtenir un échantillon vraiment représentatif de la population initiale –c'est à dire possédant *a priori* les mêmes caractéristiques– et surtout, non biaisé. C'est aussi le seul type d'échantillon pour lequel on peut calculer une marge d'erreur, et donc évaluer la qualité du test.

Échantillon raisonné : une telle façon de procéder suppose qu'on connaît à l'avance certaines propriétés de la population qu'on étudie. On peut alors seulement vérifier *a postériori* que les hypothèses émises sont vraies, et cette méthode peut mener à des résultats erronés lorsqu'on a mal analysé le problème.

Échantillon représentatif par rapport à un paramètre : un tel échantillon est un compromis entre les deux précédants, mais comporte aussi un aspect arbitraire qui peut entacher d'erreur les résultats obtenus. On décide par exemple que l'échantillon à étudier aura la même distribution par classe d'âge que la population d'origine, puis on prend un échantillon aléatoire dans chaque classe d'âge. Cela suppose que l'âge est un facteur qui influe sur le caractère à étudier, et surtout on privilégie ce paramètre par rapport aux autres.
La méthode des quotas généralise ce principe : on choisit plusieurs paramètres pour lequels l'échantillon à étudier devra avoir les mêmes propriétés que la population entière.

1.2 Taille de l'échantillon

La taille de l'échantillon étudié est un facteur essentiel qui détermine en partie la fiabilité des résultats. Le coût de l'échantillonage augmentant avec la taille de l'échantillon, il faut connaître la taille minimale nécessaire à l'obtention de résultats corrects.

2 Estimation de la moyenne et de l'écart type à partir d'un échantillon

On considère un échantillon aléatoire simple, de taille n donnée, d'une population P de taille N. Les tirages pour déterminer l'échantillon sont donc censés être faits avec remise, ce qui n'est jamais le cas en pratique. On montre (cours de proba sur les lois binômiale et hypergéométrique) que lorsque $N \to +\infty$, cela revient au même. Toutefois, on donnera le résultat des calculs pour les deux types d'échantillon (avec et sans remise) chaque fois que c'est possible. On appellera *échantillon de type 1*, un véritable échantillon aléatoire (avec remise) et *échantillon de type 2*, un échantillon déterminé par des tirages aléatoires sans remise.

On considère la variable aléatoire X, associée au caractère étudié dans la population.

2.1 Échantillon de type 1

La distribution d'un échantillon de type 1 est binômiale ; en effet chaque élément de la population a une probabilité $p = 1/N$ d'être obtenu, et ce à chaque tirage. Pour chacun des n tirages, on définit Y_k la variable aléatoire égale au résultat du k^{e} tirage ; les variables aléatoires $Y_1, \ldots, Y_n$ sont indépendantes et suivent toutes la même loi (celle de X). On note $y_1, \ldots, y_n$, les valeurs prises par $Y_1, \ldots, Y_n$. Enfin, on pose $\overline{Y} = \dfrac{Y_1 + \cdots + Y_n}{n}$.

La probabilité qu'un élément donné soit tiré i fois ($0 \leqslant i \leqslant n$) suit une loi binômiale $\mathcal{B}(n, \frac{1}{N})$, on a :

$E(\overline{Y}) = \dfrac{1}{n}\sum_{k=1}^{n} y_k$; $E(X) = \dfrac{1}{N}\sum_{k=1}^{N} x_k$ et on montre que $E(\overline{Y}) = E(X) = m$.

$V(X) = E\left((X-m)^2\right)$; $V(\overline{Y}) = E\left((\overline{Y}-m)^2\right)$ et on montre que $V(X) = \sigma^2$, $V(\overline{Y}) = .\dfrac{\sigma^2}{n}$

Autrement dit, la moyenne sur l'échantillon est la même que dans la population, la variance est divisée par n.

2.2 Échantillon de type 2

La distribution d'un échantillon de type 2 est hypergéométrique. Pour chacun des n tirages, on définit Z_k la variable aléatoire égale au résultat du k^{e} tirage ; les variables aléatoires $Z_1, \ldots, Z_n$ ne sont pas indépendantes, même si elles suivent toutes la même loi que X. On note $z_1, \ldots, z_n$, les valeurs prises par $Z_1, \ldots, Z_n$ et on pose $\overline{Z} = \dfrac{Z_1 + \cdots + Z_n}{n}$.

La probabilité qu'un élément donné soit tiré i fois ($0 \leqslant i \leqslant n$) suit une loi hypergéométrique $\mathcal{H}(n, \frac{1}{N}, N)$, on a :

$E(\overline{Z}) = \dfrac{1}{n}\sum_{k=1}^{n} z_k$ et on montre que $E(\overline{Z}) = E(X) = m$.

$V(\overline{Z}) = E\left((\overline{Z}-m)^2\right)$ et on montre que $V(\overline{Z}) = \dfrac{\sigma^2}{n}\dfrac{N-n}{N-1}$.

La moyenne de l'échantillon est conservée, mais pas la variance qui est plus faible que dans le cas de tirages avec remise.

3 Intervalle de confiance d'une moyenne

On considère toujours un échantillon de type 1 ou 2, identique à celui défini au paragraphe 2. On voit en cours de proba que la loi hypergéométrique converge en loi vers la loi binômiale (programme de 1re année), et que la loi binômiale converge en loi vers la loi normale (programme de 2e année). Pour des échantillons de taille n suffisante (en pratique $n > 30$ convient), on assimile la distribution de l'échantillon à une distribution normale de même espérance m que l'échantillon, et de même variance σ^2/n (σ^2 étant la variance de la population). On reprend les notations de l'échantillon de type 1 définies au paragraphe 2.1. Donc on considère que $\overline{Y}$ suit une loi normale de paramètres m et $\frac{\sigma}{\sqrt{n}}$, $\overline{Y} \hookrightarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$; par conséquent $\sqrt{n}\,\dfrac{\overline{Y}-m}{\sigma} \hookrightarrow \mathcal{N}(0,1)$ (loi normale centrée réduite).

La table de la fonction de répartition de la loi normale centrée réduite (F) permet d'estimer la moyenne avec un risque d'erreur connu : si $T \hookrightarrow \mathcal{N}(0,1)$, alors $p(|T| \leqslant u) = 2\,p(T \leqslant u) - 1 = 2\,F(u) - 1$. En appliquant ce résultat à $T = \sqrt{n}\,\dfrac{\overline{Y}-m}{\sigma}$, on obtient $p\left(m - u\,\dfrac{\sigma}{\sqrt{n}} \leqslant \overline{Y} \leqslant m + u\,\dfrac{\sigma}{\sqrt{n}}\right) = 2\,F(u) - 1$.

Par exemple, pour $u = 1,96$, $F(u) \simeq 0,97500$; donc $p\left(m - 1,96\,\dfrac{\sigma}{\sqrt{n}} \leqslant \overline{Y} \leqslant m + 1,96\,\dfrac{\sigma}{\sqrt{n}}\right) \simeq 0,95$.

Cela signifie que si l'on calculé l'espérance m de $\overline{Y}$, la probabilité que la moyenne de la population se situe dans l'intervalle $\left[m - 1,96\,\dfrac{\sigma}{\sqrt{n}}\,;\, m + 1,96\,\dfrac{\sigma}{\sqrt{n}}\right]$ est de $0,95$. Le risque d'erreur en prenant cet intervalle de confiance est donc de 5%. Le risque d'erreur passe à environ 1% si on prend $u = 2,58$ car $F(2,58) \simeq 0,99506$.

4 Tests statistiques

On se limite à l'étude de tests de comparaison de deux moyennes. On souhaite étudier l'influence d'un facteur (engrais, température...) sur un caractère donné. On considère deux échantillons de tailles n_1 et n_2, le premier provenant d'une population n'ayant subi aucun traitement, le second provenant d'une population soumise au facteur étudié. On note m_1, m_2, σ_1^2/n_1 et σ_2^2/n_2 les espérances et variances des variables aléatoires $\overline{Y}_1$ et $\overline{Y}_2$ associées à ces échantillons.

Le test a pour but de déterminer si les différences entre m_1 et m_2, σ_1^2/n_1 et σ_2^2/n_2 sont dues au facteur étudié, ou simplement aux fluctuations normales de la moyenne et de l'écart type résultant du caractère aléatoire des tirages.

4.1 Principe général

4.1.1 Formulation des hypothèses

On émet deux hypothèses, notées en général H_0 et H_1 qui jouent des rôles non symétriques. Le test vise à accepter ou rejeter H_0 qui est parfaitement définie contrairement à H_1 : par exemple, H_0 peut être une valeur donnée t_0 pour un pourcentage, une taille ou une mesure t quelconque ; H_1 pourra alors être formulée par $t \neq t_0$, ou bien $t > t_0$ ou encore $t = t_1$.

4.1.2 Risque d'erreur

Un test n'étant jamais parfaitement fiable, il faut évaluer les risques d'erreur en essayant de les réduire autant que possible.

Risque de première espèce : C'est la probabilité –notée α– de rejeter H_0 alors qu'elle est vraie. En général on choisit α *a priori* ; α est appelé le *seuil de signification* du test.

Risque de seconde espèce : C'est la probabilité –notée β– d'accepter H_0 alors qu'elle est fausse.

Bien sûr, α et β ne sont pas indépendants : si l'on fixe α, β est déterminé, de plus β augmente si on diminue α. Il s'agit donc de trouver un compromis ; mais le calcul de β en fonction de α n'est pas toujours facile. $1 - \beta$ est la *puissance* du test, c'est la probabilité de rejeter H_0 avec raison.

4.2 Comparaison de deux moyennes

La variable aléatoire $\overline{Y}_1 - \overline{Y}_2$ suit une loi normale –cours de 2e année– d'espérance $(m_1 - m_2)$ et d'écart type $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ (cours de proba sur l'espérance et la variance de la somme de deux variables aléatoires). On définit u_α par $F(u_\alpha) = 1 - (\alpha/2)$ –F est toujours la fonction de répartition de la loi normale centrée réduite– et on calcule $x = \dfrac{|m_2 - m_1|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$. Si $x \leqslant u_\alpha$, on accepte H_0, sinon on la rejette.

En fait, on montre qu'il faut remplacer les variances des deux échantillons pour le calcul ci-dessus par $\dfrac{\sigma_1^2}{n_1 - 1}$ et $\dfrac{\sigma_2^2}{n_2 - 1}$.

4.3 Comparaison de deux proportions

Le principe est le même que pour la comparaison des moyennes, on admet ici que la distribution d'échantillonage des proportions suit une loi normale $\mathcal{N}\left(p, \sqrt{(p\,q)/n}\right)$ où p est la proportion d'individus de la population possédant le caractère étudié, et $q = 1 - p$. Soient p_1 et p_2 les proportions observées pour deux échantillons de tailles n_1 et n_2 respectivement. On teste l'hypothèse H_0 : « $p_1 = p_2$ » contre l'hypothèse H_1 : « $p_1 \neq p_2$ » ou bien H_1 : « $p_1 > p_2$ ». La variable aléatoire $\overline{Y}_1 - \overline{Y}_2$ vérifie : $E(\overline{Y}_1 - \overline{Y}_2) = p_1 - p_2$ et $V(\overline{Y}_1 - \overline{Y}_2) = \dfrac{p_1\,(1-p_1)}{n_1} + \dfrac{p_2\,(1-p_2)}{n_2}$. Pour déterminer si H_0 est vérifiée (avec un risque d'erreur de première espèce égal à α), on compare la quantité $x = \dfrac{|p_2 - p_1|}{\sqrt{\dfrac{p_1\,(1-p_1)}{n_1} + \dfrac{p_2\,(1-p_2)}{n_2}}}$ à la valeur critique u_α définie au paragraphe précédent (4.2).

Si $x < u_\alpha$, on accepte l'hypothèse H_0, sinon on considère que la différence entre les proportions est significative au seuil α.

Remarque : pour que les calculs soient valables, il faut des échantillons de taille suffisante, en pratique $n\,p$ et $n\,q$ supérieurs à 5.

4.3.1 Estimation d'une proportion

Pour valider (ou non) l'hypothèse « $p = p_0$ », on cacule de même l'intervalle de confiance pour un seuil que l'on fixe : On calcule à nouveau $x = \dfrac{|p - p_0|}{\sqrt{\dfrac{p_0\,(1-p_0)}{n}}}$ et on compare la valeur à u_α...

5 Cas des petits échantillons

Lorsqu'on dispose d'un échantillon de taille insuffisante, on ne peut pas approximer une loi binômiale par une loi normale. On peut alors utiliser un test de Student à $n-1$ degrés de liberté.

En effet, la moyenne de l'échantillon étant donnée, on ne peut pas considérer que les X_i sont indépendants, puisque si on en connait $n-1$, le dernier s'en déduit à partir de la moyenne.

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes suivant des lois normales de moyennes respectives μ_i et d'écart-type σ_i ; $Y_i = \dfrac{X_i - \mu_i}{\sigma_i}$ les variables centrées et réduites associées, alors par définition la variable X , telle que $X := \displaystyle\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$ suit une loi du χ^2 à n degrés de liberté.

On a le théorème suivant :

Théorème 5.1 (Intervalle de confiance associé à l'espérance d'une variable de loi normale de variance inconnue)**.**

L'intervalle de confiance de μ au seuil de confiance α est donné par :
$\left[\overline{x} - t^{n-1}_{1-\alpha/2} \sqrt{\frac{S}{n}}, \overline{x} + t^{n-1}_{1-\alpha/2} \sqrt{\frac{S}{n}} \right]$ *où*
$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$*, l'estimateur de l'espérance.*
$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$*, l'estimateur non biaisé de la variance.*
t^k_γ *le quantile d'ordre γ de la loi de Student à k degrés de liberté.*

α (bilatéral)	50%	60%	70%	80%	90%	95%	98%	99%	99,5%	99,8%	99,9%
$1-\gamma$ (unilatéral)	75%	80%	85%	90%	95%	97,5%	99%	99,5%	99,75%	99,9%	99,95%
k											
1	1,000	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	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	0,920	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	1,440	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	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	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	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	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	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,197	3,610	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	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	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	1,060	1,319	1,714	2,069	2,500	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	2,060	2,485	2,787	3,078	3,450	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	3,690
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	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	0,851	1,050	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	2,000	2,390	2,660	2,915	3,232	3,460
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291