Tests de Student-Fisher

Matthieu Kowalski

1 Rappels généraux sur les tests

1.1 Un exemple

Un producteur d'ampoules flash, qui ne peuvent servir qu'une seule fois, affirme à ses clients que la proportion p d'ampoules défectueuses est inférieure à 10%. Pour éviter que les lots vendus ne lui soient renvoyés, il veut s'assurer de ses dires.

Pour cela, il prélève un échantillons de $n = 1000$ ampoules de sa chaîne de production. Le but est d'évaluer p à partir de cet échantillon. Il note X la variable aléatoire qui représente le nombre d'ampoules défectueuses sur 1000 ampoules testées. Pour estimer p, il considère chaque ampoule comme une variable aléatoire $X_i \sim \mathcal{B}(p)$ de sorte que

$$\begin{cases} X_i = 1 & \text{si l'ampoule } i \text{ est défecteuse} \\ X_i = 0 & \text{sinon.} \end{cases}$$

Ainsi, $X = \sum\limits_{i=1}^{n} X_i$ suit la loi $\mathcal{B}(n,p)$.

Intuitivement, si $\bar{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ est "grand", alors il y a peu de chance que p soit inférieur à 10%. Le producteur peut donc prendre comme règle de décision :

- Si $\bar{X}_n > t$ ("$\bar{X}_n$ assez grand"), alors $p > 10\%$;
- Si $\bar{X}_n > \leq t$ ("$\bar{X}_n$ assez petit"), alors $p \leq 10\%$.

où t est un nombre à fixer. Le problème est donc :

Comment choisir t convenablement ?

Un premier choix naturel est de prendre $t = 10\%$. La règle de décision serait alors :

- Si $\bar{X}_n > 10\%$, alors $p > 10\%$;
- Si $\bar{X}_n > \leq 10\%$, alors $p \leq 10\%$.

On calcule dans la suite les probabilités de se tromper si on adopte cette règle. Une première façon de se tromper est de décider que $p > 10\%$ alors que ce n'est pas le cas :

$$\alpha_1 = \sup_{p \leq 10\%} \mathbb{P}\{\bar{X}_n > 10\%\} = \sup_{p \leq 10\%} \mathbb{P}\{\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} > \frac{10\% - p}{\sqrt{p(1-p)/n}}\} .$$

Ici, on cherche à trouver la plus grande probabilité (ie le plus gros risque) de se tromper (ie $\bar{X}_n > 10\%$) quand l'hypothèse $p \leq 10\%$ est vérifiée.

On applique le TLC à $Z = \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$. La fonction $p \in [0, 10\%] \mapsto \mathbb{P}\left\{Z > \frac{10\% - p}{\sqrt{p(1-p)/n}}\right\}$ est croissante, et donc le sup est atteint pour $p = 10\%$. On obtient donc

$$\alpha \approx 50\% \ .$$

Dans le pire des cas, on a donc 50% de risque de se tromper en décidant que $p > 10\%$ avec cette stratégie.

L'autre erreur possible, est de décider que $p \leq 10\%$ alors qu'en réalité $p < 10\%$:

$$\alpha_2 = \sup_{p > 10\%} \mathbb{P}\{\bar{X}_n \leq 10\%\} \ .$$

Par un raisonnement similaire, on trouve :

$$\alpha_2 \approx 50\% \ .$$

On se trompe alors ici aussi une fois sur deux en décidant que $p < 10\%$ avec cette règle de décision.

La règle de décision initiale est donc très mauvaise : on se trompe une fois sur deux dans les deux cas ! Pour le producteur d'ampoule, conclure que $p > 10\%$ lorsque ce n'est pas le cas aura des conséquences moins grave que de conclure $p \leq 10\%$ si $p > 10\%$. En effet, dans le premier cas, il ne prendra pas le risque d'envoyer le lot d'ampoules, alors que dans le second, beaucoup de lots lui seront retournés (et ses clients ne seront pas satisfaits).

Si on reprent les calculs précédents de façon plus générale, en gardant un seuil t quelconque, on a :

$$\begin{aligned}
\alpha_1 &= \sup_{p \leq 10\%} \mathbb{P}\{\bar{X}_n > t\} \approx \sup_{p \leq 10\%} \mathbb{P}\{Z > \frac{t-p}{\sqrt{p(1-p)/n}}\},\ Z \sim \mathcal{N}(0,1) \ , \\
\alpha_2 &= \sup_{p > 10\%} \mathbb{P}\{\bar{X}_n \leq t\} \approx \sup_{p > 10\%} \mathbb{P}\{Z \leq \frac{t-p}{\sqrt{p(1-p)/n}}\},\ Z \sim \mathcal{N}(0,1) \ .
\end{aligned}$$

Ainsi, plut t est choisi grand, plus α_1 devient petit, mais plus α_2 devient grand. Les deux erreurs ne peuvent pas être (en général) petites en même temps.

Il faut choisir de privilégier une hypothèse par rapport à l'autre. On notera cette hypothèse (H_0). La seconde sera notée (H_1). Dans cet exemple, on teste $(H_0)\ :\ p \geq 10\%$ contre $(H_1)\ :\ p < 10\%$. Par définition de (H_0) et (H_1), l'erreur qu'on contrôle est

$$\alpha = \mathbb{P}_{(H_0)}\{\text{Rejeter } (H_0)\} \ .$$

α s'appelle le niveau du test, ou l'erreur de première espèce. La règle de décision est la suivante :

$$\begin{array}{llll} \text{Si} & \bar{X}_n < t & \text{Rejet de} & (H_0)\ . \\ \text{Si} & \bar{X}_n \geq t & \text{Acceptation de} & (H_0)\ . \end{array}$$

Si on prend $\alpha = 5\%$, le calcul de t donne :

$$\begin{aligned} 5\% &= \sup_{p \geq 10\%} \mathbb{P}(\bar{X}_n < t) \\ &\approx \sup_{p \geq 10\%} \mathbb{P}\left\{ Z \leq \frac{t-p}{\sqrt{p(1-p)/n}} \right\} \\ &\approx \mathbb{P}\left\{ Z \leq \frac{t-0.1}{\sqrt{0.1(0.9)/n}} \right\}\ . \end{aligned}$$

La table de la loi normale donne

$$t = 0.1 - 1.645\sqrt{0.09/n}\ .$$

α fixe la règle de décision, et on ne peut plus contrôler l'erreur de deuxième type $\mathbb{P}_{(H_1)}\{\text{accepter } (H_0)\}$. Ici, le calcul de cette erreur donne 95%.

1.2 Étapes de construction d'un test

1. Choix de l'hypothèse privilégiée (H_0) en fonction de l'erreur qu'on veut contrôler : $\alpha = \mathbb{P}_{(H_0)}\{\text{rejeter } H_0\}$.
2. Choix du niveau α (petit).
3. Construction d'une région de rejet de (H_0) de niveau α. On la construit en générale sur un estimateur $\hat{\theta}$ du paramètre θ sur lequel porte le test. Cette région est du type $\mathcal{R} = \left\{\hat{\theta} \in R\right\}$. Dans l'exemple précédent, on a $\mathcal{R} = \left\{\bar{X}_n < t\right\}$.
4. Conclusion au vu de l'échantillon selon la règle de décision :
   - Si $\hat{\theta} \in R$, rejet de (H_0) ;
   - Si $\hat{\theta} \notin R$, acceptation de (H_0) ;
5. Sin on accepte (H_0), calcul éventuel de l'erreur de deuxième espèce.

1.3 P-valeur

En pratique, plutôt que de fixer un niveau α et d'en déduire la région le t associée, on peut calculer la P-valeur du test. Dans notre exemple, supposons qu'on observe $\bar{X}_n = \bar{x}_n = 0.09$. La P-valeur du test est donnée par

$$\mathbb{P}_{(H_0)}\{\bar{X}_n < \bar{x}_n\} = \mathbb{P}_{p=0.1}\{\bar{X}_n < 0.09\}\ .$$

La P-valeur est probabilité de rejeter à tort (H_0) quand t est remplacé par la valeur observé sur l'échantillon. La P-valeur nous dit si ce que nous observons sur l'échantillon de données est probable sous l'hypothèse (H_0).

Ainsi, la règle de décision finale est donnée quelque soit le niveau α choisi par

1. Si $\alpha >$P-valeur, rejet de (H_0).
2. Si $\alpha \leq$P-valeur, acceptation de (H_0).

2 Test de Fisher

Soit deux échantillons X et Y suivant respectivement une loi normale $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$ et de taille respective n_X et n_Y.

Le but du test de Fisher est de fournir une règle de décision vis à vis de l'égalité des deux variance σ_X et σ_Y.

Rappel : on utilise l'estimateur de la variance convergent et asymtotiquement sans biais

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_n - \bar{X}_n)^2 \ .$$

La statistique du test de Fisher est

$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \ .$$

Ce rapport suit une loi de Fisher, ie $F \sim \mathcal{F}_{n_X-1, n_Y-1}$.

Intuitivement, F doit être proche de 1 sous H_0 et largement plus petite ou plus grande sous H_1. On considère un test de région de rejet du type

$$R = \{F < k_1\} \cup \{F > k_2\} \ ,$$

ou, de manière équivalente, de région d'acceptation

$$\bar{R} = \{k_1 < F < k_2\} \ .$$

2.1 Test unilatéral à droite

On considère le problème de test :

$$\begin{cases} H_0 & : \sigma_X^2 = \sigma_Y^2 \\ H_1 & : \sigma_X^2 > \sigma_Y^2 \ . \end{cases}$$

Dans ce cas, on rejette H_0 lorsque le rapport des variances devient grand. On cherche la région d'acceptation sous la forme

$$\bar{R} = \{F < k_2\} \ .$$

On résout l'équation

$$\alpha = \mathbb{P}_{H_0}\{F > k_2\} \ ,$$

ce qui donne

$$k_2 = \mathcal{F}_{n_X-1,n_Y-1;1-\alpha} \ .$$

La règle décision est donc
– Si $F > k_2$, rejet de H_0
– Si $F < k_2$, acceptation de H_0 .
La P-valeur du test est donnée par :

$$\text{P-valeur} = \mathbb{P}\{Z > F\} = 1 - \mathbb{P}\{Z \leq F\} \ ,$$

où $Z \sim \mathcal{F}_{n_X-1,n_Y-1}$. La règle de décision associée est :

1. Si $\alpha >$ P-valeur, rejet de H_0.
2. Si $\alpha \leq$ P-valeur, rejet de H_0.

2.2 Test unilatéral à gauche

On considère le problème de test :

$$\begin{cases} H_0 & : \sigma_X^2 = \sigma_Y^2 \\ H_1 & : \sigma_X^2 < \sigma_Y^2 \ . \end{cases}$$

Dans ce cas, on rejette H_0 lorsque le rapport des variances devient petit. On cherche la région d'acceptation sous la forme

$$\bar{R} = \{F > k_1\} \ ,$$

On résout l'équation

$$\alpha = \mathbb{P}_{H_0}\{F < k_1\} \ ,$$

ce qui donne

$$k_1 = \mathcal{F}_{n_X-1,n_Y-1;\alpha} \ .$$

La règle décision est donc
– Si $F < k_1$, rejet de H_0.
– Si $F > k_1$, acceptation de H_0.
La P-valeur du test est donnée par :

$$\text{P-valeur} = \mathbb{P}\{Z < F\}$$

où $Z \sim \mathcal{F}_{n_X-1,n_Y-1}$. La règle de décision associée est :

1. Si $\alpha >$ P-valeur, rejet de H_0
2. Si $\alpha \leq$ P-valeur, rejet de H_0

2.3 Test bilatéral

On considère le problème de test :

$$\begin{cases} H_0 & : \sigma_X^2 = \sigma_Y^2 \\ H_1 & : \sigma_X^2 \neq \sigma_Y^2 \end{cases} .$$

On cherche k_1 et k_2 tels que

$$\alpha = \mathbb{P}_{H_0}\{F < k1\} \quad \text{et} \quad \mathbb{P}_{H_0}\{F > k_2\} .$$

On impose de plus $\mathbb{P}_{H_0}\{F < k1\} = \mathbb{P}_{H_0}\{F > k_2\}$, ce qui donne

$$k_1 = \mathcal{F}_{n_X-1,n_Y-1;\frac{\alpha}{2}} \quad \text{et} \quad k_2 = \mathcal{F}_{n_X-1,n_Y-1;1-\frac{\alpha}{2}} .$$

La règle décision est donc

- Si $F < k_1$ ou $F > k_2$, rejet de H_0.
- Si $k_1 < F < k_2$, acceptation de H_0.

Soit $Z \sim \mathcal{F}_{n_X-1,n_Y-1}$. On peut définir deux types de $P-valeurs$:

1. Accepter un test bivarié revient à accepter les deux tests unilatéraux. Si on se réfère à la définition de la P-valeur pour les tests unilatéraux, doit avoir $\mathbb{P}Z < F > \frac{\alpha}{2}$ et $\mathbb{P}Z > F > \frac{\alpha}{2}$ pour accepter le test bilatéral. En prenant $P - valeur = \mathbb{P}Z < F$, on obtient comme règle de décision :
   - Si $\frac{\alpha}{2} >$ P-valeur ou $\frac{\alpha}{2} <$ P-valeur, rejet de H_0.
   - Si $\frac{\alpha}{2} \leq$ P-valeur $\leq 1 - \frac{\alpha}{2}$, rejet de H_0.
2. Une autre façon de définir la P-valeur (*qui est celle utilisée par* R) dans le cas d'un test bilatéral est

   $$\text{P-valeur} = \begin{cases} 2\mathbb{P}\{Z < F\} & \text{si} \quad \mathbb{P}\{Z < F\} \leq \frac{1}{2} \\ 2(1 - \mathbb{P}\{Z < F\}) & \text{si} \quad \mathbb{P}\{Z < F\} \geq \frac{1}{2} \end{cases}$$

   Et la règle de décision associée est
   - Si $\alpha >$ P-valeur, rejet de H_0.
   - Si $\alpha \leq$ P-valeur, rejet de H_0.

3 Test de Student

On considère deux échantillons X et Y suivant respectivement une loi normale $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$ et de taille respective n_X et n_Y.

Le but du test de Student est de fournir une règle de décision vis à vis de l'égalité des deux moyennes μ_X et μ_Y, *lorsque les variances sont inconnues mais égales*, ie $\sigma_X^2 = \sigma_T^2 = \sigma$.

La statistique du test de Student est

$$T = \frac{\hat{\mu}_X \bar{-} \hat{\mu}_Y}{\hat{\sigma}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} ,$$

avec

$$\begin{aligned}\hat{\sigma} &= \frac{\sum_{k=1}^{n_X}(X_k - \hat{\mu}_X)^2 + \sum_{k=1}^{n_Y}(Y_k - \hat{\mu}_Y)^2}{n_X + n_Y - 2} \\ &= \frac{n_X - 1}{n_X + n_Y - 2}\hat{\sigma}_X^2 + \frac{n_Y - 1}{n_X + n_Y - 2}\hat{\sigma}_Y^2 \ .\end{aligned}$$

T suit une loi de student-t de $n_X + n_Y - 2$ degrés de liberté.

Intuitivement, T doit être proche de 0 sous H_0, et beaucoup plus grande ou plus petite sous H_1. On considère un test de région de rejet du type

$$R = \{|T| > t\} \ ,$$

ou, de manière équivalente, de région d'acceptation

$$\bar{R} = \{|T| < t\} \ .$$

3.1 Test unilatéral à droite

On considère le problème de test :

$$\begin{cases} H_0 & : \mu_X = \mu_Y \\ H_1 & : \mu_X > \mu_Y \ . \end{cases}$$

Dans ce cas, on rejette H_0 lorsque T devient grand. On cherche la région d'acceptation sous la forme

$$\bar{R} = \{T < t\} \ .$$

On résout l'équation

$$\alpha = \mathbb{P}_{H_0}\{T > t\} \ ,$$

ce qui donne

$$t = t_{n_X+n_Y-2;1-\alpha} \ .$$

La règle décision est donc

- Si $T > t$, rejet de H_0.
- Si $T < t$, acceptation de H_0.

3.2 Test unilatéral à gauche

On considère le problème de test :

$$\begin{cases} H_0 & : \mu_X = \mu_Y \\ H_1 & : \mu_X < \mu_Y \ . \end{cases}$$

Dans ce cas, on rejette H_0 lorsque T devient très négatif. On cherche la région d'acceptation sous la forme

$$\bar{R} = \{T > t\} \ .$$

On résout l'équation

$$\alpha = \mathbb{P}_{H_0}\{T < t\} \ ,$$

ce qui donne

$$t = t_{n_X+n_Y-2;\alpha} = -t_{n_X+n_Y-2;1-\alpha} \ .$$

La règle décision est donc
- Si $T < t$, rejet de H_0.
- Si $T > t$, acceptation de H_0.

3.3 Test bilatéral

On considère le problème de test :

$$\begin{cases} H_0 & : \mu_X = \mu_Y \\ H_1 & : \mu_X \neq \mu_Y \end{cases}$$

Dans ce cas, on rejette H_0 lorsque T devient très négatif. On cherche la région d'acceptation sous la forme

$$\bar{R} = \{|T| < t\} \ .$$

On résout l'équation

$$\alpha = \mathbb{P}_{H_0}\{|T| > t\} \ ,$$

ce qui donne

$$t = t_{n_X+n_Y-2;1-\frac{\alpha}{2}} \ .$$

La règle décision est donc
- Si $|T| > t$, rejet de H_0.
- Si $|T| < t$, acceptation de H_0.

Soit $Z \sim t_{n_X-1,n_Y-1}$. On peut définir deux types de $P - valeurs$:

1. Accepter un test bivarié revient à accepter les deux tests unilatéraux. Si on se réfère à la définition de la P-valeur pour les tests unilatéraux, doit avoir $\mathbb{P}Z < F > \frac{\alpha}{2}$ et $\mathbb{P}Z > F > \frac{\alpha}{2}$ pour accepter le test bilatéral. En prenant $P - valeur = \mathbb{P}Z < F$, on obtient comme règle de décision :
   - Si $\frac{\alpha}{2} >$ P-valeur ou $\frac{\alpha}{2} <$ P-valeur, rejet de H_0.
   - Si $\frac{\alpha}{2} \leq$ P-valeur $\leq 1 - \frac{\alpha}{2}$, rejet de H_0.
2. Une autre façon de définir la P-valeur (*qui est celle utilisée par* R) dans le cas d'un test bilatéral est

$$\text{P-valeur} = \begin{cases} 2\mathbb{P}\{Z < F\} & \text{si} \quad \mathbb{P}\{Z < F\} \leq \frac{1}{2} \\ 2(1 - \mathbb{P}\{Z < F\}) & \text{si} \quad \mathbb{P}\{Z < F\} \geq \frac{1}{2} \end{cases}$$

   Et la règle de décision associée est
   - Si $\alpha >$ P-valeur, rejet de H_0.
   - Si $\alpha \leq$ P-valeur, rejet de H_0.