Introduction à la statistique - $2^{\text{ème}}$ partie

Les tests

Préparation au Capes - Université Rennes 1

Hélène Guérin, helene.guerin@univ-rennes1.fr

Idée : On estime un paramètre inconnu θ à l'aide d'observations $(X_1, ..., X_n)$. Notre intuition ou un autre échantillon, nous laisse à penser que la valeur réelle de θ est une certaine valeur connue θ_0. La question qui se pose alors est : est-ce que cette valeur théorique θ_0 est proche de la réalité ? Pour y répondre, on va tester à l'aide de l'échantillon $(X_1, ..., X_n)$ si l'hypothèse "$\theta = \theta_0$" est réaliste. Ce test permet de comparer la valeur liée à l'expérience à une certaine valeur désirée θ_0.

On considère un échantillon $X_1, X_2, \ldots$ dont la loi $\mathbb{P}_\theta$ dépend du paramètre inconnu θ.

Définition

Faire un test de l'**hypothèse nulle** H_0 contre l'**hypothèse alternative** H_1 au niveau de risque α (de l'ordre de 5%), c'est fixer une **zone de rejet** $D = D(X_1, ..., X_n)$ ne dépendant que de l'échantillon, telle que

$$\mathbb{P}_{H_0}(D) = \sup_{\theta \in H_0} \mathbb{P}_\theta(D) = \alpha.$$

Alors la conclusion du test est la suivante :

si D est réalisé, on rejette l'hypothèse H_0

si D n'est pas réalisé, on accepte l'hypothèse H_0.

Il y a deux erreurs possibles :

• **Erreur de $1^{\text{ère}}$ espèce** : Rejetter à tord l'hypothèse H_0.
La probabilité de cet événement est $\mathbb{P}_{H_0}(D)$, c'est à dire α.

• **Erreur de 2^{nde} espèce** : Accepter H_0 alors que l'hypothèse est fausse.
La probabilité de cet événement est $\beta = \mathbb{P}_{H_1}(\overline{D}) = \sup_{\theta \in H_1} \mathbb{P}_\theta(D)$.

La première erreur α étant en générale fixée, une bonne zone de rejet minimise la valeur de β.

Remarque Plus on diminue le niveau de risque α, plus on diminue la probabilité de se tromper, i.e. de rejeter H_0 alors que H_0 est vrai, plus on diminue la taille de la zone de rejet.

Comment trouver une zone de rejet ?

On limite toujours ce cours au cas où le paramètre inconnu θ est la moyenne $m = E[X_1]$ et où $\sigma^2 = Var(X_1)$ existe.

I Cas où H_0 est une hypothèse simple

Soit m_0 une valeur connue.

On souhaite tester l'hypothèse nulle $H_0 : m = m_0$ contre l'hypothèse alternative $H_1 : m \neq m_0$ au niveau de risque α. Il faut donc trouver une zone de rejet D telle que $I\!P_{H_0}(D) = P_{m_0}(D) = \alpha$.

$\rightsquigarrow$ **Utilisation des intervalles de confiance**

Soit I un intervalle de confiance pour m au niveau $1-\alpha$: $\forall m$ $I\!P_m(m \in I) = 1-\alpha$. Posons $D = \{m_0 \notin I\}$.
Alors D est un test de $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ au niveau de risque α.

On choisit t_α tel que $\mathbb{P}(|Z| > t_\alpha) = \alpha$.

Si σ^2 est connu, on choisit comme zone de rejet

$$D = \left\{ |\overline{X}_n - m_0| > t_\alpha \sqrt{\frac{\sigma^2}{n}} \right\}$$

Si σ^2 est inconnu, on l'estime par la variance empirique $\hat{\sigma}^2$ et on choisit la zone de rejet

$$D = \left\{ |\overline{X}_n - m_0| > t_\alpha \sqrt{\frac{\hat{\sigma}^2}{n}} \right\}$$

Exemple Un sondage dénombre 221023 garçons sur 429440 naissances. Y a t'il d'équiprobabilité de naissance des garçons et des filles ?

On estime la proportion p de naissance de garçon par la proportion de l'échantillon $\hat{p} = \frac{221023}{429440}$.

Il est évident qu'on va rejetter l'hypothèse d'équiprobabilité si la proportion estimée est trop petite ou trop grande, donc on va chercher une zone de rejet de la forme $D = \{\hat{p} \notin [a, b]\}$.

L'intervalle $I = [0.512, 0.517]$ est un intervalle de confiance à 95%, alors $D = \{1/2 \notin I\}$ est une zone de rejet au niveau de risque 5%.

On est dans la zone de rejet. Donc on rejette l'hypothèse d'équiprobabilité des naissances.

II Cas où H_0 est une hypothèse composite

Soit m_0 une valeur connue.

On souhaite tester $H_0 : m \leq m_0$ contre $H_1 : m > m_0$ au niveau de risque α. Il faut donc trouver une zone de rejet D telle que $\mathbb{P}_{H_0}(D) = \sup_{m \leq m_0} \mathbb{P}_m(D) = \alpha$.

Plus généralement, on peut vouloir tester

- $H_0 : m \geq m_0 \quad \Rightarrow \quad \alpha = \sup_{m \geq m_0} \mathbb{P}_m(D)$

- $H_0 : m \in [m_1, m_2] \quad \Rightarrow \quad \alpha = \sup_{m \in [m_1, m_2]} \mathbb{P}_m(D) \ldots$

Remarque Il n'est en général pas simple de trouver une zone de rejet qui vérifie $\mathbb{P}_{H_0}(D) = \alpha$. Dans ce cas, on cherche une zone de rejet qui a une erreur de première espèce inférieure : $\mathbb{P}_{H_0}(D) \leq \alpha$.

Exemple initial L'entreprise n'accepte le lot de pièces mécaniques que si la proportion de pièces défectueuses est inférieure à 5%. Si on note p la proportion de pièces défectueuses dans tout le lot, on souhaite donc tester $H_0 : p \leq 5\%$ contre $H_1 : p > 5\%$. Fixons le niveau de risque à $\alpha = 1\%$.

Il est naturel de rejetter l'hypothèse H_0 si la proportion de pièces défectueuses de l'échantillon $\hat{p}$ est trop importante. On va donc chercher une zone de rejet de la forme $D = \{\hat{p} \geq d\}$.

Pour $d = 0.086$, on a $\sup_{p \leq 0.05} I\!P_p(D) \leq 0.01$.

Comme $\hat{p} = 0.075$, on n'est pas dans la zone de rejet. L'entreprise accepte le lot.

III - Égalité des moyennes pour des échantillons non indépendants

On considère n individus sur lesquels on réalise deux séries de mesures concernant les variables X et Y : par exemple X est la quantité observée avant traitement et Y est la même quantité après traitement.

On se demande si le traitement a eu un effet sur la moyenne. On veut par conséquent effectuer le test suivant :

$$H_0 : E[X] = E[Y] \text{ contre } H_1 : E[X] \neq E[Y].$$

Les variables X et Y étant mesurées sur les mêmes individus, elles sont dépendantes. On se ramène au cadre d'un échantillon de variables indépendantes et de même loi en considérant la variable Z définie par

$$Z_1 = Y_1 - X_1, \ldots, Z_n = Y_n - X_n$$

On effectue alors le test : $H_0 : E[Z] = 0$ contre $H_1 : E[Z] \neq 0$ en utilisant les notions introduites dans la section I.

Remarque Si on veut tester

$$H_0 : E[X] \leq E[Y] \quad \text{contre} \quad H_1 : E[X] > E[Y],$$

alors on effectue le test $H_0 : E[Z] \leq 0$ contre $H_1 : E[Z] > 0$ en utilisant les notions introduites dans la Section II.

IV -Comparaison de 2 moyennes pour des échantillons indépendants

On veut comparer deux séries statistiques observées sur deux populations distinctes, ou bien aux résultats obtenus dans deux expériences réalisées indépendamment sur une même population.

On a donc deux échantillons indépendants :

$X_1, \ldots, X_{n_1}$ des variables iid, avec $E[X_1] = m_1$, $Var(X_1) = \sigma_1^2$

$Y_1, \ldots, Y_{n_2}$ des variables iid, avec $E[Y_2] = m_2$, $Var(Y_2) = \sigma_2^2$

On veut tester $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$ au niveau de risque α. D'après le théorème centrale limite, on a

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \underset{n \to +\infty}{\longrightarrow} Z \text{ avec } Z \sim \mathcal{N}(0,1)$$

On choisit t_α tel que $\mathbb{P}(|Z| > t_\alpha) = \alpha$.

Si σ_1^2 et σ_2^2 sont connus, on choisit comme zone de rejet

$$D = \left\{ \frac{|\overline{X}_{n_1} - \overline{Y}_{n_2}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > t_\alpha \right\}$$

Si σ_1^2 et σ_2^2 sont inconnus, on les estime par les variances empiriques $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ et on choisit la zone de rejet

$$D = \left\{ \frac{|\overline{X}_{n_1} - \overline{Y}_{n_2}|}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} > t_\alpha \right\}$$

Exemple Age des étudiants de Capes.

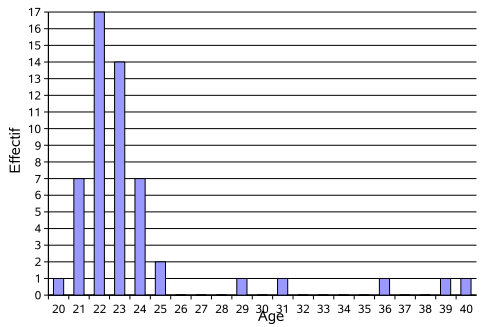
On suppose qu'il n'y a pas de redoublement !

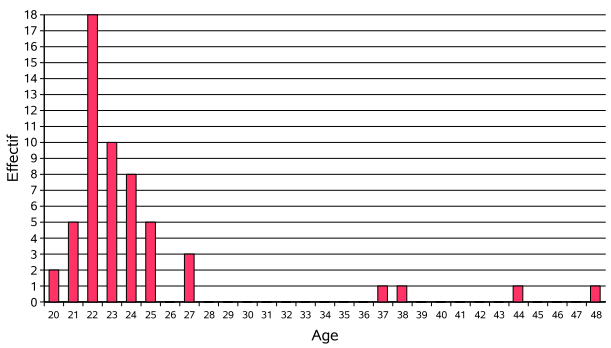
En 2005-2006, on a relevé les âges suivant sur un échantillon de 53 personnes :

Age	20	21	22	23	24	25	29	31	36	39	40
Effectif	1	7	17	14	7	2	1	1	1	1	1

En 2006-2007, on a relevé les âges suivant sur un échantillon de 55 personnes :

Age	20	21	22	23	24	25	27	37	38	44	48
Effectif	2	5	18	10	8	5	3	1	1	1	1





Effectif en 2005-2006

$\overline{X} = 23.7$, $\hat{\sigma}_X^2 = 16.5$

Effectif en 2006-2007

$\overline{Y} = 24.3$, $\hat{\sigma}_Y^2 = 28.3$

On teste au niveau de risque 5% si la moyenne d'âge a évoluée :

$$H_0 : m_X = m_Y \quad \text{contre} \quad H_1 : m_X \neq m_Y.$$

La zone de rejet est alors $D = \left\{ |\overline{X} - \overline{Y}| > 1.96\sqrt{\frac{\hat{\sigma}_X^2}{53} + \frac{\hat{\sigma}_Y^2}{55}} \right\}$.

Au vu des sondages : $|\overline{X} - \overline{Y}| = 0.6$ et $1.96\sqrt{\frac{\hat{\sigma}_X^2}{53} + \frac{\hat{\sigma}_Y^2}{55}} = 1.78$.

On accepte donc H_0 : l'âge moyen des étudiants de capes n'a pas évolué entre 05/06 et 06/07.