Probabilités et Statistiques

Tests Statistiques

Julian Tugaut

Télécom Saint-Étienne

Sommaire

Plan

Définition

Définition

Un test d'hypothèse est un algorithme qui permet de décider entre deux hypothèses. Pour ce faire, le test d'hypothèse analyse une statistique élaborée à partir d'un échantillon aléatoire, tout en quantifiant le risque d'erreur. Ou plutôt : les risques d'erreurs.

Définition

Définition

Un test d'hypothèse est un algorithme qui permet de décider entre deux hypothèses. Pour ce faire, le test d'hypothèse analyse une statistique élaborée à partir d'un échantillon aléatoire, tout en quantifiant le risque d'erreur. Ou plutôt : les risques d'erreurs.

Remarque

Il y a deux grandes familles de tests statistiques : les tests paramétriques et les tests non paramétriques.

Tests paramétriques

Les tests paramétriques présupposent que l'on connaît le genre de loi que suivent les variables aléatoires sous-jacentes : Bernoulli $\mathcal{B}(p)$, Poisson $\mathcal{P}(\lambda)$, normale $\mathcal{N}(m, \sigma^2)$...

Tests paramétriques

Les tests paramétriques présupposent que l'on connaît le genre de loi que suivent les variables aléatoires sous-jacentes : Bernoulli $\mathcal{B}(p)$, Poisson $\mathcal{P}(\lambda)$, normale $\mathcal{N}(m, \sigma^2)$...

Ces hypothèses sont généralement difficiles à vérifier (et nécessitent donc un test non paramétrique pour commencer). Il convient de noter que les tests paramétriques sont puissants. Ainsi, dans le cas des grands échantillons ou si l'on est exactement dans le modèle gaussien, il est facile de décider si oui ou non les paramètres proposés sont les bons.

Tests non paramétriques

Les tests non paramétriques ne présupposent rien. Il n'y a en particulier aucune hypothèse de normalité à satisfaire au préalable. Toutefois, leur puissance est plus modérée.

Tests non paramétriques

Les tests non paramétriques ne présupposent rien. Il n'y a en particulier aucune hypothèse de normalité à satisfaire au préalable. Toutefois, leur puissance est plus modérée.

De fait, on les emploie quand on ne peut faire d'autres tests. Leur avantage est de pouvoir être utilisés même avec des échantillons de faible taille.

Test de conformité

Pour commencer, on veut savoir si une valeur pré-établie du paramètre de la loi sous-jacente est bonne ou non. On confronte alors la valeur théorique de la statistique liée à cette valeur pré-établie à la valeur pratique de la statistique prise sur un échantillon.

Test de conformité

Pour commencer, on veut savoir si une valeur pré-établie du paramètre de la loi sous-jacente est bonne ou non. On confronte alors la valeur théorique de la statistique liée à cette valeur pré-établie à la valeur pratique de la statistique prise sur un échantillon.

Notamment, on peut effectuer des tests sur la moyenne, la variance, la proportion...

Test d'ajustement

Ici, on vérifie si la distribution empirique dans l'échantillon est représentée de façon crédible ou non par une distribution pré-établie. En particulier, on peut effectuer un test de normalité pour savoir si la distribution respecte bien l'hypothèse de normalité. Suite à cela, un test paramétrique peut être appliqué.

Test d'ajustement

Ici, on vérifie si la distribution empirique dans l'échantillon est représentée de façon crédible ou non par une distribution pré-établie. En particulier, on peut effectuer un test de normalité pour savoir si la distribution respecte bien l'hypothèse de normalité. Suite à cela, un test paramétrique peut être appliqué.

Il convient de noter que la loi sous-jacente peut être continue ou discrète.

Test d'ajustement

Ici, on vérifie si la distribution empirique dans l'échantillon est représentée de façon crédible ou non par une distribution pré-établie. En particulier, on peut effectuer un test de normalité pour savoir si la distribution respecte bien l'hypothèse de normalité. Suite à cela, un test paramétrique peut être appliqué.

Il convient de noter que la loi sous-jacente peut être continue ou discrète.

On parle aussi de test d'adéquation.

Test d'homogénéité

Dans un test d'homogénéité, on essaie de vérifier si oui ou non divers échantillons sont issus d'une même population et donc si l'hypothèse d'homogénéité est vérifiée. Dit autrement, on cherche à vérifier si la loi sous-jacente des différentes variables aléatoires est commune aux différents échantillons.

Test d'homogénéité

Dans un test d'homogénéité, on essaie de vérifier si oui ou non divers échantillons sont issus d'une même population et donc si l'hypothèse d'homogénéité est vérifiée. Dit autrement, on cherche à vérifier si la loi sous-jacente des différentes variables aléatoires est commune aux différents échantillons.

On parle aussi de test de comparaison.

Test d'indépendance

Le dernier type de test que nous présentons consiste à décider si deux variables aléatoires sous-jacentes sont indépendantes dans l'échantillon considéré. Dit autrement, les deux caractères sont-ils indépendants (la non-validité du Théorème donné en statistiques descriptives vient en fait des fluctuations d'échantillonnage) ou non ?

Test d'indépendance

Le dernier type de test que nous présentons consiste à décider si deux variables aléatoires sous-jacentes sont indépendantes dans l'échantillon considéré. Dit autrement, les deux caractères sont-ils indépendants (la non-validité du Théorème donné en statistiques descriptives vient en fait des fluctuations d'échantillonnage) ou non ?

On parle aussi de test d'association.

Plan

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).
4. Établir la distribution de ladite statistique.

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).
4. Établir la distribution de ladite statistique.
5. Choisir le niveau de signification du test, α.

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).
4. Établir la distribution de ladite statistique.
5. Choisir le niveau de signification du test, α.
6. Déterminer la région critique associée.

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).
4. Établir la distribution de ladite statistique.
5. Choisir le niveau de signification du test, α.
6. Déterminer la région critique associée.
7. Calculer la valeur empirique de la statistique (dans l'échantillon).

Liste des étapes

L'algorithme régissant la mise en place d'un test d'hypothèse est une succession d'étapes simples, précises et chacune de ces étapes est d'une grande importance :

1. Formuler l'hypothèse nulle (H_0).
2. Définir l'hypothèse alternative (H_1).
3. Contrôler avec une statistique liée à (H_0) (et à (H_1) de fait).
4. Établir la distribution de ladite statistique.
5. Choisir le niveau de signification du test, α.
6. Déterminer la région critique associée.
7. Calculer la valeur empirique de la statistique (dans l'échantillon).
8. Décider en fonction des étapes **6.** et **7.**.

Formuler (H_0)

La première étape consiste à définir l'hypothèse nulle, que l'on note (H_0). Il s'agit de l'hypothèse que l'on cherche à contrôler. Cela peut être une conformité à la loi normale par exemple. L'hypothèse nulle consiste à dire qu'il n'y a pas suffisamment de différence entre les valeurs théoriques et les valeurs empiriques pour affirmer que l'hypothèse que l'on a formulée est fausse.

Formuler (H_0)

La première étape consiste à définir l'hypothèse nulle, que l'on note (H_0). Il s'agit de l'hypothèse que l'on cherche à contrôler. Cela peut être une conformité à la loi normale par exemple. L'hypothèse nulle consiste à dire qu'il n'y a pas suffisamment de différence entre les valeurs théoriques et les valeurs empiriques pour affirmer que l'hypothèse que l'on a formulée est fausse.

Il est important de comprendre que l'objectif du test d'hypothèse n'est pas d'accepter (H_0) mais plutôt de ne pas la rejeter ; ou le cas contraire de la rejeter.

Définir (H_1)

Une fois formulée l'hypothèse nulle (H_0), une manière classique de procéder est de prendre comme hypothèse alternative (H_1) le complémentaire de (H_0). Cependant, cette dernière n'est pas obligatoirement le complémentaire de (H_0).

Définir (H_1)

Une fois formulée l'hypothèse nulle (H_0), une manière classique de procéder est de prendre comme hypothèse alternative (H_1) le complémentaire de (H_0). Cependant, cette dernière n'est pas obligatoirement le complémentaire de (H_0).

Par exemple, si l'on fait un test sur une proportion et que les deux seules proportions plausibles sont $p := \frac{2}{7}$ et $p := \frac{4}{7}$, alors si (H_0) correspond à $p = \frac{2}{7}$, il s'ensuit que (H_1) n'est pas $p \neq \frac{2}{7}$ mais $p = \frac{4}{7}$.

Définir (H_1)

Une fois formulée l'hypothèse nulle (H_0), une manière classique de procéder est de prendre comme hypothèse alternative (H_1) le complémentaire de (H_0). Cependant, cette dernière n'est pas obligatoirement le complémentaire de (H_0).

Par exemple, si l'on fait un test sur une proportion et que les deux seules proportions plausibles sont $p := \frac{2}{7}$ et $p := \frac{4}{7}$, alors si (H_0) correspond à $p = \frac{2}{7}$, il s'ensuit que (H_1) n'est pas $p \neq \frac{2}{7}$ mais $p = \frac{4}{7}$.

Par suite, si l'on rejette (H_0), on accepte (H_1).

Contrôler avec une statistique

Pour que les tests statistiques aient un sens, il convient que ces tests soient fondés sur le calcul des probabilités et sur les théorèmes des statistiques. C'est pourquoi, après que (H_0) et (H_1) ont été formulées, on va devoir contrôler ces deux hypothèses au moyen d'une statistique S à partir de l'échantillon considéré.

Contrôler avec une statistique

Pour que les tests statistiques aient un sens, il convient que ces tests soient fondés sur le calcul des probabilités et sur les théorèmes des statistiques. C'est pourquoi, après que (H_0) et (H_1) ont été formulées, on va devoir contrôler ces deux hypothèses au moyen d'une statistique S à partir de l'échantillon considéré.

Par exemple, pour un test du χ^2 d'ajustement, on considère

$$\chi^2 := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \,.$$

Contrôler avec une statistique

Pour que les tests statistiques aient un sens, il convient que ces tests soient fondés sur le calcul des probabilités et sur les théorèmes des statistiques. C'est pourquoi, après que (H_0) et (H_1) ont été formulées, on va devoir contrôler ces deux hypothèses au moyen d'une statistique S à partir de l'échantillon considéré.

Par exemple, pour un test du χ^2 d'ajustement, on considère

$$\chi^2 := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \,.$$

La plupart du temps, dans les cas concrets, on se ramène à des tests connus et cette étape consiste à identifier le bon test (et donc la bonne statistique à utiliser).

Établir la distribution de la statistique de contrôle

On dispose de deux hypothèses principales : (H_0) et (H_1) ; ces dernières étant éventuellement complémentaires. Selon ces hypothèses, la loi de la statistique va différer.

Établir la distribution de la statistique de contrôle

On dispose de deux hypothèses principales : (H_0) et (H_1) ; ces dernières étant éventuellement complémentaires. Selon ces hypothèses, la loi de la statistique va différer.

Dans cette étape, on doit établir la loi de variabilité de la statistique de contrôle conditionnellement à (H_0). Et, si possible on fait de même conditionnellement à l'hypothèse alternative (H_1) ; bien que dans le cas d'un test sur la moyenne, la loi forte des grands nombres nous permet de nous affranchir de cette possibilité.

Choisir α

Un test d'hypothèse amène nécessairement à des risques d'erreur. L'on aimerait que les deux erreurs qui peuvent survenir (celle de première espèce et celle de seconde espèce) soient aussi petites que possibles.

Choisir α

Un test d'hypothèse amène nécessairement à des risques d'erreur. L'on aimerait que les deux erreurs qui peuvent survenir (celle de première espèce et celle de seconde espèce) soient aussi petites que possibles.

L'erreur qui nous intéresse le plus est la première à savoir la probabilité de rejeter (H_0) conditionnellement au fait que (H_0) soit vrai. Cette erreur, que l'on quantifie par α est celle que l'on va contrôler.

Choisir α

Un test d'hypothèse amène nécessairement à des risques d'erreur. L'on aimerait que les deux erreurs qui peuvent survenir (celle de première espèce et celle de seconde espèce) soient aussi petites que possibles.

L'erreur qui nous intéresse le plus est la première à savoir la probabilité de rejeter (H_0) conditionnellement au fait que (H_0) soit vrai. Cette erreur, que l'on quantifie par α est celle que l'on va contrôler.

On se fixe donc $\alpha > 0$ en fonction de l'étude que l'on fait avec le test statistique en question. Généralement, on prend $\alpha = 0.1\%$ ou 1% ou 2% voire 5%.

Déterminer la région critique associée - 1

Par définition de la région critique associée, on doit trouver le seuil critique $q_c(\alpha)$ tel que

$$\mathbb{P}_{(H_0)}\left(S > q_c(\alpha)\right) = \alpha\,,$$

si l'hypothèse (H_0) est de la forme $p = p_0$ et (H_1) de la forme $p > p_0$.

Déterminer la région critique associée - 2

Si jamais (H_1) est de la forme $p < p_0$, alors $q_c(\alpha)$ satisfait

$$\mathbb{P}_{(H_0)}\left(S < q_c(\alpha)\right) = \alpha\,.$$

Déterminer la région critique associée - 3

Enfin, si (H_1) est de la forme $p \neq p_0$, $q_c(\alpha)$ satisfait

$$\mathbb{P}_{(H_0)}\left(|S| > q_c(\alpha)\right) = \alpha \,.$$

Déterminer la région critique associée - 4

La région critique avec laquelle on rejette l'hypothèse (H_0) sera donc de la forme $[q_c(\alpha); +\infty[$, $]-\infty; q_c(\alpha)]$ ou $]-q_c(\alpha); q_c(\alpha)[^c$.

Calculer la valeur empirique de la statistique

On calcule la valeur de la statistique $S(X_1, \cdots, X_n)$ dans l'échantillon. On obtient une valeur $s(x_1, \cdots, x_n)$ où $x_1, \cdots, x_n$ sont les réalisations dans l'échantillon.

Décider

Si $s(x_1, \cdots, x_n)$ est dans la région critique, on rejette (H_0) donc on accepte l'hypothèse alternative (H_1). Si la valeur empirique n'est pas dans la région critique, on ne rejette pas (H_0).

Décider

Si $s(x_1, \cdots, x_n)$ est dans la région critique, on rejette (H_0) donc on accepte l'hypothèse alternative (H_1). Si la valeur empirique n'est pas dans la région critique, on ne rejette pas (H_0).

Il est essentiel de comprendre que l'on n'accepte pas (H_0). On ne la rejette simplement pas.

Plan

Erreur de première espèce

L'erreur de première espèce consiste à rejeter (H_0) alors que (H_0) est vraie. On contrôle cette erreur par α.

Erreur de première espèce

L'erreur de première espèce consiste à rejeter (H_0) alors que (H_0) est vraie. On contrôle cette erreur par α.

Il est essentiel de bien comprendre que le paramètre α n'est pas un paramètre de précision.

Erreur de première espèce

L'erreur de première espèce consiste à rejeter (H_0) alors que (H_0) est vraie. On contrôle cette erreur par α.

Il est essentiel de bien comprendre que le paramètre α n'est pas un paramètre de précision.

Ainsi, au premier abord, on pourrait penser que plus α est petit, moins on se trompe.

Erreur de première espèce

L'erreur de première espèce consiste à rejeter (H_0) alors que (H_0) est vraie. On contrôle cette erreur par α.

Il est essentiel de bien comprendre que le paramètre α n'est pas un paramètre de précision.

Ainsi, au premier abord, on pourrait penser que plus α est petit, moins on se trompe.

Ceci n'est pas exact.

Erreur de première espèce

L'erreur de première espèce consiste à rejeter (H_0) alors que (H_0) est vraie. On contrôle cette erreur par α.

Il est essentiel de bien comprendre que le paramètre α n'est pas un paramètre de précision.

Ainsi, au premier abord, on pourrait penser que plus α est petit, moins on se trompe.

Ceci n'est pas exact.

Plus α est petit et moins l'on ne rejettera (H_0).

Erreur de seconde espèce

L'erreur de seconde espèce consiste à ne pas rejeter (H_0) alors que (H_1) est vraie. Cette erreur est notée β. On n'a pas de prise sur β. On peut éventuellement calculer cette valeur si l'on dispose de statistiques simples.

Erreur de seconde espèce

L'erreur de seconde espèce consiste à ne pas rejeter (H_0) alors que (H_1) est vraie. Cette erreur est notée β. On n'a pas de prise sur β. On peut éventuellement calculer cette valeur si l'on dispose de statistiques simples.

La quantité $1 - \beta$ est appelée la puissance du test. Ainsi, les tests paramétriques ont une erreur de seconde espèce plus faible que celle des tests non paramétriques.

Différences entre les erreurs

Il est crucial de bien comprendre qu'*a priori*, α et β n'ont rien à voir. En effet,

$$\alpha := \mathbb{P}_{(H_0)}\left(\text{rejeter } (H_0)\right) = \frac{\mathbb{P}\left((H_0),\ \text{rejeter } (H_0)\right)}{\mathbb{P}((H_0))}\,,$$

Différences entre les erreurs

Il est crucial de bien comprendre qu'*a priori*, α et β n'ont rien à voir. En effet,

$$\alpha := \mathbb{P}_{(H_0)}\left(\text{rejeter } (H_0)\right) = \frac{\mathbb{P}\left((H_0),\ \text{rejeter } (H_0)\right)}{\mathbb{P}((H_0))} ,$$

tandis que

$$\beta := \mathbb{P}_{(H_0)^c}\left(\text{ne pas rejeter } (H_0)\right) = \frac{\mathbb{P}\left((H_0)^c,\ \text{ne pas rejeter } (H_0)\right)}{\mathbb{P}((H_0)^c)}$$

dans le cas où $(H_1) = (H_0)^c$.

Plan

Remarque

Il est crucial de comprendre qu'on ne verra dans ce cours qu'une infime partie de la théorie des tests statistiques. Pour plus d'informations, voir le polycopié.

Justification théorique du test - 1

Donnons-nous des variables aléatoires $(X_i)_{i\in[\![1;n]\!]}$ indépendantes et identiquement distribuées suivant la loi discrète suivante :

$$\mathbb{P}(X_i = k) = p_k \text{ pour } k \in [\![1;r]\!]\,,\ \text{avec } \sum_{k=1}^{r} p_k = 1\,.$$

Ainsi, il y a r classes différentes.

Justification théorique du test - 1

Donnons-nous des variables aléatoires $(X_i)_{i\in[\![1;n]\!]}$ indépendantes et identiquement distribuées suivant la loi discrète suivante :

$$\mathbb{P}(X_i = k) = p_k \text{ pour } k \in [\![1;r]\!]\,,\ \text{avec } \sum_{k=1}^{r} p_k = 1\,.$$

Ainsi, il y a r classes différentes.

Maintenant, pour toute classe $k \in [\![1;r]\!]$, on pose :

$$N_k := \sum_{i=1}^{n} \mathbb{1}_{\{X_i=k\}}\,.$$

Justification théorique du test - 2

N_k correspond au nombre d'occurrences de la valeur k dans l'échantillon considéré. C'est donc une variable aléatoire. On sait que $\frac{1}{n}N_k$ converge presque sûrement vers p_k quand n tend vers l'infini d'après la loi forte des grands nombres. Et, d'après le théorème central de la limite, $\frac{\frac{1}{n}N_k - p_k}{\sqrt{\frac{p_k(1-p_k)}{n}}}$ tend en loi vers la loi normale centrée réduite.

Justification théorique du test - 3

Il est ainsi naturel **d'imaginer** que $\frac{(N_k - np_k)^2}{np_k(1-p_k)}$ puisse tendre vers une loi du χ^2 à 1 degré de liberté puis l'on peut **envisager** que la somme pour k allant de 1 à r de cette quantité tend en loi vers une loi du χ^2 quand n tend vers l'infini. Le cas échéant, on pourrait créer un test qui discrimine si oui ou non le r-uplet $\left(p_1^0, \cdots, p_r^0\right)$ correspond à la distribution réelle.

Résultat théorique - 1

Plus rigoureusement, on introduit

$$\chi_n^2 := \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k} = n \sum_{k=1}^{r} \frac{\left(\frac{N_k}{n} - p_k\right)^2}{p_k} \,.$$

Résultat théorique - 2

Théorème de Pearson

Quand n tend vers l'infini, la variable aléatoire χ_n^2 tend en loi vers une loi du Khi-deux à $r-1$ degrés de liberté.

Résultat théorique - 2

Théorème de Pearson

Quand n tend vers l'infini, la variable aléatoire χ_n^2 tend en loi vers une loi du Khi-deux à $r-1$ degrés de liberté.

La preuve du théorème est omise.

Résultat théorique - 2

Théorème de Pearson

Quand n tend vers l'infini, la variable aléatoire χ_n^2 tend en loi vers une loi du Khi-deux à $r-1$ degrés de liberté.

La preuve du théorème est omise.

Remarque

En pratique, on se contente d'avoir $np_k \geq 5$ pour tout $k \in [\![1; r]\!]$. Mais, il est préférable que l'on ait $np_k \geq 10$.

Résultat théorique - 3

Remarque

Lorsque l'effectif attendu np_k d'une classe est plus petit que 5, il est recommandé de regrouper cette classe avec une autre qui lui est adjacente avant de procéder au test du χ^2. Le test d'ajustement porte alors sur la distribution dans les classes obtenues après le regroupement.

Plan

Mécanique générale pour une loi discrète - 1

Soit $p_k^0 \in [0;1]$ pour tout $k \in [\![1;r]\!]$, avec $\sum_{k=1}^{r} p_k^0 = 1$.

Mécanique générale pour une loi discrète - 1

Soit $p_k^0 \in [0;1]$ pour tout $k \in [\![1;r]\!]$, avec $\sum_{k=1}^r p_k^0 = 1$.

Étape 1. On se donne l'hypothèse (H_0) suivante : $p = p^0$ c'est-à-dire que l'on a $p_k = p_k^0$ pour tout $k \in [\![1;r]\!]$.

Mécanique générale pour une loi discrète - 1

Soit $p_k^0 \in [0;1]$ pour tout $k \in [\![1;r]\!]$, avec $\sum_{k=1}^r p_k^0 = 1$.

Étape 1. On se donne l'hypothèse (H_0) suivante : $p = p^0$ c'est-à-dire que l'on a $p_k = p_k^0$ pour tout $k \in [\![1;r]\!]$.

Étape 2. L'hypothèse alternative (H_1) est $p \neq p^0$ c'est-à-dire qu'il existe $k_0 \in [\![1;r]\!]$ tel que $p_k \neq p_{k_0}^0$.

Mécanique générale pour une loi discrète - 2

Étape 3. On introduit la statistique $Z_n := \sum_{k=1}^{r} \frac{(N_k - np_k^0)^2}{np_k^0} = n \sum_{k=1}^{r} \frac{\left(\frac{N_k}{n} - p_k^0\right)^2}{p_k^0}$ où N_k est le nombre d'occurrences de la classe k dans l'échantillon considéré.

Mécanique générale pour une loi discrète - 2

Étape 3. On introduit la statistique $Z_n := \sum_{k=1}^{r} \frac{(N_k - np_k^0)^2}{np_k^0} = n \sum_{k=1}^{r} \frac{\left(\frac{N_k}{n} - p_k^0\right)^2}{p_k^0}$ où N_k est le nombre d'occurrences de la classe k dans l'échantillon considéré.

Étape 4. On sait alors d'après le Théorème de Pearson que sous (H_0), Z_n converge en loi vers une loi du χ^2 à $r-1$ degrés de liberté. Et, par la loi forte des grands nombres, il est assez facile de montrer que Z_n tend presque sûrement vers l'infini sous (H_1).

Mécanique générale pour une loi discrète - 2

Étape 3. On introduit la statistique
$Z_n := \sum_{k=1}^{r} \frac{(N_k - np_k^0)^2}{np_k^0} = n \sum_{k=1}^{r} \frac{\left(\frac{N_k}{n} - p_k^0\right)^2}{p_k^0}$ où N_k est le nombre d'occurrences de la classe k dans l'échantillon considéré.

Étape 4. On sait alors d'après le Théorème de Pearson que sous (H_0), Z_n converge en loi vers une loi du χ^2 à $r-1$ degrés de liberté. Et, par la loi forte des grands nombres, il est assez facile de montrer que Z_n tend presque sûrement vers l'infini sous (H_1).

Étape 5. On se donne ensuite un niveau de signification (erreur de première espèce) α.

Étape 6. La région critique consiste ici à résoudre

$$\mathbb{P}(Z_n > q_c(\alpha)) = \alpha\,.$$

Mécanique générale pour une loi discrète - 3

Étape 6. La région critique consiste ici à résoudre

$$\mathbb{P}(Z_n > q_c(\alpha)) = \alpha\,.$$

On trouve ainsi $q_c(\alpha) := q_{1-\alpha}$, le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $r-1$ degrés de liberté. Ceci nous donne la région critique suivante :

$$\mathcal{J}_\alpha := [q_{1-\alpha}; +\infty[\ .$$

Mécanique générale pour une loi discrète - 3

Étape 6. La région critique consiste ici à résoudre

$$\mathbb{P}(Z_n > q_c(\alpha)) = \alpha\,.$$

On trouve ainsi $q_c(\alpha) := q_{1-\alpha}$, le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $r-1$ degrés de liberté. Ceci nous donne la région critique suivante :

$$\mathcal{J}_\alpha := [q_{1-\alpha}; +\infty[\ .$$

Étape 7. On calcule ensuite la valeur de Z_n dans l'échantillon.

Étape 8. Ainsi, si $Z_n \in \mathcal{J}_\alpha$, on rejette l'hypothèse (H_0) et on accepte (H_1) ; sinon on ne la rejette pas.

Un exemple - 1

Exemple

On simule cent lancers d'un dé à six faces par un programme. On obtient la répartition suivante :

Valeurs isolées x_k	1	2	3	4	5	6
Effectifs N_k	17	22	18	14	13	16

Un exemple - 1

Exemple

On simule cent lancers d'un dé à six faces par un programme. On obtient la répartition suivante :

Valeurs isolées x_k	1	2	3	4	5	6
Effectifs N_k	17	22	18	14	13	16

Peut-on considérer, au risque de 5%, que la simulation est bien celle d'un dé non pipé ?

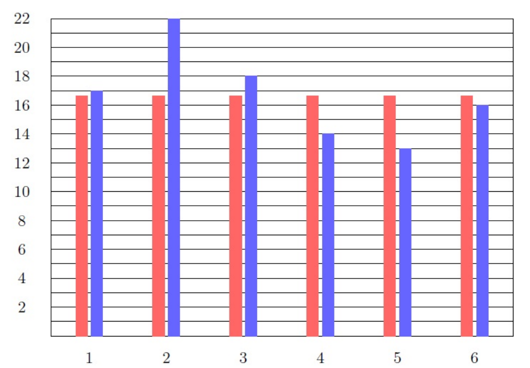
Un exemple - 2

Considérer le dé comme équilibré revient à supposer l'hypothèse (H_0) : "l'échantillon est celui d'une loi équirépartie sur un ensemble à $r = 6$ éléments". On a ici $p_1^0 = p_2^0 = \cdots = p_6^0 = \frac{1}{6}$.

Un exemple - 2

Considérer le dé comme équilibré revient à supposer l'hypothèse (H_0) : "l'échantillon est celui d'une loi équirépartie sur un ensemble à $r = 6$ éléments". On a ici $p_1^0 = p_2^0 = \cdots = p_6^0 = \frac{1}{6}$.

Regardons par ailleurs les deux distributions (effectifs théoriques et effectifs réels) sur un diagramme :



Un exemple - 3

On organise les calculs de la manière suivante :

x_k	1	2	3	4	5	6	Σ
N_k	17	22	18	14	13	16	100
np_k^0	16.67	16.67	16.67	16.67	16.67	16.67	100
$\frac{(N_k - np_k^0)^2}{np_k^0}$	0.0067	1.7067	0.1067	0.4267	0.8067	0.0267	$\chi^2_{obs} = 3.08$

Un exemple - 4

Pour la loi du Khi-deux à $r-1=5$ degrés de liberté, une lecture de la table donne le quantile d'ordre 0.95 égal à $\chi^2_{0.95}(5) \approx 11.1$.

Un exemple - 4

Pour la loi du Khi-deux à $r-1=5$ degrés de liberté, une lecture de la table donne le quantile d'ordre 0.95 égal à $\chi^2_{0.95}(5) \approx 11.1$.

Comme $\chi^2_{obs} < 11.1$, on ne rejette pas (H_0).

Un exemple - 4

Pour la loi du Khi-deux à $r-1=5$ degrés de liberté, une lecture de la table donne le quantile d'ordre 0.95 égal à $\chi^2_{0.95}(5)\approx 11.1$.

Comme $\chi^2_{obs}<11.1$, on ne rejette pas (H_0).

Il s'agit, au risque d'erreur de 5%, d'une simulation d'un dé équilibré.

Plan

Mécanique générale pour une loi continue - 1

Si l'on suppose maintenant que la loi suivie par X est continue et qu'on veut la confronter à une densité f, on se ramène alors à une discrétisation de ladite loi. Pour ce faire, on se donne $r-1$ réels deux à deux distincts $a_1, \cdots, a_{r-1}$. On suppose par ailleurs que l'on a $a_1 < a_2 < \cdots < a_k < a_{k+1} < \cdots < a_{r-1}$.

Mécanique générale pour une loi continue - 1

Si l'on suppose maintenant que la loi suivie par X est continue et qu'on veut la confronter à une densité f, on se ramène alors à une discrétisation de ladite loi. Pour ce faire, on se donne $r-1$ réels deux à deux distincts $a_1, \cdots, a_{r-1}$. On suppose par ailleurs que l'on a $a_1 < a_2 < \cdots < a_k < a_{k+1} < \cdots < a_{r-1}$. On pose ensuite

$$p_k^0 := \int_{a_{k-1}}^{a_k} f(x)\mathrm{d}x\,,$$

pour tout $k \in [\![1; r]\!]$ où $a_0 := -\infty$ et $a_r := +\infty$.

Mécanique générale pour une loi continue - 2

On peut ainsi se ramener au cas discret en posant

$$N_k := \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in]a_{k-1};a_k[\}} .$$

Plan

Pour une loi à paramètre

Nous souhaitons maintenant mettre en place un test permettant de décider si la loi de l'échantillon appartient ou non à une famille de lois $(p^{(\theta)})_{\theta\in\Theta}$ indexée par un paramètre θ à valeurs dans $\Theta \subset \mathbb{R}^d$. On suppose que l'on dispose de $\widehat{\theta_n}$, l'estimateur du maximum de vraisemblance.

Pour une loi à paramètre

Nous souhaitons maintenant mettre en place un test permettant de décider si la loi de l'échantillon appartient ou non à une famille de lois $(p^{(\theta)})_{\theta\in\Theta}$ indexée par un paramètre θ à valeurs dans $\Theta \subset \mathbb{R}^d$. On suppose que l'on dispose de $\widehat{\theta_n}$, l'estimateur du maximum de vraisemblance.

Résultat

Pour chaque paramètre estimé, on perd un degré de liberté. Ainsi, dans le cas de la loi normale, estimer (μ, σ^2) fait perdre deux degrés de liberté.

Plan

Autres tests

On peut aussi faire des tests du χ^2 d'homogénéité ou d'indépendance.

Autres tests

On peut aussi faire des tests du χ^2 d'homogénéité ou d'indépendance.

Également, il existe d'autres types de tests comme ceux de Kolmogorov-Smirnov.