

Rappels

L'objectif d'un test statistique est de déterminer si une hypothèse H_0 doit être rejetée ou non. Un test comporte deux risques :

- le risque de première espèce α qui consiste à déclarer que H_0 est fautive (c'est à dire conclure à une différence) alors qu'en réalité H_0 est vraie
- le risque de deuxième espèce β qui consiste à ne pas déclarer que H_0 est fautive (conclure qu'il n'y a pas de différence) alors qu'en réalité H_0 est fautive.

Dans la pratique on construit un test en trois étapes:

- On Formule l'hypothèse que vous voulez tester H_0 et l'hypothèse alternative H_1 . La formulation unilatérale ou bilatérale de H_1 va déterminer les bornes de la région de rejet.
- On Choisit le type de test à implémenter. Il existe souvent plusieurs tests pour répondre à la même question. Le choix du test est notamment guidé par ses conditions d'application. Exemple : si vous souhaitez tester que deux échantillons indépendants sont issus de la même population, vous pouvez utiliser un test de Student ou un test non paramétrique tel que Mann Whitney. Comment choisir? Vous choisissez un test de Student si vos données sont gaussiennes ou si la taille de vos 2 échantillons est assez grande ($n_1 > 30$ et $n_2 > 30$), afin que le théorème central limite s'applique. Sinon vous utilisez le test de Mann Whitney.
- On interprète les sorties du logiciel : on rejette H_0 , si la p-value est inférieure au risque de 1ère espèce (en général $\alpha = 5\%$).

Application : les cambriolages en France

On souhaite comparer des données statistiques de la délinquance en Ile de France (IDF) avec celles de la région Provence-Alpes-Côte d'Azur (PACA). Les données, publiées par le journal l'Express, fournissent le nombre de vols de voitures (*VolVoiture*) et de cambriolages (*Cambriolage*) dans différentes circonscriptions. Suivant les indicateurs, les délits sont exprimés pour 1000 ou 10000 habitants.

1 Etapes préliminaires

- Charger les données

```
df=pd.read_csv("DataSecurite3.csv",encoding = "ISO-8859-1", engine='python',delimiter=";")
```

- Créer 2 nouveaux tableaux : l'un contenant les données pour les circonscriptions d'Ile de France (IDF, départements : 75, 77, 78, 91, 92, 93, 94, 95) et l'autre pour celles de la région Provence Alpes Côte d'Azur (PACA, départements : 04, 05, 06, 13, 83 , 84)

```
# Exemple de selection: IDF = df.loc[(df['Dpt'] == '75' | (df['Dpt'] == '77'))]
```

1. Faites une analyse descriptive de la variable cambriolage en des IDF et PACA.

2 Comparaison d'une moyenne à une valeur théorique

2.1 On veut savoir si le nombre moyen de cambriolages en IDF est différent de la la moyenne nationale

2.1.1 Formulez les hypothèses H_0 et H_1 .

2.1.2 Quelle est la loi de la statistique de test sous H_0 ?

2.1.3 Effectuez le test et interprétez la p-value au risque 5%.

```
Utiliser la fonction scipy.stats.ttest_1samp( , )
```

2.2 On veut savoir si le nombre moyen de cambriolages en IDF est supérieur à la moyenne nationale

2.2.1 Formulez les hypothèses H_0 et H_1 .

2.2.2 Effectuez le test et interprétez la p-value au risque 5%.

2.2.3 On veut identifiez les différences entre les résultats du test bilatéral (dans 2.1) et le test unilatéral (dans 2.2). Est-ce que la valeur de la statistique et le nombre de degrés de libertés df sont différents?

3 Comparaison de deux échantillons indépendants

On s'intéresse au nombre de cambriolages moyen par circonscription. En particulier on veut savoir si les moyennes dans les régions IDF et PACA sont-elles statistiquement différentes

3.1 Avant de procéder au test, nous nous intéressons d'abord aux variances des cambriolages en IDF et PACA sont égales (leurs égalité).

3.1.1 Pourquoi a-t-on besoin de tester l'égalité des variances?

3.1.2 Rappelez la statistique de ce test, les hypothèses H_0 et H_1 . Ainsi que la loi de la statistique sous H_0 et la zone de rejet .

3.2 On suppose que les variance sont différentes. Maintenant qu'on a une réponse sur l'égalité des variances, on peut faire le test de comparaison du nombre de cambriolage moyen entre les régions IDF et PACA

3.2.1 Formulez les hypothèses H_0 et H_1 permettant de répondre à cette question.

3.2.2 Quel test choisiriez-vous pour la comparaison des moyennes ? Donner la statistique de test et sa loi sous H_0 .

3.2.3 Réalisez le test de comparaison de moyennes et concluez lorsque le risque α est de 5%, 1%.

```
Utiliser la fonction stats.ttest_ind( , , )
```

Sources

Méthode Statistiques Médecine - Biologie -Jean Bouyer, ESTEM éditions Inserm

L'Express - Sécurité 2013 - Classements des communes de France

<https://lexpress.opendatasoft.com/explore/dataset/statistiques-securite-france-2013/?flg=fr>

http://www.lexpress.fr/actualite/societe/insecurite-le-palmares-des-villes-de-france_1300974.html#m2ZPvIj0jcHs7c1g.

99