POLYTECH GRENOBLE

Statistiques

Nathalie GUYADER
nathalie.guyader@gipsa-lab.fr

Année 2021-2022

GRENOBLE INP UGA

POLYTECH GRENOBLE

Chapitres précédents

- Séance 1:
  - Introduction générale
  - Chapitre 1: Probabilités
  - Chapitre 2: Variables aléatoires
  - Chapitre 3: Statistiques descriptives
- Séance 2:
  - Introduction à Python
  - Prise en main d'un Jupyter Notebook
  - TP1: Python et Statistiques descriptives
- Séance 3:
  - TP2: Statistiques descriptives (données sur les véhicules) et compte-rendu sous Jupyter Notebook
- Séance 4
  - Chapitre 4: Estimation par intervalle de confiance Partie 1 : théorie de l'échantillonnage
  - TP3
- Séance 5
  - Chapitre 4: Estimation par intervalle de confiance Partie 2 : estimation
  - Fin du TP3
- Séance 6
  - Chapitre 5: Tests d'hypothèse – Introduction – Comparaison d'une moyenne (d'une proportion) à une norme et comparaison de deux moyennes (deux proportions)

Retour sur les comparaisons de moyennes à une norme et les comparaisons des 2 moyennes

- Comparaison d'une moyenne à une norme
  - Exercice 1:

    On considère dans cet exercice la longueur de pièces usinées. Les réglages sont faits pour assurer une longueur de 27 mm. Un échantillon de 100 pièces est alors extrait pour vérifier cette valeur de 27 mm. Quelle conclusion pouvez-vous faire?
- Vous réaliserez le test « à la main » mais également en utilisant la fonction stat.ttest_1samp

Attention 1: un test de student est ici réalisé; vous pouvez tester un ztest avec la fonction statsmodels.stats.weightstats.ztest

Attention 2: avant d'effectuer le test il est important de vérifier la « normalité » des données..

Retour sur les comparaisons de moyennes à une norme et les comparaisons des 2 moyennes

- Comparaison de deux moyennes
  - Exercice 2:

Ces données correspondent à des erreurs de mesure sur des longueurs de tubes en verre en sortie de 2 chaines de production. Les 2 chaines théoriquement identiques produisent des tubes en verre. Un contrôle de la qualité des tubes est réalisé en sortie de chaque chaine. Une des variables analysées est l'écart entre la longueur attendue et la longueur réelle (en millimètres). L'ensemble des données se trouvent dans le fichier Exercice2.xlsx

Vous réaliserez le test « à la main » mais également en utilisant la fonction stat.ttest_2samp ou avec la fonction ztest

Attention 1: avant d'effectuer le test il est important de vérifier la « normalité » des données et l'homogénéité des variances

POLYTECH GRENOBLE

Chapitre 5 - Tests d'hypothèse (suite)

POLYTECH

Comparaison de 2 moyennes ou plus

L'analyse de variance (ANOVA) permet de comparer **les moyennes de plusieurs échantillons**.
Les conditions préalables sont :
- les échantillons soient indépendants
- les distributions des mesures étudiées (les échantillons) soient issues de distributions parentes normales: *normalité des données*.
- les échantillons sont extraits de distributions parente de même variance (les variances observées sont homogènes): *homogénéité des variances*.

Remarque: ces conditions sont identiques à celles du test de Student dans la comparaison de 2 moyennes (voir tableau de comparaison – doc Comp1et2Moy.pdf)

Si l'une de ces hypothèses n'est pas remplie, l'utilisation de l'ANOVA risquerait d'aboutir à des conditions erronées. L'exemple suivant est repris d'un cours consultable sur internet (http://www.univ-tours.fr/ash/psycho)

Analyse de variance à 1 facteur:

Supposons que l'on étudie les notes obtenues à une épreuve de math par 4 écoles différentes.

Groupes	**A**	**B**	**C**	**D**	
Notes	x_{A1} …	x_{B1} …	x_{C1} … x_i …	x_{D1} …	
Effectif	n_A	n_B	n_C	n_D	$N = n_A + n_B + n_C + n_D$
Total	T_A	T_B	T_C	T_D	$T_G = T_A + T_B + T_C + T_D$
Moyenne	$\bar{x}_A$	$\bar{x}_B$	$\bar{x}_C$	$\bar{x}_D$	$\bar{x}_G = T_G / N$

Les trois conditions sont supposées vérifiées :
- « Les écoles sont indépendantes »
- La variable (note en math) se distribue normalement dans les ensembles parents des 4 classes
- Les ensembles parents des 4 classes ont les mêmes variances

Les hypothèses statistiques:

(H0) : moyennes identiques $\mu_A = \mu_B = \mu_C = \mu_D = \mu_G$

(H_1) : Au moins l'une des moyennes est différentes de μ_G

La solution repose sur la décomposition de la variation de la variable en une variation 'intra' groupe et une variation 'inter' groupe (variation ou somme des carrés).

La variation totale:

$$SC_T = \sum_{i=1}^{N} (x_i - \overline{x_G})^2$$

La variation entre les classes (variation inter-groupe ou factorielle)

$$SC_F = n_A(\overline{x_A} - \overline{x_G})^2 + \cdots + n_D(\overline{x_D} - \overline{x_G})^2$$

La variation à l'intérieur de chaque classe (variation intra-groupe ou résiduelle)

$$SC_r = \sum_{i=1}^{n_A} (x_{Ai} - \overline{x_A})^2 + \cdots + \sum_{i=1}^{n_D} (x_{Di} - \overline{x_D})^2$$

On en déduit les variances (ou carrés moyens):

$$CM_T = \frac{SC_T}{ddl_T} = \frac{SC_T}{N-1}$$

$$CM_F = \frac{SC_F}{ddl_F} = \frac{SC_F}{k-1}$$

$$CM_r = \frac{SC_r}{ddl_r} = \frac{SC_r}{N-k}$$

Remarques:

$SC_T = SC_F + SC_r$

$ddl_T = ddl_F + ddl_r$

La statistique de décision:

Le rapport entre la variance inter-groupe et la variance intra-groupe suit une loi de Fisher Snédécor avec les degrés de liberté (ddl Inter, ddl Intra):

$$F = \frac{\text{variance inter}}{\text{variance intra}} = \frac{CM_F}{CM_r}$$

A partir des résultats que l'on observe sur les échantillons, on calcule une valeur de F et on compare cette valeur à une valeur critique (choisie avec un certain seuil, généralement 5% ou 1%).

- Si le f que l'on calcule est inférieur au f critique alors on est dans la zone de non rejet de H_0, le test est non significatif.

- Par contre si le f calculé est supérieur au f critique alors on est dans la zone de rejet de H_0. On accepte donc H_1 et on en conclut qu'une moyenne diffère des autres avec le risque d'erreur α. Il y a un effet du facteur étudié.

Source de variation	Somme des carrés des écarts	Nombre de ddl	Carrés moyens (variances)	F
Entre les groupes (inter)	$SC_F = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T_G^2}{N}$	k-1	$CM_F = \frac{SC_F}{k-1}$	$\frac{CM_F}{CM_r}$
A l'intérieur des groupes (intra)	$SC_r = \sum_{i=1}^{N} x_i^2 - \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right)$	N-k	$CM_r = \frac{SC_r}{N-k}$	
Total	$SC_T = \sum_{i=1}^{N} x_i^2 - \frac{T_G^2}{N}$	N-1		

Remarque: soit vous pouvez utiliser ces formules soit vous pouvez reprendre celles données 3 transparents avant avec les SC et les CM

Exemple : On reprend l'exemple précédent avec les notes suivantes

Groupes	A	B	C	D
Notes	6	8	7	4
	3	8	4	3
	7	5	8	6
	5	6	6	3
	4	7	5	6
	5	6	9	7
	3	2	2	8

Supposons que nous ayons vérifié les hypothèses de normalité et d'homogénéité des variances pour les 4 groupes.
Tester l'effet du facteur « écoles » sur les notes obtenues en math.

1/ La méthode utilisée sera bien entendu une méthode d'analyse de variance. Posez les hypothèses nulle et alternative à tester.

2/ Faites ensuite tous les calculs nécessaires (vous pouvez bien entendu faire ces calculs avec Python)

Source de variation	Somme des carrés des écarts	Nombre de ddl	Carrés moyens (variances)	F
Entre les groupes (inter)				
A l'intérieur des groupes (intra)				
Total				

3/ Conclusion

Exercice

On récupère ici les données qui se trouvent dans le fichier Exercice3.xls
Chaque vecteur X1, X2 et X3 contient les temps d'exécution en millisecondes de 3 programmes d'analyse d'images. Ces 3 programmes permettent les mêmes analyses mais avec des implémentations différentes. Nous souhaitons ici décider si l'une des implémentations est meilleure car plus rapide (en temps d'exécution).

1. Tracer pour les comparer les histogrammes (et les boxplot) des mesures (temps d'exécution) pour les 3 programmes. Donner également les moyennes des temps d'exécution pour les 3 programmes et les variances.

2. Quel test proposez-vous dans un premier temps pour comparer les moyennes des temps d'exécution des 3 programmes ?

3. Réalisez ce test et donner une conclusion.

4. Effectuer ensuite les tests de comparaison des moyennes des différents programmes 2 à 2.

Test du khi deux

Ce test s'utilise pour évaluer statistiquement le degré d'ajustement entre une distribution d'effectifs observés et une distribution fictive (répartition théorique). Un outil statistique qui permet de vérifier la concordance entre une distribution expérimentale et une distribution théorique est le **Test du Khi-Deux**.

Remarque :

Cette statistique sert non seulement à vérifier la qualité de **l'ajustement entre une distribution théorique et une distribution expérimentale** mais également à tester **l'indépendance de deux variables** dénombrées suivant diverses modalités (tableau de contingence).

Principe général du test:

Ce test permet de juger de la qualité de l'ajustement d'une distribution théorique à une distribution expérimentale. Pour ce faire, il s'agit de prélever un échantillon suffisamment important et de répartir les observations suivant les diverses valeurs possibles de la variable statistique observée si celle-ci est discrète ou selon une répartition en classes si elle est continue. On veut alors vérifier si cette distribution des effectifs expérimentales s'apparente à une distribution théorique particulière.

Remarque :

Dans la pratique, on doit avoir, pour que l'utilisation de la loi du Khi-Deux soit valide, un nombre suffisant d'observations (comme règle pratique, on utilise 50 observations et plus) pour que les effectifs théoriques des différentes classes soient d'au moins 5. Dans le cas où cette condition n'est pas satisfaite, il y a lieu de regrouper deux ou plusieurs classes adjacentes.

Hypothèses statistiques et règle de décision :

Les seules conditions d'application qui sont requises pour effectuer le test sont :

- Echantillon prélevé au hasard de la population (échantillon aléatoire simple).
- Une taille d'échantillon suffisamment importante (fréquences théoriques de chaque classe soit 5 et plus).

Une fois les fréquences déterminées, il faut par la suite décider, à l'aide de la statistique du khi deux, si les écarts entre les fréquences théoriques et celles qui résultent des observations permettent :

- de ne pas rejeter l'hypothèse nulle émise ; si tel est le cas, les écarts entre les effectifs observées et les effectifs théoriques ne sont pas significatifs
- de rejeter l'hypothèse nulle émise ; si tel est le cas, les écarts sont plutôt attribuables au fait que la distribution théorique, suivie effectivement par les observations, est différente de celle que nous avons supposée. Les écarts sont significatifs.

Les hypothèses statistiques peuvent s'énoncer comme suit :

(H_0) : Les observations suivent la loi de distribution théorique spécifiée.
(H_1) : Les observations ne suivent pas la distribution théorique spécifiée.

En acceptant de courir un risque α (seuil de signification) de refuser l'hypothèse H_0 alors qu'elle est vraie, on en déduit la règle de décision suivante :

On rejette (H_0) ssi :

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(E_{o_i} - E_{t_i}\right)^2}{E_{t_i}} > \chi^2_{\alpha;\nu}$$

$$\nu = k - 1 - l$$

Avec k le nombre de classes et l le nombre de paramètres estimés pour calculer les effectifs théoriques

Exercice 1 :

On souhaite vérifier si un dé est bien équilibré. On jette pour cela 120 fois le dé et on enregistre les résultats obtenus. Il sont résumés dans le tableau ci-dessous:

Résultats	1	2	3	4	5	6
Fréquences observées	14	16	28	30	18	14

Peut-on conclure, au seuil de signification α= 0.05, que le dé est bien équilibré ?

Exercice 2 : Ajustement d'une loi binomiale

Dans une entreprise fabriquant des tubes de verre, on effectue un contrôle visuel sur des échantillons de 20 tubes de verre prélevés après chaque heure de production. La répartition du nombre d'échantillons sans tube défectueux, avec 1 tube défectueux, …, 6 tubes défectueux par échantillons de 20, est présenté dans le tableau ci-dessous. On a observé au total 80 échantillons de taille 20 sur une période de 2 semaines.

Nombres de tubes défectueux	**Nombres d'échantillons**
0	13
1	21
2	19
3	12
4	9
5	4
6	2

La variable « nombre de tubes défectueux » correspond aux conditions d'application de la loi binomiale. On aimerait, à l'aide du test du Khi-Deux, au seuil de signification de 5%, tester l'hypothèse selon laquelle les observations se comportent d'après une loi binomiale.

Le test du Khi-Deux est fréquemment employé lors de sondage où les données se présentent sous forme de fréquences absolues et sont compilées selon deux caractères dans un tableau à double entrée.

Exercice : Sondage auprès des employés de Giscom : existe-t-il un lien entre le niveau de satisfaction vis –à-vis de leur travail et la catégorie salariale de l'employé ?
La responsable des ressources humaines de l'entreprise Giscom a effectué un sondage auprès de 200 employés, choisis au hasard à partir du fichier de l'entreprise, pour connaître leur niveau de satisfaction vis-à-vis de leur travail. Ce caractère a été réparti selon trois modalités : élevé, moyen, faible. On a également noté la catégorie salariale à laquelle appartenait l'employé soit : $20 000 et moins ; plus de $20 000 mais moins de $30 000, $30 000 et plus. A partir des résultats de cette enquête, on construit le tableau de contingence suivant :

	Catégorie salariale			
	$20 000 et moins	**Plus de $20 000 moins de $30 000**	**$30 000 et plus**	**Total des lignes**
Elevé	**13**	**19**	**25**	**57**
Moyen	**28**	**29**	**28**	**85**
Faible	**24**	**18**	**16**	**58**
Total des colonnes	**65**	**66**	**69**	**200**

On voudra vérifier si ces caractères sont indépendants ou si au contraire, le fait d'appartenir à une catégorie salariale nous permet de déduire le niveau de satisfaction de l'employé vis-à-vis de son travail