TD SAS n°5 : Statistiques inférentielles, estimations ponctuelles et par intervalle, tests de conformité à un standard

Introduction

Combien y a-t-il de pays sur terre ? Actuellement, un peu moins de 200 états sont reconnus dans le monde, 197 Etats pour être précis selon la liste officielle reconnue par l'Organisation des Nations Unies (ONU). Certains états n'y figurent pas (ex : Taïwan, Kosovo, Transnistrie, Haut-Karabagh, Ossétie du Sud, Abkhazie, République turque de Chypre du Nord, Azad Cachemire, République populaire de Donetsk…).

… Où sommes-nous le plus heureux ?

Chaque année depuis 2011, l'ONU publie dans le "World Happiness Report" les résultats du sondage Gallup World Poll (GWP) qui estime un score de bonheur dans différents pays. Le GWP recueille les données par interrogatoires téléphoniques*. Le score varie entre 0 et 10 points. Nous vous proposons d'étudier aujourd'hui les données issues du rapport de 2017 à partir du fichier "Happiness2017.csv".

** dans les pays dont les réseaux de télécommunication couvrent au moins 80% de la population ; les tailles des échantillons varient entre 500 et 1000 enquêtés par pays.*

Descriptif des variables

- Country : identifiant
- Region : région du pays
- Happiness_Score : somme des mesures du bien-être subjectif
- Happiness_Rank : rang du Hapiness_Score
- Lower_Confidence_Interval : Borne inférieure de l'intervalle de confiance de l'estimation du Happiness score
- Upper_Confidence_Interval : Borne supérieure de l'intervalle de confiance de l'estimation du Happiness score
- Economy_GDP_per_Capita, Family, Health_Life_Expectancy, Freedom, Trust_Government_Corruption, Generosity, Dystopia_Residual : mesures du bien-être subjectif

PARTIE 0 : Importation des données à partir de SAS ON DEMAND et création de nouvelles variables

0.1 Créez une nouvelle bibliothèque « TD » ainsi qu'un nouveau folder "TD5" (onglet Files(Home). Importez les données sur l'estimation du score de bonheur du fichier .csv : "Happiness2017.csv" (PROC IMPORT).

0.2 Combien y a-t-il d'observations et de variables ? Affichez un résumé de la table. (PROC CONTENTS).

0.3 Discrétisation de variables :

0.3.1 Créez une variable catégorielle « happy_cat » à partir de la variable « Happiness_Score ». Arbitrairement nous choisissons de coder les nouvelles modalités de la façon suivante (étape DATA, branchement conditionnel IF ELSE, spécifier le nombre de caractères à afficher grâce à l'instruction LENGTH) :

- [0;4[= unhappy,
- [4;6[= moderate,
- [6;max] = happy.

0.3.2 Faites de même à partir de la variable « generosity ». Codez ses modalités de la façon suivante :

- [0;0.05[= very_selfish,
- [0.05;0.10[= bit_selfish,
- [0.10;0.15[= indifferent,
- [015;0.20[= generous,
- [0.20;max] = very_generous.

PARTIE 1 : Estimations ponctuelles & par intervalle

1.1 Quelle est la modalité de la variable « happy_cat » la plus fréquente ? (PROC FREQ)

1.2 Croisons les variables « happy_cat » et « generosity_cat ». A quelle modalité de "happy_cat" correspond la plus importante proportion de "happy" ? (PROC FREQ)

1.3 Estimer une proportion et son intervalle de confiance :

1.3.1 En ne considérant que les pays possédant une grande liberté (« Freedom » > 0,5) quel est l'intervalle de confiance à 99% de l'estimation de la fréquence « moderate » (variable « happy_cat ») dans notre échantillon ? (PROC FREQ + WHERE)

1.3.2 Quel est la fréquence de "happy" (modalité de la variable "happy_cat") et son IC à 95% ?

1.4 Quelle est la moyenne mondiale de la variable « Happiness_Score », ainsi que sa variance ? (PROC MEANS)

1.5 Quelle est la moyenne du « happiness_score » pour la catégorie « indifferent » de « generosity_cat » en précisant les bornes supérieurs et inférieurs de son intervalle de confiance à 95% ? (PROC MEANS). Par rapport au résultat obtenu à la question précédente, quelle remarque pouvez-vous soulevez ?

1.6 Quelle est la moyenne et l'écart type du score du bonheur pour les pays dont « Economy_GDP_Per_Capita" est inférieure à 0.8 ? (PROC MEANS + std* + clm**)

**standard deviation*

***confidence limits of the mean*

PARTIE 2 : Test centraux et de conformité à un standard

2.1 On ne considère ici que les données des pays plutôt généreux (i.e. de Generosity>0.15) . Sachant qu'en 2015 le score mondial de bonheur était de 5.21, est-ce que ce score a évolué entre 2015 et 2017 ? (PROC TTEST)

Attention : Avant d'effectuer le test de student, vérifier les conditions d'utilisation: l'échantillon est-il assez grand?

2.2 Calculez la variance du score mondial de bonheur de 2017.

2.3 Même question qu'en 1 mais en utilisant la connaissance de la variance du score de bonheur. (PROC IML pour effectuer les calculs, fonction quantile avec option 'NORMALE' pour déterminer le quantile de la loi normale centrée réduite).

2.4 Tracer l'histogramme des distributions de l'indice de générosité des pays. Qu'observez-vous ? (PROC UNIVARIATE avec option histogram)

2.5 Peut-on considérer que la distribution de l'indice de générosité suit une loi normale ? (PROC UNIVARIATE avec option normal)

PARTIE 3 : ANOVA et Proportion

3.1 La moyenne de Dystopia_Residual est-elle significativement différente entre les pays de chaque region au risque 5% ? 10% ? (PROC ANOVA ou PROC GLM)

3.2 Réalisez à nouveau le test en utilisant un test non-paramétrique et la variable Freedom à la place de Dystopia_Residual.. (PROC NPAR1WAY)

3.3 Peut-on affirmer avec une certitude de 95% que les proportions des régions sont toutes identiques ? (PROC FREQ)

PARTIE 4 : Test d'indépendance entre deux variables aléatoires

4.1 Peut-on conclure à un lien entre la variable "Economy_GDP_per_Capita" et la variable "generosity" ? (PROC CORR)

4.2 Idem pour les variables "happy_cat" et "Dystopia_Residual" ? (PROC ANOVA)

4.3 Enfin, faites de même pour les variables "happy_cat" et "generosity_cat" (PROC FREQ avec option CHISQ).