Exercices de préparation des examens

1 Types de variables

2 Méthodes de collecte de données

Exercice 1 (périodique 2013) 20 min (/3)

Identifier trois techniques (modes) d'enquêtes différents: pour chacun, spécifier 2 avantages et 2 inconvénients.

Exercice 2 (final 2014) 15 min (/1 pt)

Décrire deux avantages et deux inconvénients des enquêtes Origine-Destination réalisées dans la grande région de Montréal tous les cinq ans.

Solution

- Quelques avantages des enquêtes Origine-Destination réalisées dans la grande région de Montréal: taille de l'échantillon, approche désagrégée (déplacements individuels des usagers)
- Quelques inconvénients des enquêtes Origine-Destination réalisées dans la grande région de Montréal: période figée dans le temps, information partielle ou incomplète (car rapportée par une personne pour tous les membres du ménage)

Exercice 3 (quiz 2012)

1. À partir de quelle liste est constituée la base de sondage de l'enquête ménage Origine-Destination de la région de Montréal?
2. Donner des exemple de biais attendus lors de la réalisation de l'enquête ménage Origine-Destination de la région de Montréal, et des raisons pour ces biais?
3. À quoi sert le facteur d'expansion ?
   (a) À ce qu'un répondant faisant partie d'un groupe sous-représenté dans l'échantillon compte pour plus qu'un autre répondant faisant partie d'un groupe sur-représenté dans l'échantillon
   (b) À pondérer l'échantillon afin qu'il représente la taille de la population de référence
   (c) À éliminer le taux de non-réponse
   (d) À visualiser plus précisément les données recueillies
   (e) Toutes ces réponses

Solution

1. La liste de téléphone fixe
2. 1) Sous-représentation des jeunes qui ont de moins en moins de téléphones fixes; 2) Sous-représentation des familles occupées, qui n'ont pas le temps de répondre au téléphone.
3. Réponse b

Exercice 4 (méthodes d'enquêtes) 25 min (/4 pts)

Le conseil intermunicipal de transport (CIT) Laurentides est la société de transport qui dessert un territoire de banlieue au nord de Montréal. Vous désirez recueillir des informations concernant le profil des usagers de la ligne d'autobus 23, ainsi que leur profil d'utilisation de cette ligne d'autobus. Le trajet de la ligne d'autobus relie la gare de train Sainte-Thérèse et la municipalité de Sainte-Anne-des-Plaines, en passant par le Collège Lionel-Groulx (cégep).

1. Quelle est la population de référence pour cette enquête ?
2. Quelle technique de collecte de données suggérez-vous ? Expliquer le fonctionnement de cette collecte.
3. Quel sera le format du questionnaire ?
4. Identifier un cadre temporel approprié pour cette enquête ainsi que l'unité temporelle. Justifier.
5. Quelle est la taille minimale de l'échantillon que vous devez recueillir ? Vous désirez un niveau de confiance de 95 % et vous acceptez une marge d'erreur de 4 %. Selon les données de transactions de carte à puce, vous savez que votre population de référence est composée de 2000 individus, et vous voulez vous assurer de respecter la proportion de 75 % d'étudiants dans la clientèle de la ligne.

Solution

1. Les usagers qui utilisent la ligne d'autobus 23
2. Interception, enquête à bord des autobus de la ligne 23
3. Papier ou iPad
4. Un jour de semaine en automne (durant le calendrier scolaire et de travail)
5. La taille minimale de l'échantillon est $n = (1.96^2 \times 0.75(1 - 0.75))/0.04^2 = 450$ (correction pop limitée $n' = 450/(1 + 450/2000) = 367$ individus)

3 Traitement de données

Exercice 1 (quiz) (/1 Pt)

On note que chaque déplacement d_i est constitué de n_i points $(x_{i,j}, y_{i,j})$, où $1 \leq j \leq n_i$, dans un système de coordonnées cartésiennes. Écrire un algorithme qui calcule la longueur d'un déplacement. Utilisez le format de pseudo-code montré en cours en spécifiant bien les entrées et les sorties.

4 Bases de données

Exercice 1 (périodique 2012, 2019) 45 min (/6 pts)

Nous nous intéressons à créer un système d'information pour une compagnie aérienne. Pour cela, il est nécessaire de modéliser les différents objets et concepts nécessaires à son fonctionnement. Ces entités sont les suivantes: aéroport, vol, avion, employé, type d'employé, garage, passager, billet.

1. Proposer un modèle de données sous forme d'un diagramme Entité/Association impliquant toutes les entités listées ci-dessus. Ajouter des attributs (en indiquant les identifiants) et les associations entre entités avec leurs cardinalités, minimale et maximale, et les fonctionnalités. (4 pts)

2. Traduire le schéma Entité/Association en schéma relationnel. Indiquer clairement les clefs primaires et externes (et à quoi les clefs externes font référence), et proposer des types pour les attributs. (2 pts)

Solution

1. Les entités et leurs attributs sont les suivants (l'identifiant de chaque entité est en **gras**):

   aéroport id, nom, ville, nombre de pistes

   vol id, origine, destination, date et heure de départ, date et heure d'arrivée

   avion id, marque, modèle, date de fabrication

   employé id, nom, date de naissance

   type d'employé id, description

   garage id, adresse, taille

   passager id, nom, genre, date de naissance

   billet id, prix

   Les associations sont les suivantes (il est souhaitable de nommer les associations et de faire un schéma):

   - avion-vol: un vol implique 1-1 avion, un avion est impliqué dans 1-n vols. La fonctionnalité est 1-n.
   - vol-aéroport: un vol implique 2-2 aéroports (arrivée, départ), un aéroport est le point de départ ou d'arrivé de 1-n vols. La fonctionnalité est 2-n.
   - vol-billet: un billet permet de voler sur 1-1 vol, 1-n billets sont vendus pour un vol. La fonctionnalité est 1-n.
   - billet-passager: un billet permet à 1-1 passager de prendre le vol, un passager a pu acheter 0-n billets (dans sa vie). La fonctionnalité est 1-n.
   - vol-employé: un vol implique 1-n employés (sur le vol), un employé travaille/participe à 0-m vols (0 si reste au sol). La fonctionnalité est n-m.
   - employé-type d'employé: un employé a 1-n type d'emplois (historiquement), un type d'emploi peut correspondre à 1-m employés. La fonctionnalité est n-m.

- garage-aéroport: un garage est situé à proximité de 1-1 aéroport, un aéroport peut avoir 0-n garages. La fonctionnalité est 1-n.

On pouvait ajouter des associations entre garage et avion (le garage typiquement utilisé par un avion) et entre employé et garage (pour des employés travaillant à la maintenance dans un garage) ou employé et aéroport.

2. Chaque entité devient une table (Aéroports, Vols, Avions, Employés, Types d'employé, Garages, Passagers, Billets). Il faut ajouter deux tables pour représenter les associations n-m, VolEmployés et EmployéTypeEmployés, qui sont constituées de deux clefs externes vers les clefs primaires des tables concernées par les associations (respectivement Vols et Employés, et Employés et Types d'employés). Il faut ajouter des clefs externes suivantes pour les associations n-m:
   - aéroportDépartId, aéroportArrivéeId et avionId dans Vols faisant références respectivement à Aéroports.id (2) et Avions.id;
   - volId et passagerId dans Billets faisant références à Vols.id et Passagers.id;
   - aéroportId dans Garages faisant référence à Aéroports.id.

Exercice 2 (modèles de données) 45 min (/6 pts)

Nous nous intéressons à créer un système d'information pour une compagnie maritime. Pour cela, il est nécessaire de modéliser les différents objets et concepts nécessaires à son fonctionnement. Ces entités sont les suivantes: port, trajet, bateau, employé, type d'employé, port d'attache (port unique de référence pour un bateau), conteneur ("container").

1. Proposer un modèle de données sous forme d'un diagramme Entité/Association impliquant toutes les entités listées ci-dessus. Ajouter des attributs (en indiquant les identifiants) et les associations entre entités avec leurs cardinalités, minimale et maximale, et les fonctionnalités. (4 pts)

2. Traduire le schéma Entité/Association en schéma relationnel. Indiquer clairement les clefs primaires et externes (et à quoi les clefs externes font référence), et proposer des types pour les attributs. (2 pts)

Solution
Il n'y a pas d'associations n-m dans ce diagramme E/A. Une solution a été présentée en cours.

Exercice 3 (final 2014) 30 min (/3.5 pts)

Une trajectoire est un ensemble de mesure de positions (x, y) à des instants t.

1. Proposer un modèle de données relationnel permettant de stocker des trajectoires et de faire des requêtes sur leurs positions. Indiquer clairement la clef primaire et les types des attributs. (1 pt)

2. On désire effectuer un projet pilote des déplacements d'un petit groupe de conducteurs par récepteur GPS. Modifier la base de données précédente pour enregistrer des caractéristiques de chaque conducteur participant au projet et ses trajectoires GPS (par exemple nom, âge, sexe, lieu de domicile, etc.). Indiquer toujours la clef primaire et les types des attributs. (1 pt)

3. Écrire une requête SQL pour l'extraction des déplacements de l'usager Paul, en ordonnant les positions de ses trajectoires temporellement. (1 pt)

4. Quelle caractéristique des systèmes de gestion de base de données permet de protéger les informations individuelles des usagers participant au projet? (0.5 pt)

Exercice 4 (quiz) (/2 pts)

Nous voulons concevoir un modèle de données pour l'enquête sur les déplacements effectuée par la Ville de Montréal à l'aide de l'application mobile MTL trajet. Après avoir installé l'application sur leur téléphone, les participants répondent à un premier questionnaire sur leurs caractéristiques socio-démographiques et leurs habitudes de transport. L'application enregistre ensuite leurs déplacements pendant 30 jours. Pour chaque déplacement, lorsqu'elle détecte la fin du déplacement, l'application demande au répondant des informations complémentaires comme le mode et le motif du déplacement.

1. Proposer un modèle pour les données collectées avec l'application mobile MTL trajet sous forme d'un diagramme Entité/Association impliquant au minimum les entités suivantes: répondant, déplacement, point GPS. Ajouter des attributs (incluant l'identifiant) et les associations entre entités, avec leurs cardinalités minimale et maximale, et les fonctionnalités.

2. Traduire le schéma Entité/Association en schéma relationnel. Indiquer clairement les clefs primaires et externes.

Exercice 5 (final 2012) 30 min (/4)

Présenter un modèle de données pour un système de transport. Vous avez deux choix: une agence de location de voitures et camions ou un service de covoiturage dont la gestion est centralisée. Votre modèle doit comprendre au moins cinq entités et constituer un ensemble cohérent (qui ne manque pas un élément important nécessaire à la fonctionnalité principale du système). Présentez le modèle relationnel pour un tel système. Indiquez clairement:

1. les clefs primaires;

2. les clés étrangères;

3. des attributs pertinents;

4. les types de données de ces attributs;

5. la fonctionnalité de chaque relation;

6. vous devez avoir au moins une relation de type *plusieurs-à-plusieurs* (*n-m*).

Exercice 6 (périodique 2010) 45 min (/6 pts)

La ville de Montréal aimerait créer un système d'information pour gérer les places de stationnement public dans la ville. En particulier, ce système devrait permettre un inventaire exhaustif des emplacements, payants ou non, et de leur utilisation. On suppose que les emplacements payants sont équipés d'un appareil de collecte de l'argent de stationnement (parcomètre) et que le coût de stationnement par heure est fixe dans

le temps. Lorsqu'un conducteur se gare sur un de ces emplacements, il doit payer au parcomètre pour une certaine durée. Il faut aussi noter que certains stationnements ont des restrictions particulières, par exemples certains sont réservés aux personnes handicapées.

1. Proposer un modèle de données sous forme d'un diagramme Entité/Association qui permette d'enregistrer toutes les informations sur les emplacements de stationnement public et leur utilisation tels que décrits ci-dessus. Ajouter des attributs (en indiquant les identifiants) et les associations entre entités avec leurs cardinalités, minimale et maximale, et les fonctionnalités. (4 pts)

2. Traduire le schéma Entité/Association en schéma relationnel. Indiquer clairement les clefs primaires et externes (et à quoi les clefs externes font référence), et proposer des types pour les attributs. (2 pts)

Exercice 7 (final 2014) 50 min (/8 pts)

Cet exercice repose sur un ensemble de données de circulation enregistrées par des boucles magnétiques dans la région de Portland, importées dans le fichier SQLite `14-freeway_loopdata1hr.sqlite`. Ce fichier contient les données de comptage et de vitesse agrégées sur des périodes d'une heure pour plusieurs stations de comptage. Les colonnes utiles de la table "loopdata" sont les suivantes:

- "detectorid": identifiant de la station de comptage;
- "starttime": début (jour, heure et fuseau horaire) de l'intervalle d'une heure sur lequel les données de circulation sont agrégées;
- "volume": débit horaire (nombre de véhicule par heure);
- "speed": vitesse moyenne sur l'heure (en mile par heure);
- "occupancy": taux d'occupation (pourcentage du temps pendant lequel le capteur est occupé par un véhicule);
- "date": date correspondant à "starttime";
- "time": heure correspondant à "starttime";
- "daytype": jour de la semaine (nombre entier: 0 correspond à dimanche, 1 à lundi, ... et 6 à samedi).

Veuillez répondre aux questions suivantes:

1. Quelle est la clef primaire de la table "loopdata"? La table "loopdata" suit-elle les trois formes normales ? Justifier votre réponse. (1 pt)

2. Pour la station de comptage 1732, écrire la requête SQL permettant de calculer la vitesse moyenne et le nombre de mesures de vitesse par jour de la semaine (lundi, mardi, etc.). Effectuer le test statistique approprié pour déterminer si le jour de la semaine à un impact sur la vitesse moyenne à cette station. (3 pts)

3. Écrire la requête permettant de calculer le débit horaire moyen par heure pour chaque heure de la journée sur l'ensemble des jours de semaine (lundi au vendredi inclus) pour chaque station. (0.5 pt)

4. Soit quatre périodes horaires de la journée (nuit: minuit à 6h; matin: 6h à midi; après midi: midi à 18h; soirée: de 18h à minuit).

   (a) Donner une des requêtes pour créer une des quatre nouvelles tables (ou vues) calculant pour chaque station le débit moyen par période (une table/vue par période) pour les jours de semaine. (0.5 pt)

   (b) Écrire la requête pour joindre les quatre tables/vues pour obtenir le débit moyen par période de la journée par station. (0.5 pt)
   Le résultat ressemble à quelque chose comme:

Station	Débit nuit	Débit matin	Débit après midi	Débit soirée
1345	123	456	789	123
1346	456	789	123	123
...				

   (c) Chaque station est maintenant caractérisée par quatre débits moyens par période de la journée: appliquer l'algorithme des k-moyennes pour identifier des groupes homogènes de stations ayant des débits similaires selon la période de la journée. Présenter les résultats en quelques lignes. Représenter les centroïdes des groupes sur une figure. (2.5 pts)

Solution

1. La clef primaire de la table "loopdata" est la clef composite (detectorid, starttime). La table suit la première forme normale, mais pas la seconde car les attributs date, time et daytype se rapportent seulement à une partie de la clef primaire (starttime).

2. `SELECT daytype, AVG(speed), COUNT(speed) FROM loopdata WHERE detectorid=1732 GROUP BY daytype`

daytype	AVG(speed)	COUNT(speed)
0	48.8148333333333	120
1	50.1735338345865	133
2	49.0002097902098	143
3	49.8603539823009	113
4	50.9990517241379	116
5	50.5410743801653	121
6	50.1680172413793	116

   Le test statistique approprié pour déterminer si le jour de la semaine à un impact sur la vitesse moyenne à cette station est l'analyse de variance à un facteur (ANOVA). On peut utiliser Excel ou Tanagra pour faire ce test. Il faut exporter les données avec la requête suivante:
   `SELECT daytype, speed FROM loopdata WHERE detectorid=1732 ORDER BY daytype;.`
   L'hypothèse nulle est que la moyenne des vitesses est identique pour tous les groupes (hypothèse alternative: au moins une moyenne est différente). La statistique du test est $F = 3.13$, qui correspond à un risque de première espèce de 0.0048, ce qui est très faible. On peut rejeter l'hypothèse nulle et conclure que la vitesse moyenne change selon le jour de la semaine.

3. `SELECT detectorid, time, AVG(volume) FROM loopdata WHERE daytype BETWEEN 1 AND 5 GROUP BY detectorid, time ORDER BY detectorid, time;`

4.

(a) Voici l'exemple de la requête pour créer la première vue:
```
CREATE VIEW qsoiree AS SELECT detectorid, AVG(volume) AS volume
FROM loopdata WHERE (daytype BETWEEN 1 AND 5) AND (time BETWEEN
"18:00:00" and "23:00:00")
GROUP BY detectorid;
```

(b) Soient qnuit, qmatin, qapresmidi et qsoiree les quatre vues. La requête pour joindre les vues estl a suivante:
```
SELECT qnuit.detectorid, qnuit.volume
as debit_nuit,
qmatin.volume as debit_matin, qapresmidi.volume as debit_apresmidi,
qsoiree.volume as debit_soiree
FROM qnuit, qmatin, qapresmidi, qsoiree
WHERE qnuit.detectorid = qmatin.detectorid and qmatin.detectorid
= qapresmidi.detectorid
and qapresmidi.detectorid = qsoiree.detectorid
```

(c) En choisissant trois groupes, les stations se répartissent en débits élevés, moyens et faibles pour les quatres périodes considérées. Le troisième groupe de 21 stations a des débits moyens systématiquements plus élevés, tandis que le second groupe de 18 stations a des débits moyens systématiquement plus faibles. Le premier groupe contient le plus de stations (30) et présente des débits moyens entre les débits des deux autres groupes pour toutes les périodes, légèrement plus élevées que la moyenne globale pour le matin et l'après-midi et plus faibles pour la nuit et la soirée. Les centroïdes sont les suivants et peuvent êtres représentés sur un graphique à coordonnées parallèles (les moyennes des débits de chaque groupe en fonction de la période sur l'axe des abscisses).

	Groupe 1	Groupe 2	Groupe 3
debit_nuit	293.059309	146.422476	443.714401
debit_matin	1008.166874	470.038251	1327.814308
debit_apresmidi	1096.418644	534.436908	1338.707613
debit_soiree	598.190248	327.042045	874.911848

5 Données spatiales

Exercice 1 (quiz 2012) Quel système est le plus précis: MTM ou UTM? Pourquoi?

Solution MTM: chaque zone correspond à un angle de 3°. Aux bords de chaque zone, c'est plus précis que les zones UTM correspondant à un angle de 6°.

Exercice 2 (périodique 2014)

1. Quelle est la différence entre un géoïde, un datum et un ellipsoïde? (1 pt)
2. Projection de Mercator
   (a) Quel est le problème principal de la projection Mercator (celle qui est utilisée couramment pour représenter la terre sur une carte). (0.5 pt)
   (b) À quel(s) endroit(s) sur le globe est-ce le plus problématique? (0.5 pt)

Solution

1. Un datum est un système de référence permettant d'exprimer les positions au voisinage de la Terre, impliquant un modèle de la forme de la terre, généralement des coordonnées en unités d'angle (par ex. degrés), et une origine $(0, 0)$.
   Le modèle de la terre est généralement un ellipsoïde de révolution conventionnel (choisi de manière à approcher le géoïde) dont les paramètres de définition sont généralement son centre (choisi à proximité du centre de gravité terrestre) et trois axes orthonormés définis par leur orientation.

2. Projection de Mercator

   (a) La terre est projetée sur un cylindre, coïncidant avec la terre à l'équateur, et dont l'erreur sur les distances augmente avec la distance à l'équateur.

   (b) Près des pôles.

Exercice 3 (final 2017) 45 min (/6 pts)

Vous disposez des tables suivantes.

- Table arrondissements:

Champ	Type
id_arrond	Integer
nom_arrond	VARCHAR(255)
Geom	Geometry(MultiPolygon,32188)

- Table reseau_routier:

Champ	Type
id_lien_routier	Integer
Geom	Geometry(MultiLinestring,32188)

- Table reseau_cyclable:

Champ	Type
id_lien_cyclable	Integer
Geom	Geometry(MultiLinestring,32188)

Proposer une méthode, par exemple sous forme de requête SQL avec des fonctions spatiales, afin de déterminer, par arrondissement, la proportion du réseau routier qui contient une piste cyclable. La liste de fonctions spatiales est présentée dans le tableau 1.

Solution
Les étapes de la méthode sont les suivantes:

1. Création d'une table des liens routiers par arrondissement:
```
CREATE TABLE public.reseau_routier_arrond AS SELECT l.*, r.id_arrond,
ST_Intersection(l.geom,r.geom) as geom_intersection FROM public.reseau_r
l INNER JOIN public.arrondissements r ON ST_Intersects(l.geom,r.geom);
```

2. Création d'une table des liens cyclables par arrondissement:
```
CREATE TABLE public.reseau_cyclable_arrond AS SELECT l.*, r.id_arrond,
ST_Intersection(l.geom,r.geom) as geom_intersection FROM public.reseau_c
l INNER JOIN public.arrondissements r ON ST_Intersects(l.geom,r.geom);
```

3. Extraction des pistes cyclables qui sont à une certaine distance du réseau routier (ici 10m):
```
CREATE TABLE public.reseau_cyclable_ arrond_within10m AS SELECT
```

Fonction	Description
ST_Area (g1)	Returns the area of the surface if it is a Polygon or MultiPolygon
ST_Dwithin (g1,g2,distance_of_srid)	Returns true if the geometries are within the specified distance of one another
ST_Intersection (geomA,geomB)	Returns a geometry that represents the shared portion of geomA and geomB
ST_Intersects (geomA,geomB)	Returns TRUE if the Geometries/Geography "spatially intersect in 2D"
ST_Length (g1)	Returns the 2d length of the geometry if it is a linestring or multilinestring
ST_X (g1)	Return the X coordinate of the point
ST_Y (g1)	Return the Y coordinate of the point

Tableau 1: Liste de fonctions spatiales

```
DISTINCT ON (a.id_lien_cyclable) a.* FROM public.reseau_cyclable_
arrond a, public.reseau_routier_arrond b WHERE ST_DWithin(a.geom_interse
b.geom_intersection, 10);
```

4. Calcul des longueurs de réseau cyclable sélectionné et de réseau routier par arrondissement:

```
CREATE TABLE arrondissements_longueurs AS SELECT l.id_arrond,
sum(ST_Length(c.geom_intersection))/sum(ST_Length(r.geom_intersection))
as pourcentage_reseau_cyclable FROM arrondissements l LEFT JOIN
reseau_cyclable_arrond_within10m c ON l.id_arrond = c.  id_arrond
LEFT JOIN reseau_routier_arrond r ON l.id_arrond = r.  id_arrond
GROUP BY l.nom_arron;
```

6 Analyse statistique

Exercice 1 (quiz 2012) Sachant que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ (où $\bar{X}$ est la moyenne empirique de n échantillons de X, μ et σ la moyenne et l'écart-type de X) tend vers la loi normale centrée réduite, calculer un intervalle de confiance de la moyenne de 100 échantillons de vitesses de moyenne empirique 55 km/h et d'écart-type 8 km/h. Spécifier le niveau de confiance de l'intervalle (si Z est une variable aléatoire réelle de loi normale centrée réduite, $P(Z \leq 1.96) = 0.975$ et $P(Z \leq 1.645) = 0.95$).

Solution L'intervalle de confiance à 95 % est $[53.43, 56.57]$ ($\mu \pm 1.96 * \sigma/\sqrt{n} = 55 \pm 1.96 * 8/10$).

Exercice 2 (quiz 2012) On teste l'hypothèse H_0 si un échantillon de vitesse suit la loi normale: la variable de décision du test du χ^2 est calculée et vaut 14.3574. Le nombre de degrés de libertés est 9 et les valeurs seuils pour une distribution du χ^2_9 sont 14.68 et 16.92 pour des risques de première espèce respectifs de 10 % et 5 %. Conclure.

Solution Le risque est trop grand (supérieur à 10 %) de rejeter l'hypothèse nulle, donc on ne peut pas rejeter l'hypothèse nulle que les vitesses suivent la loi normale. Donc il semble que les vitesses suivent la loi normale.

Exercice 3 (final 2010)

Après un élargissement des voies sur la route étudiée à une autre question (section fouille de données, fichier `vitesse-debit.csv` (dans `10-vitesse-debit.zip`)), on effectue un nouveau relevé de données contenues dans le fichier `vitesse-debit2.csv` (dans `10-vitesse-debit.zip`). On aimerait savoir si cet aménagement a eu un impact sur les vitesses pratiquées par les conducteurs sur cette route. Après transformation des données en nombre d'observations par intervalles de vitesse (par exemple à l'aide de la fonction histogramme d'Excel), indiquer si l'aménagement à eu un impact significatif sur la distribution des vitesses (avec un niveau de confiance de 95 %).

Solution

Il faut tester l'hypothèse H_0: la distribution des vitesses est identique avant et après l'aménagement. Pour cela, il faut transformer les données en nombre d'observation par intervalle (l'outil histogramme d'Excel en constitue 15), en regroupant les intervalles avec moins de 5 observations. En prenant les données recueillies avant comme référence, on calcule la variable de décision du test du χ^2: la valeur obtenue est 63,18, supérieur à la valeur de 14.07 correspondant à un niveau de confiance de 95 % pour une variable aléatoire suivant une loi du χ^2_7 à 8-1=7 degrés de liberté. Nous pouvons donc conclure que l'aménagement a eu un impact significatif sur la distribution des vitesses sur cette route.

Exercice 4 (périodique 2013) 50 min (/7 pts)

Le nombre d'accidents sur une route est comptabilisé dans le tableau suivant pendant 15 jours (période 1):

Jour	Nombre d'accidents
1	1
2	1
3	1
4	0
5	2
6	0
7	0
8	0
9	2
10	0
11	1
12	1
13	1
14	1
15	1

1. Écrire l'algorithme du calcul de la médiane d'un ensemble de n nombres réels x_i. (1 pt)

2. Calculer la moyenne et la médiane du nombre d'accidents par jour. (1 pt)
3. Calculer un intervalle de confiance à 95 % pour la moyenne du nombre d'accidents par jour. (1 pt)
4. Calculer le nombre de jours d'observation nécessaires afin d'obtenir la moyenne du nombre d'accidents par jour avec une précision (tolérance) de 0.15 accidents par jour pour un niveau de confiance de 90 et 95 %. (1 pt)
5. Tracer l'histogramme de la loi de distribution du nombre d'accidents par jour (et non la série temporelle du nombre d'accidents en fonction du jour). (1 pt)
6. Pour améliorer la sécurité routière, un policier est placé de façon visible sur le coté de la route pendant 15 jours. Pendant cette période 2, le nombre moyen d'accidents par jour est 0.45 et l'écart-type empirique n'a pas changé (on suppose que les variances sont les mêmes pour les périodes 1 et 2 et que le nombre d'accidents par jour suit une loi normale). Déterminer si le nombre d'accidents a baissé avec un risque d'erreur de première espèce de 5 %. (2 pts)

Solution

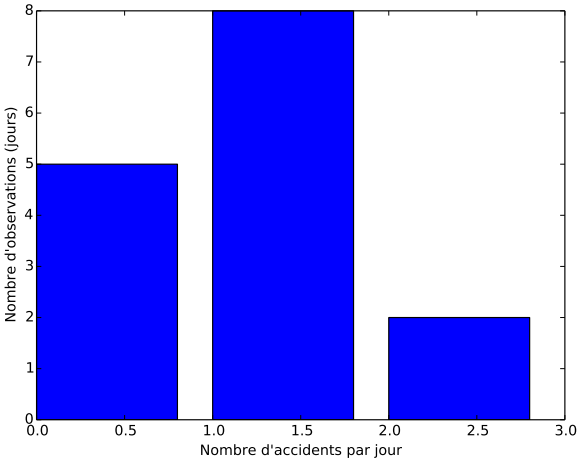
1. Voici un algorithme (supposant une fonction de tri tri existante sur les nombres réels, vue en cours):

   entrée: n nombres réels x_i
   sortie: la médiane des n nombres réels x_i

   début
   $liste\_triee = tri(x_i)$
   si n pair
   renvoyer l'élément en position $n/2$ de $liste\_triee$
   sinon
   renvoyer l'élément en position $(n-1)/2$ de $liste\_triee$
   fin

2. La moyenne et la médiane du nombre d'accidents par jour sont respectivement 0.80 et 1 accidents par jour.
3. L'expression $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ suit la loi de Student à 14 degré de liberté, et la probabilité qu'une telle variable soit respectivement dans l'intervalle $[-2.145, +2.145]$ et $[-1.761, +1.761]$ est 95 % et 90 %. L'écart-type corrigé s est 0.68 et l'intervalle de confiance de la moyenne du nombre d'accidents par jour est donc respectivement $0.8 \pm 2.14\frac{0.68}{\sqrt{15}} = [0.42, 1.18]$ et $[0.49, 1.11]$ pour des niveaux de confiance de 95 et 90 %.
4. On suppose que l'écart-type empirique est proche du vrai écart-type. Le nombre d'observation nécessaire est respectivement $n = 1.64^2\frac{0.68^2}{0.15^2} = 55$ et $n = 1.96^2\frac{0.68^2}{0.15^2} = 79$ pour des niveaux de confiance de de 90 et 95 %.

5. L'histogramme de la loi de distribution du nombre d'accidents par jour ci-dessous est obtenu par le code Python à la fin de l'exercice:



6. On teste l'hypothèse H_0: le nombre moyen d'accidents n'a pas changé contre H_1: le nombre d'accidents à baissé. La statistique du test est $\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = (0.8 - 0.45)/(0.68\sqrt{2/15}) = 1.41$ ($n_1 = n_2 = 15$, $s_1 = s_2 = 0.68$). La statistique suit la loi de Student à $n = 15+15-2$ degrés de liberté. La valeur seuil de la distribution pour un risque de première espèce de 0.05 est 1.701 (soit la probabilité qu'une variable suivant la loi de Student à 28 degrés de liberté soit supérieure ou égale à 1.701 est 0.05). On ne peut donc rejeter H_0, le nombre d'accidents ne semble pas avoir été affecté. On peut trouver dans la table que la valeur p (ou risque de première espèce) pour 1.41 est entre 0.10 et 0.05, ce qui pourrait être accepté avec un niveau de confiance de 90 %.

Exercice 5 (final 2013) 55 min (/8 pts)

Cet exercice repose sur un ensemble de 3000 accidents impliquant un piéton et un véhicule entre 2003 et 2006 dans la ville de Montréal (le fichier `13-accidents-pietons-montreal.txt` est disponible sur moodle). Les données se présentent sous la forme d'un fichier texte (avec séparation des champs par une tabulation), et chaque accident est décrit par les attributs décrits dans le tableau 2.

1. Discuter les modèles statistiques pouvant être utilisés pour étudier l'association de ces attributs avec la gravité des accidents. (1 pt)

2. Décrire les traitements nécessaires pour utiliser des données nominales dans une analyse de régression (par exemple une régression linéaire). (1 pt)

3. En créant une nouvelle variable binaire représentant les accidents mortels et avec blessures graves (la variable vaut 1 si l'accident est mortel ou avec blessures graves, 0 sinon), choisir un modèle statistique pour étudier les facteurs contribuant à la probabilité d'un accident mortel ou grave: estimer le modèle (avec Tanagra), présenter clairement les attributs significatifs et discuter les résultats. (4 pts)

Attribut	Description
EVENT	numéro d'accident
RDCLASS	classification de la route (1: autoroute; 2: route à numéro; 3: collectrice; 4: artère; 5: locale)
SPD_KM	limite de vitesse selon la classification de la route
MED_INC	revenu médian dans la zone de l'accident
pop_dens_200	densité de population dans un rayon de 200 m
veh_type	type de véhicule ("car": voiture; moto; "VTB": van, camion ("truck") ou bus; "EMS": véhicule d'urgence))
BAD_WEAT	variable indicatrice de mauvais temps
SEVERITY	gravité de l'accident (3: mortel; 2: blessure grave; 1: blessure légère; 0: sans blessure)
DARK	variable indicatrice de la nuit
Park_10	présence d'un parc à 10 m de l'accident
hosp_50	présence d'un hôpital dans un rayon de 50 m
veh_mvt	mouvement du véhicule impliqué ("straight"; "backup"; "leftturn"; "righturn")
Comm_Per	pourcentage d'activité commerciale
Res_Per	pourcentage d'activité résidentielle
Inter_Acc	occurrence de l'accident dans un carrefour

Tableau 2: Attributs des accidents

4. Tracer un histogramme des distributions selon la classe de la route du nombre d'accidents mortels et graves d'une part, et du nombre d'accidents avec blessures légères et sans blessure d'autre part: appliquer un test statistique pour déterminer si les deux distributions sont différentes. (2 pts)

Solution

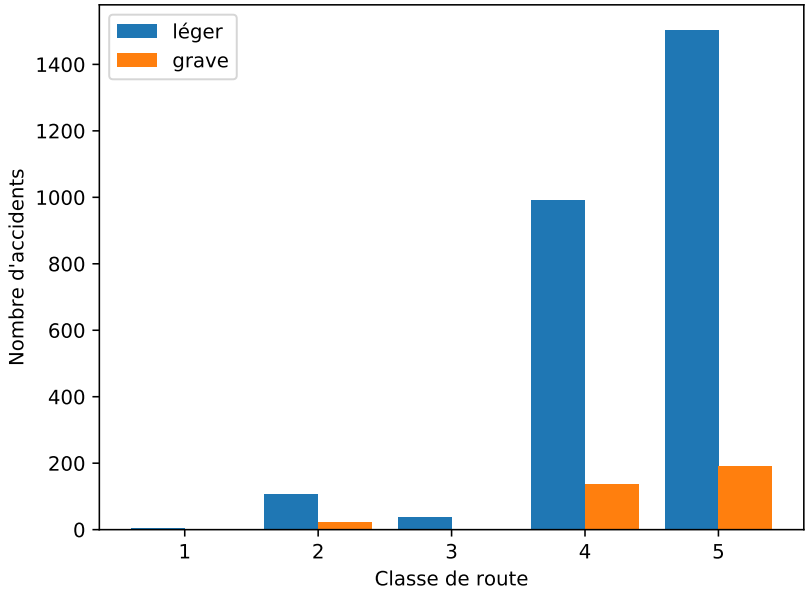
1. La gravité d'un accident est représentée par une variable nominale ordonnée. Un modèle logit ordonné est le plus adapté pour étudier l'association entre les attributs des accidents et leur gravité (variable dépendante). Un modèle multinomial pourrait aussi être utilisé, mais n'utiliserait pas l'information d'ordre des niveaux de gravité.

2. Des données nominales prenant K valeurs où $K \geq 3$ doivent être remplacées par $K - 1$ variables binaires représentant chacune des valeurs prises (par exemple, la variable de classe de la route sera représentée par quatre variables pour les autoroutes, les routes à numéro, les collectrice et les artères, les rues locales étant représentées par la valeur nulle (faux) de ces quatres variables binaires).

3. On crée la variable binaire "severity0" pour représenter les accidents mortels et avec blessures graves: la variable vaut 1 si l'accident est mortel ou avec blessures graves, 0 sinon (pour Tanagra, il peut être avantageux d'utiliser du texte pour les valeurs de cette variable de sorte qu'elle soit directement reconnue comme catégorielle (binaire)). Un exemple de fichier Tanagra est fourni. On voit dans le modèle que par exemple la variable de la limite de vitesse est significative (niveau de confiance de 95 %) et négativement corrélée, ce qui correspond aux connaissances en sécurité

routière (plus la vitesse est élevée, plus un accident est grave).
Réponse à compléter.

4. Le test statistique est le test du χ^2 qui permet de comparer deux échantillons de données. Il faut choisir un échantillon de référence, le normaliser et multiplier par le nombre d'observations de l'autre échantillon pour avoir les effectifs attendus. L'histogramme des nombres d'accidents selon le type de route est le suivant:



Le tableau pour le test du χ^2 en considérant le nombre d'accidents graves comme réference est le suivant (après regroupement des catégories de route pour lesquelles on a moins de cinq observations (seulement 4 accidents légers sur les route de catégorie 1):

Classe de route	Nombre d'accidents légers attendus	Nombre d'accidents légers observés
1 et 2	179.39	112
3	22.42	39
4	1016.54	991
5	1427.64	1504

On calcule le nombre d'accidents légers attendus s'ils suivaient la même distribution que les accidents graves (en divisant le nombre d'accidents graves de chaque catégories de route par le nombre total d'accidents graves et en multipliant par le nombre total d'accidents légers). L'hypothèse nulle du test est que les distributions sont identiques. Sous l'hypothèse nulle, la statistique du test suit la loi du χ^2 à $d = n - 1 - p = 4 - 1 - 0 = 3$ degrés de liberté. La statistique du test est $X^2 = 42.29$, ce qui correspond à un risque de première espèce inférieur à 10^{-6}. On rejette l'hypothèse nulle, les distributions des accidents légers et graves selon les classes des routes sont différentes. On peut conclure que la gravité des accidents n'est pas la même sur les différentes classes de routes.

7 Régression et modélisation économétrique

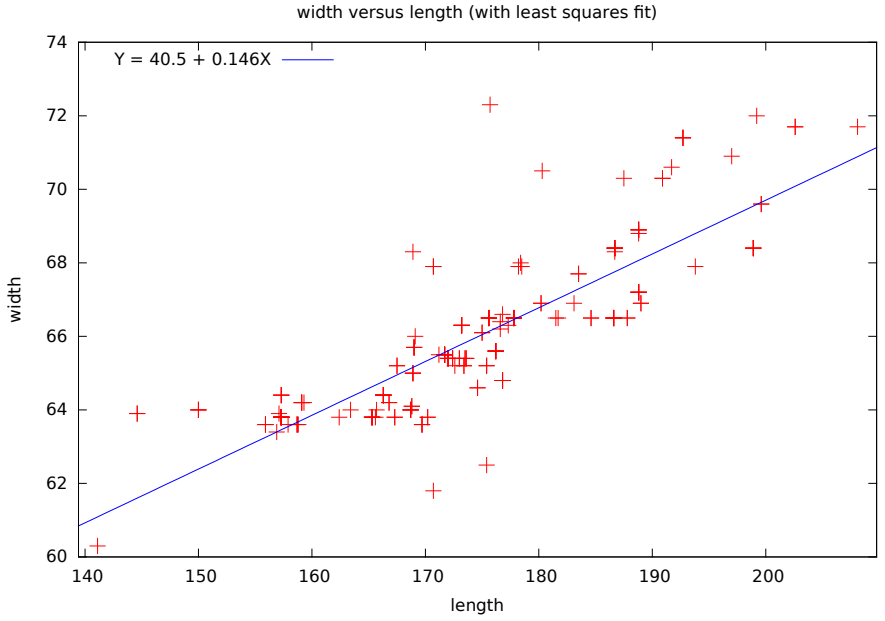
Exercice 1 (final 2014) 30 min (/4.5 pts)

Cet exercice repose sur le jeu de données de caractéristiques de voitures contenu dans le fichier `autos.txt`. Il vise à étudier la relation entre les deux variables longueur et largeur des voitures (colonnes "length" et "width"). Veuillez répondre aux question suivantes:

1. Tracer le nuage de points de la largeur en fonction de la longueur et calculer le coefficient de corrélation: commenter. (1 pt)
2. En utilisant un des logiciels à votre disposition, estimer la droite de régression linéaire de la largeur en fonction de la longueur:
   (a) Discuter la significativité du modèle et calculer (sans reprendre de Excel) l'intervalle de confiance à 90 % et 95 % du coefficient a de la longueur en notant que la statistique $\frac{\hat{a}-a}{s_{\hat{a}}}$ suit une loi de Student à $n-2$ degrés de liberté (avec $s_{\hat{a}} = \sqrt{\frac{\frac{1}{n-2}\sum_i (y_i-\hat{y}_i)^2}{\sum_i (x_i-\bar{x})^2}}$, y_i et x_i respectivement la largeur et la longueur du véhicule i, $\bar{x}$ la longueur moyenne empirique et $\hat{}$ désignant les termes estimés ou prédits). (2.5 pts)
   (b) En se basant sur l'étude graphique des résidus, indiquer si les hypothèses de régression linéaire sont respectées. (1 pt)

Solution

1. Le nuage de points de la largeur en fonction de la longueur est le suivant:



Le coefficient de corrélation linéaire entre les deux variables est 0.841, ce qui est très élevé. Plus un véhicule est large, plus il est long (dans cet ensemble de données).

2. (a) Les paramètres du modèle sont $\hat{a} = 0.146253$ et $\hat{b} = 40.452511$. Le modèle est significatif (risque de première espèce très faible, de l'ordre de 10^{-56}). L'intervalle de confiance de a est $[\hat{a} - t_{\alpha/2}s_{\hat{a}}, \hat{a} + t_{\alpha/2}s_{\hat{a}}]$ où $t_{\alpha/2}$ est telle que $P(|t| < t_{\alpha/2}) = 1 - \alpha$ pour une variable aléatoire t suivant la loi de Student à 203 degrés de liberté (la loi de Student tend vers la loi normale lorsque le nombre de degrés de libertés devient grand). On trouve $[0.1353, 0.1571]$ et $[0.1332, 0.1592]$ respectivement pour les niveaux de confiance de 90 % et 95 %.

   (b) Les hypothèses de régression linéaire concernant les résidus semblent respectées puisque les résidus sont également répartis de part et d'autres de l'axe des abscisses (moyenne nulle et variance constante). Il y a quelques points aberrants avec des erreurs plus grandes.

Exercice 2 (final 2020) 40 min (/5.5 pts)

Un modèle de régression linéaire multiple donne les résultats présentés dans les tableaux 3 et 4.

Tableau 3: Résultats globaux

R^2	1.000
R^2 ajusté	1.000
Statistique F	3309
Prob (>F)	0.0128
Observations	5

Tableau 4: Coefficients du modèle

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Constante	3.5347	0.162	21.759	0.029	1.471	5.599
x1	7.3775	0.239	30.870	0.021	4.341	10.414
x2	-4.9703	0.132	-37.551	0.017	-6.652	-3.289
x3	0.4180	0.165	2.531	0.240	-1.681	2.517

1. Est-ce une bonne idée d'ajouter une quatrième variable au modèle, sachant que sa corrélation avec la variable dépendante est très forte? Justifier. (1 pt)

2. Décrire la qualité du modèle, s'il est significatif, et indiquer les variables significatives du modèle. Justifier. (1 pt)

3. Une nouvelle collecte de données est effectuée incluant la quatrième variable et fournissant 100 observations fournies dans le fichier `exercice1.csv`.

   (a) En s'appuyant sur une visualisation de la distribution de la variable dépendante y, décrire la forme de la distribution (0.5 pt)

   (b) Proposer un modèle de la variable y en fonction des quatre variables indépendantes $x1$, $x2$, $x3$ et $x4$.

      i. Décrire la qualité de votre modèle et indiquer les variables significatives. (2 pts)

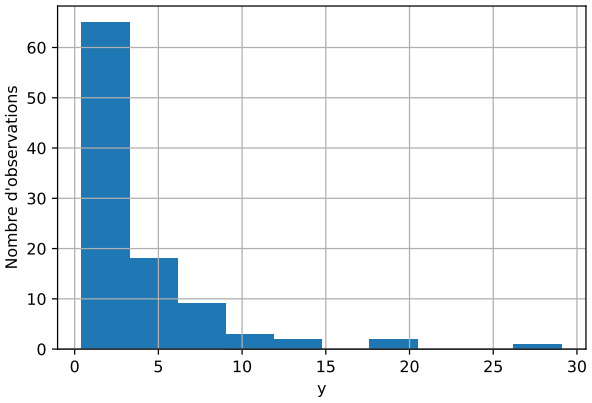
ii. En s'appuyant sur des visualisation des résidus, vérifier et commenter si les conditions d'estimations du modèle sont vérifiées. (1 pt)

Solution

1. Non, ce n'est pas une bonne idée car le nombre de variables p, incluant la constante, serait alors égal au nombre d'observations n et on doit avoir $p > n + 1$ (hypothèse H_9 d'application du modèle). Le modèle ne pourrait pas être estimé.

2. Le modèle à un R^2 élevé, en fait, "parfait", démontrant une relation linéaire entre les variables. Le modèle est "significatif", c'est-à-dire qu'on peut rejeter l'hypothèse nulle que tous les coefficients sont nuls avec un risque de se tromper de 0.0128. On cherche ensuite les variables pour lesquelles la valeur p du test pour les hypothèses nulles que chaque coefficient est nul est inférieure à une petite valeur, par exemple 0.05. Les variables $x1$ et $x2$ sont ainsi significatives. Le risque de rejeter l'hypothèse nulle que le coefficient pour la variable $x3$ est nulle est trop élevé (près d'une chance sur quatre de se tromper).

3. (a) L'histogramme de la variable y est présenté ci-dessous. La distribution n'est pas symétrique, la variable y ne peut donc pas suivre la loi normale (notons qu'il n'est pas nécessaire que la variable dépendante suive la loi normale).



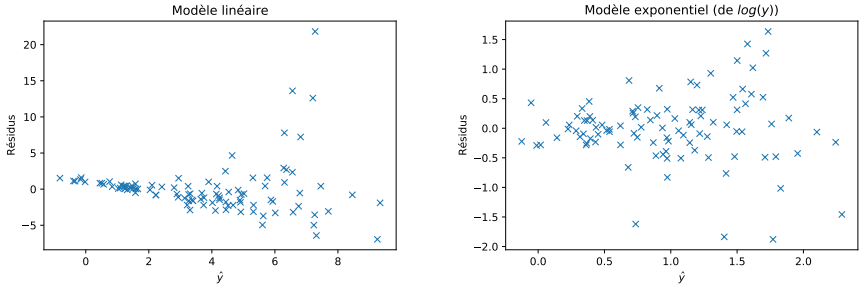
(b) i. On teste un modèle linéaire et un modèle exponentiel (avec le logarithme de y) de façon à avoir une variable distribuée plus symétriquement et le second modèle est meilleur (voir les graphiques des résidus ci-dessous). Les résultats du modèle exponentiel sont présentés dans le tableau ci-dessous. Le modèle est significatif dans son ensemble (valeur p très, très faible). Le R^2 est modérément élevé, sa valeur ajoutée étant 0.463, mais bien plus élevée que le R^2 ajusté du modèle linéaire (0.267). Les variables $x1$, $x2$ et $x3$ sont significatives avec une valeur p très faible, mais pas la variable $x4$.

Dep. Variable:	np.log(y)	**R-squared:**	0.484
Model:	OLS	**Adj. R-squared:**	0.463
Method:	Least Squares	**F-statistic:**	22.30
Date:	Tue, 14 Dec 2021	**Prob (F-statistic):**	5.27e-13
Time:	17:07:05	**Log-Likelihood:**	-86.029
No. Observations:	100	**AIC:**	182.1
Df Residuals:	95	**BIC:**	195.1
Df Model:	4		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.4046	0.198	2.044	0.044	0.012	0.798
x1	0.9597	0.194	4.936	0.000	0.574	1.346
x2	-1.0748	0.202	-5.327	0.000	-1.475	-0.674
x3	0.9990	0.205	4.862	0.000	0.591	1.407
x4	0.2493	0.200	1.248	0.215	-0.147	0.646

Omnibus:	12.804	**Durbin-Watson:**	2.259
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	26.155
Skew:	-0.429	**Prob(JB):**	2.09e-06
Kurtosis:	5.354	**Cond. No.**	6.78

ii. Les graphiques ci-dessous présentent respectivement à gauche et à droite les résidus du modèle de y et $log(y)$ en fonction des variables indépendantes. On voir clairement que la condition H_6 n'est pas respectée: la variance augmente avec $\hat{y}$. Ce problème est beaucoup moins prononcé pour le modèle exponentiel, mais on voit par contre quelques grandes erreurs pour un petit nombre d'observations (erreur supérieure à 1 en valeur absolue) qu'il serait intéressant de vérifier.



Exercice 3 (final 2021) 20 min (/3 pts)

Les résultats d'un modèle de choix discret de type logit sont présentés dans le tableau 5. La variable à prédire est le choix de l'avion pour un déplacement (les alternatives sont le train, le bus et la voiture). Les variables explicatives du modèle sont les suivantes:

- invt: temps de déplacement (en min)
- hinc: revenu du ménage (1000$)
- psize: nombre de personnes se déplaçant ensemble

Dep. Variable:	car	**No. Observations:**	210
Model:	Logit	**Df Residuals:**	206
Method:	MLE	**Df Model:**	3
Date:		**Pseudo R-squ.:**	0.8240
Time:		**Log-Likelihood:**	-21.775
converged:	True	**LL-Null:**	-123.76

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	9.7860	2.312	4.232	0.000	5.254	14.318
invt	-0.0469	0.009	-5.185	0.000	-0.065	-0.029
hinc	0.0291	0.020	1.460	0.144	-0.010	0.068
psize	-1.0656	0.536	-1.988	0.047	-2.116	-0.015

Tableau 5: Résultat d'un modèle logit

Veuillez répondre aux questions suivantes:

1. Décrire la qualité du modèle, s'il est significatif, et indiquer les variables significatives du modèle. Justifier. (1 pt)

2. Expliquer comment comparer les poids relatifs des différentes variables indépendantes. (0.5 pt)

3. Quel modèle permettrait d'étudier les facteurs associés au choix du mode de transport parmi au moins trois modes. (0.5 pt)

4. Discuter une méthode d'enquête permettant de recueillir de telles données (population de référence, type d'enquête et technique d'enquête). (1 pt)

8 Visualisation de données

9 Fouille de données et apprentissage automatique

Exercice 1 (final 2010) 35 min (/5 pts)

Le fichier `vitesse-debit.csv` (dans `10-vitesse-debit.zip`) contient des observations des vitesses (en km/h) et débits (en nombre de véhicules par heure) par intervalle de 15 min pour une direction d'une route rurale à deux voies.

1. Décrire deux méthodes, une intrusive et une non-intrusive, de collecte de données de vitesse sur une route (citer un avantage et un inconvénient pour chacune). (0.5 pt)

2. Après exploration visuelle des données, proposer une répartition des données en groupes "homogènes": justifier vos choix, caractériser ces groupes par leurs statistiques descriptives sommaires et proposer une courte description qualitative des groupes. (2 pts)

3. En prenant les dix premières observations du fichier comme exemple, illustrer en quelques étapes (au moins 3 étapes, dont l'initialisation et l'étape finale) le fonctionnement d'un algorithme de segmentation des données. Discuter le ou les traitement(s) nécessaire(s) préalable(s) à la segmentation. (2 pts)

4. Indiquer comment il serait possible de représenter une troisième variable (par exemple la proportion de poids lourds par intervalle de 15 min sur cette route) dans un nuage de points dans l'espace des vitesses et des débits. (0.5 pt)

Exercice 2 (final 2020) 70 min (/9.5 pts)

Cet exercice repose sur un jeu de données de circulation collectées sur une autoroute de la région métropolitaine de Portland pour sept jours consécutifs en septembre 2011, disponible dans le fichier `portland-1395.csv`. Les données sont agrégées par intervalle de 20 s et les attributs sont les suivants:

- detectorid: identifiant du détecteur
- starttime: date et heure du début de l'intervalle
- volume: nombre de véhicules détectés dans l'intervalle de 20 s
- speed: vitesse moyenne des véhicules
- occupancy: taux d'occupation (proportion du temps que le capteur est occupé par un véhicule)
- status: statut du détecteur (non-utilisé)
- dqflags: indicateur de qualité (non-utilisé)
- date: date déduite de l'attribut starttime

Veuillez répondre aux questions suivantes:

1. Décrire une technologie de capteur pour collecter les trois attributs volume, speed et occupancy, et indiquer un avantage et un inconvénient. (1 pt)
2. Choisir deux journées et faire les analyses suivantes:
   (a) calculer les intervalles de confiance à 95 % des vitesses pour chaque journée; (1 pt)
   (b) comparer les vitesses moyennes à l'aide d'un test statistique; (1 pt)
   (c) tester l'adéquation de la distribution des vitesses d'une de ces journées à la loi normale. (1.5 pts)
3. Expliquer (sans le faire) comment comparer à l'aide d'un test statistique les moyennes de la variable volume selon les jours et les conditions pour appliquer le test. (1 pt)
4. À l'aide d'une méthode de segmentation, regrouper les conditions de circulation (décrites par les trois attributs volume, speed et occupancy) et décrire les groupes résultants. Choisir un petit nombre de groupes (2 à 4). (3 pts)
5. Proposer (sans la réaliser avec les données) une visualisation graphique des groupes et des trois attributs utilisés pour créer les groupes. (0.5 pt)
6. Décrire (sans l'appliquer) une méthode d'apprentissage supervisé permettant d'identifier les variables importantes dans les groupes obtenus par la méthode de segmentation. (0.5 pt)

Solution

1. Les boucles magnétiques installées dans la chaussée permettent de collecter des données de débit (volume) et taux d'occupation (occupancy). Des paires de boucles sont nécessaires pour mesurer la vitesse (speed).

2. On choisit les deux premières journées, des 15 et 16 septembre 2011.

   (a) Les intervalles de confiance de la moyenne des vitesses sont respectivement $49.12 \pm 1.96\frac{11.10}{\sqrt{3305}} = [48.75, 49.50]$ et $49.63 \pm 1.96\frac{11.60}{\sqrt{4153}} = [49.28, 49.99]$.

   (b) On utilise l'approximation normale pour comparer les moyennes des vitesses entre les deux journées du fait du grand nombre d'observations. L'hypothèse nulle est que la vitesse moyenne est la même pour les deux jours, alors que l'hypothèse alternative est qu'elle est différente (on n'a pas d'ordre privilégié). La statistique du test est $Z_0 = -1.9356$, ce qui est plus petit (en valeur absolue) que 1.96, ce qui correspond à un risque de première espèce supérieur à 0.05. On ne peut pas rejeter l'hypothèse nulle que les moyennes sont égales.

   (c) On teste l'adéquation des vitesses du 15 septembre 2011 à la loi normale à l'aide du test du χ^2. Il faut compter les observations dans des intervalles de valeurs de vitesse. Après calcul des échantillons attendus sur la vitesse suivait la loi normale (moyenne et écart-type de mêmes valeurs que les vitesses) et regroupement des intervalles avec cinq observations ou moins, on a le tableau suivant:

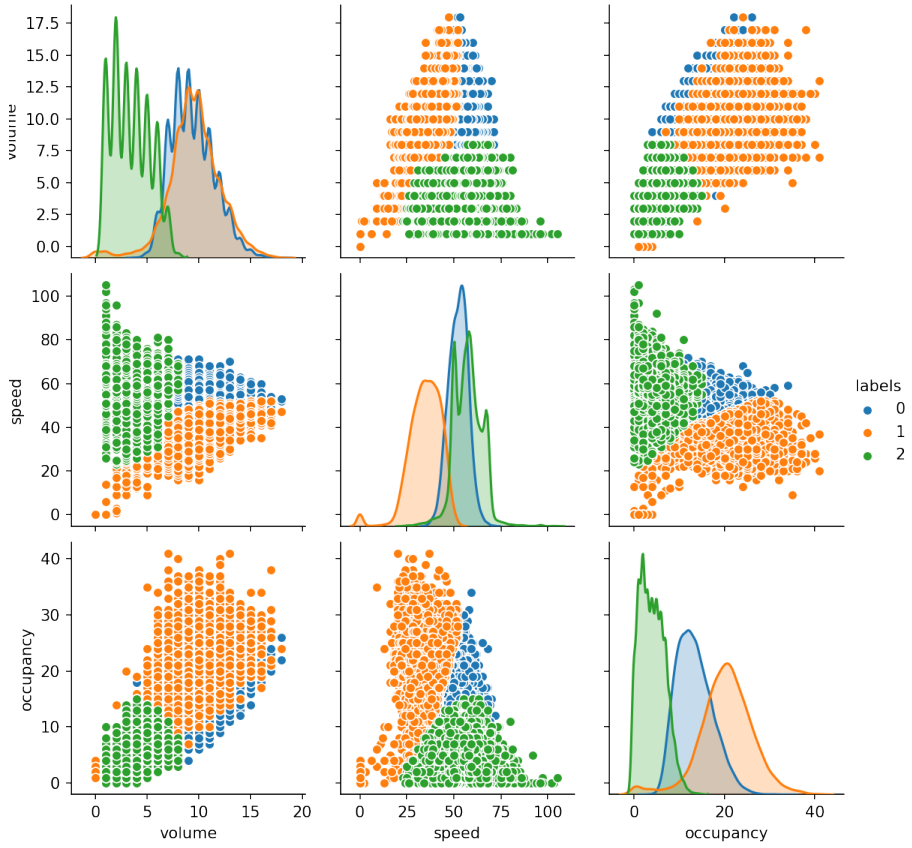
Intervalles	Nombre d'observations	Nombre attendus
$[0.0, 9.6]$	15	11.62
$[9.6, 19.2]$	142	99.35
$[19.2, 28.8]$	458	441.09
$[28.8, 38.4]$	576	967.21
$[38.4, 48.0]$	1409	1050.10
$[48.0, 57.6]$	602	564.64
$[57.6, 67.2]$	93	150.06
$[67.2, 76.8]$	10	20.94

   La statistique du test vaut 330. La probabilité qu'une variable aléatoire suivant le test du χ^2 à $d = 8 - 1 - 2 = 5$ degrés de liberté est 0.0 (la valeur seuil pour 0.05 est 11.07). On peut donc rejeter avec une grande confiance l'hypothèse nulle, les vitesses ne suivent pas la loi normale le 15 septembre 2011.

3. Il faut appliquer le test ANOVA. Pour cela, on calcule les moyennes du débit (variable volume) pour chaque jour de la semaine, avec un tableau croisé dynamique par exemple. Il faut ensuite calculer la statistique du test à partir des sommes des carrés totaux et expliqués, et comparer la valeur obtenue à la loi de Fisher. Les conditions d'applications sont les suivantes:

   - *Normalité* de la distribution: on suppose, sous l'hypothèse nulle, que les échantillons de chaque groupe suivent une loi normale (équivalent à ce que les résidus suivent une loi normale)
   - *Homoscédasticité*: homogénéité des variances entre chaque groupe
   - *Indépendance* des échantillons: on suppose que chaque échantillon analysé est indépendant des autres échantillons

4. Il faut appliquer une méthode de segmentation, par exemple l'algorithme des k-moyennes. On peut tracer les nuages de points suivants pour étudier la répartiton des trois groupes selon les trois variables considérées. En s'appuyant sur ces graphiques et sur le calcul des centres de chaque groupe, on voit clairement que le groupe 2 est celui des vitesses les plus basses et taux d'occupation les plus élevés, donc des conditions les moins fluides. Le groupe 1 semble avoir les vitesses les plus élevées et a les taux d'occupation et débits les plus faibles, il s'agit des conditions de circulation les plus fluides. Le groupe 0 correspond à des conditions intermédiaires entre ces deux groupes.



5. On peut faire une matrice de nuage de points avec la couleur pour représenter les groupes comme fait ci-dessus.

6. Il est possible d'apprendre un arbre de décision pour prédire à le groupe auquel appartient chaque observation à partir des variables explicatives. La première variable choisie sera la plus importante pour prédire le groupe.

Exercice 3 (final 2021) 30 min (/4 pts)

Un arbre de décision a été appris à partir du jeu de données des survivants du Titanic (fichier `titanic.txt` disponible sur Moodle). Chaque personne est décrite par quatre attributs: son statut sur le bateau (équipage, 1ère, 2ème ou 3ème classe), son âge (adulte ou enfant), son genre (homme ou femme) et s'il a survécu ou pas. L'arbre de décision est présenté dans la figure 1 et sur Moodle pour la lisibilité.

1. Veuillez décrire les types des attributs du jeu de données: à quelle catégorie appartient la tâche de prédire la survie d'un passager? (1 pt)

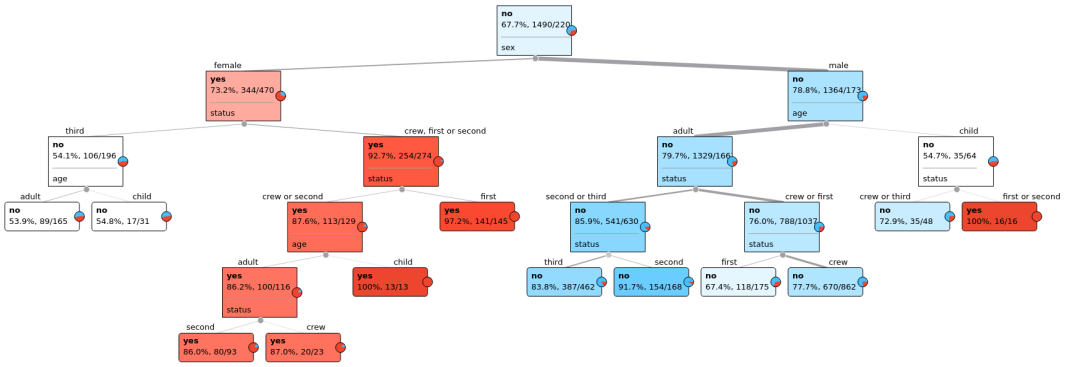


Figure 1: Arbre de décision de la survie d'un passager du Titanic.

2. Quelle est l'attribut le plus important pour prédire la survie des passagers selon l'arbre de décision? Justifier la réponse. (1 pt)

3. Écrire deux règles complètes de décision de l'arbre (de la racine à la feuille) de niveaux de "confiance" très différents. (1 pt)

4. Quel autre modèle pourrait être utilisé pour prédire la survie des passagers? Quels sont les avantages des différents modèles? (1 pt)

Solution

1. La variable prédite (si un passager a survécu ou pas au naufrage du Titanic) est catégorielle, binaire. Prédire cette variable est donc une tâche de classification.

2. Le sexe (genre) est l'attribut le plus important pour prédire la survie des passagers puisque c'est le premier choisi par l'arbre pour la classification.

3. Si sex = female et status = third et age = adulte, alors survie = no pour 53.9 % des personnes concernées
   Si sex = male et age = child et status = first or second, alors survie = yes pour 100 % des personnes concernées

4. Un autre modèle pouvant faire de la classification est un classifieur bayésien naïf, avec l'avantage qu'il fonctionne directement avec des attributs catégoriels.

10 Analyse spatiale

Exercice 1 (périodique 2013) 45 min (/6 pts)

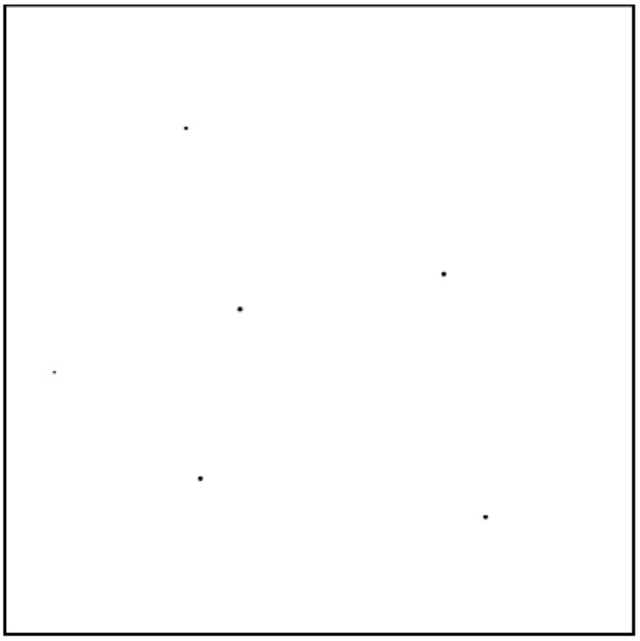
1. Donner un exemple de mesure de centralité en analyse spatiale et expliquer à quoi cela peut servir dans un contexte de transport. (1 pt)

2. Donner un exemple de mesure de dispersion en analyse spatiale et expliquer à quoi cela peut servir dans un contexte de transport. (1 pt)

3. À quoi servent l'indice Moran I global et l'indice Geary C global? Dans quel contexte pourrait-on les utiliser en transport? (1 pt)

4. Est-ce possible d'obtenir une projection de la terre sur un plan qui conserve les distances, les formes, les angles et les superficies? Pourquoi? (1 pt)

5. Dessiner à quoi ressembleraient le diagramme de Voronoi (selon la distance euclidienne) autour des points ci-dessous (ne pas mesurer exactement, faire seulement une esquisse). (2 pts)



Exercice 2 (périodique 2014) 30 min (/3 pts)

1. Types de distribution:

   (a) Dessiner à quoi ressemblerait un motif de points généré par un processus ponctuel avec une structure spatial totalement aléatoire dans le cadre ci-dessous. (0.5 pt)

(b) Quelle serait la valeur de l'indice de Moran I et de l'indice de Geary C pour cette distribution? (0.5 pt)

2. Donner deux exemples pertinents d'utilisation des polygones de Thiessen (ou diagrammes de Voronoï) dans le domaine du transport. (1 pt)

Solution

1. Types de distribution:
   (a) Tracer des points selon la loi uniforme indépendamment pour les deux coordonnées.
   (b) $I = 0$ et $C = 1$.
2. Par exemple, les polygones permettent de convertir des données ponctuelles en données zonales, d'estimer la population cible pour un arret de bus, de delimiter les frontieres d'une zone pour un menage.

Exercice 3 (final 2021) 35 min (/4.5 pts)

Trois couches de données spatiales au format shapefile sont fournies pour cet exercice dans l'archive `donnees-spatiales.zip` (sur Moodle), soit les limites des zones de l'enquête OD de la grande région de Montréal de 2013 et deux couches de points. Veuillez répondre aux questions suivantes:

1. Comparer les ensembles de points des deux couches à l'aide de mesures de centralité et de dispersion (2 pts)
2. Décrire visuellement si les points des deux couches sont répartis de façon similaire: commenter le ou lesquels pourraient avoir une structure spatiale totalement aléatoire. (1 pt)
3. Décrire une méthode permettant de caractériser l'intensité des ensembles de points. (0.5 pt)

4. Décrire une procédure de traitement spatial pour caractériser les points selon les communes. (1 pt)

Solution

1. Il faut calculer le centre des points (disponible dans QGIS) et l'écart-type selon chaque dimension ou l'écart-type de la distance au centre (aucun outil ne semble disponible dans QGIS, une façon de faire est d'exporter les données en fichier csv pour traitement dans un outil externe comme Excel).

2. Les points ne sont pas répartis de façon similaire dans les deux couches de points fournies (points1 et points2). Les points de la couche « points2 » semblent répartis de façon uniformes selon les deux dimensions, ces points pourraient donc avoir une structure spatiale totalement aléatoire.

3. L'intensité est le nombre de points par unité de surface : il faut donc définir une partition de l'espace, puis faire le calcul pour chaque élément. On a vu trois exemples dans le cours, les quadrats (on fait une partition, puis il faudrait calculer l'intentité par quadrat), les cartes de chaleurs (similaire aux quadrats dans le découpage de l'espace, mais avec une fonctions qui peut lisser le nombre moyens de points en chaque case) et le diagramme de Voronoi (par définition, il y a un point par zone du diagramme, donc l'intensité y est l'inverse de la surface de chaque zone).

4. Pour caractériser les points par commune, il faut faire une jointure spatiale qui identifie pour chaque point la zone (commune) dans laquelle il se situe.

Exercice 4 (final 2021) 30 min (/4 pts)

La figure 2 présente le nuage de points de Moran pour le nombre d'accidents par zone (arrondissement ou municipalité) de l'Île de Montréal pour l'année 2018 selon le critère de proximité de la tour ("Rook"). L'indice I de Moran global vaut 0.362 avec une pseudo valeur p de 0.001.

1. Est-ce que la variable du nombre d'accidents par zone présente une auto-corrélation spatiale? Justifier la significativité. Est-ce surprenant? (1.5 pt)

2. Décrire les quatre quadrants du nuage de points de Moran, en particulier lesquels démontrent une autocorrélation positive, et le lien avec l'indice I de Moran. (1 pt)

3. Quelle est la différence avec le critère de proximité de la reine ("Queen")? Pensez-vous que l'indice I de Moran sera différent avec le critère de la reine? Justifier. (1 pt)

4. Quelle mesure permettrait de déterminer les zones avec des similarités ou dissimilarités particulières? (0.5 pt)

5. Expliquer comment la pseudo valeur p est obtenue (0.5 pt bonus)

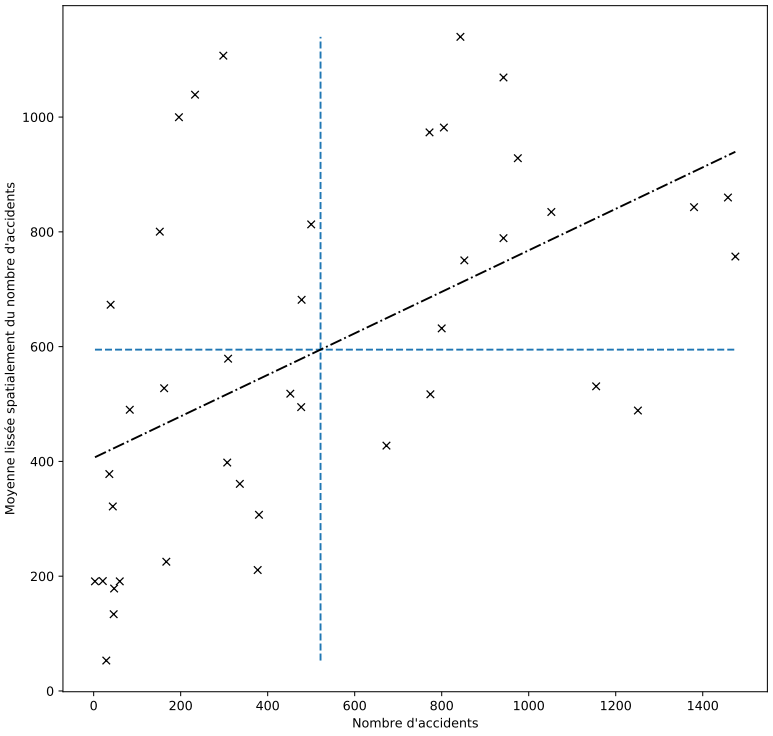


Figure 2: Nuage de points de Moran du nombre d'accident par zone de l'Île de Montréal.

Solution

1. Oui. La valeur p de 0.001 indique un petit risque de se tromper en rejetant l'hypothèse nulle d'absence d'auto-corrélation spatiale des résidus. Ce n'est pas surprenant puisque les accidents sont liés à des facteurs spatiaux, comme les aménagements routiers (limites de vitesse, longueur d'autoroutes, mesures d'apaisement de la circulation), qui sont aussi corrélés spatialement.

2. les quadrants en haut à droite et en bas à gauche montre une auto-corrélation positives (respectivement des zones avec un nombre d'accidents et un nombre lissé spatialement plus grands ou plus bas que les moyennes respectives). Les deux autres quadrant dénotent des zones qui se démarquent de leurs voisins (nombre d'accidents plus élevé que la moyenne et nombre lissé plus bas (quadrant bas droite) ou réciproquement (quadrant haut gauche)). L'indice I de Moran est lié à la pente de la droite de régression du nuage de points de Moran.

3. les zones "en coin" seront aussi considérées commes voisines
peu de différence attendue avec le critère queen (vu en classe)

4. L'indice de Moran I Local

5. Par simulation de Monte Carlo en faisant des permutations des valeurs du nombre d'accidents parmi toutes les zones
   Si les étudiants mentionnent que c'est obtenu par simulation, on pourrait donner le demi point