Contexte

- un individu est un bien proposé à la location
- la population est l'ensemble des biens proposés et sa taille est 135
- on trouve 17 variables.
  - Les modalités de id et host_id sont des nombres entiers, donc à priori des variables quantitatives discrètes, mais la nomenclature indique que ce sont des identifiants. Il faut donc les considérer comme des variables qualitatives nominales car la valeur numérique d'une observation n'a pas de sens.
  - host_response_time et host_is_superhost sont des variables qualitatives. host_response_time peut être considérée comme ordinale en ordonnant les valeurs de la réponse la plus rapide à la plus lente. host_is_superhost est binaire, donc nominale.
  - host_response_rate et host_acceptance_rate sont des valeurs entières de pourcentage. On devrait les traiter comme des variables quantitatives discrètes, mais on pourra aussi les considérer comme continues si les valeurs observées sont trop nombreuses.
  - host_listings_count, host_total_listings_count, accommodates, bathrooms_text, bedrooms, beds, number_of_reviews prennent des valeurs numériques discrètes. Le nombre de modalités effectivement observées est suffisamment faible pour les traiter comme des variables quantitatives discrètes. Cependant, on observe quelques modalités extrêmes pour certaines de ces variables.
  - latitude, longitude, price, review_scores_rating sont des variables quantitatives. Pour price, on pourrait aussi considérer que les prix sont des entiers, mais le grand nombre de modalités distinctes incite à considérer ce caractère comme continu.

host_response_time

Après tri, la série comporte 67 valeurs et donc 68 données manquantes. Le tableau complet s'écrit :

xi	ni	fi	ni+	fi+	ni-	fi-
within an hour	36	53,73 %	36	53,73 %	67	100,00 %
within a few hours	18	26,87 %	54	80,60 %	31	46,27 %
within a day	8	11,94 %	62	92,54 %	13	19,40 %
a few days or more	5	7,46 %	67	100,00 %	5	7,46 %
total	**67**	**100,00 %**				

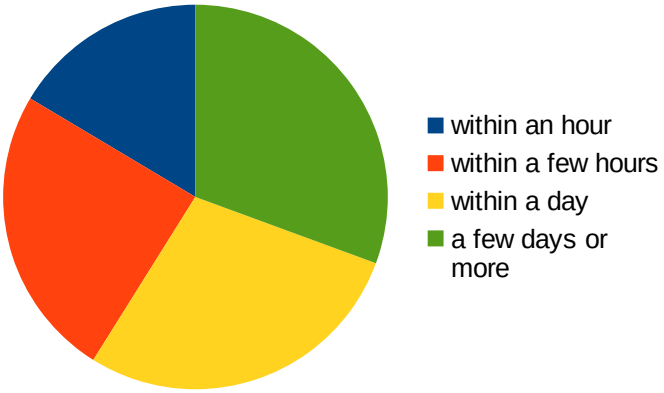
On en déduit que

- 62 hôtes répondent au maximum dans la journée (n_i croissant)
- 46,27% des hôtes ne répondent pas dans l'heure qui suit.

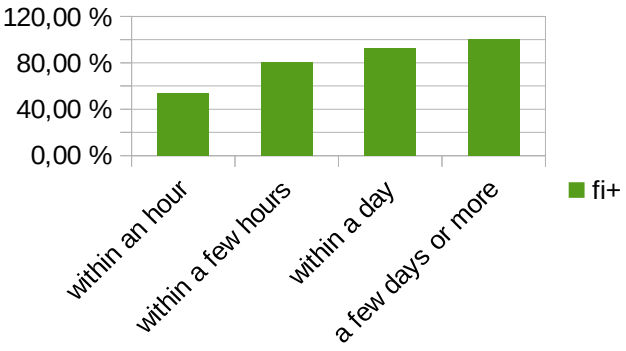
Pour les graphiques en secteur, l'information apportée est la même pour la fréquence et pour les effectifs car ce type de graphique ne rend compte que des fréquences, par nature. Chaque secteur est proportionnel à l'effectif et la surface totale est constante quelle que soit la taille de la population.

Pour les graphiques en bâtons des effectifs, l'échelle renseigne sur les valeurs des effectifs. Si l'on souhaite faire des comparaisons entre populations de tailles différentes, il faudra préférer le diagramme des effectifs et adopter une échelle commune, éventuellement sur le même diagramme.

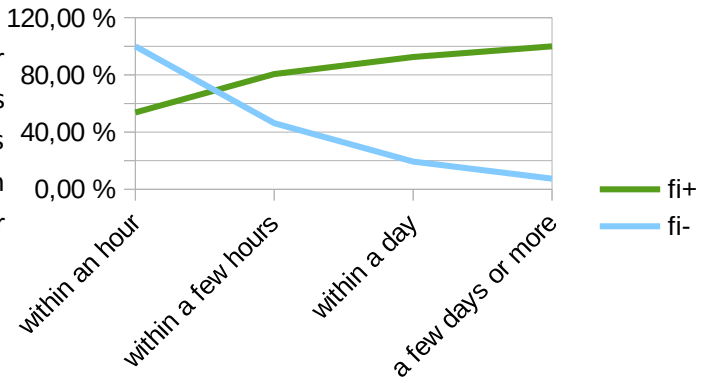
Le diagramme en secteur des fréquences cumulées croissantes ne permet pas de visualiser une information pertinente ou facile à lire : il est inapproprié et ne doit pas être utilisé. Le diagramme en secteur doit être limité aux variables qualitatives nominales ayant peu de modalités.



Le diagramme en bâton permet de lire l'information du type "quel est le délai à partir duquel on est sûr que 75% (par exemple) des hôtes répondent. Il est donc légitime de produire un tel diagramme. Cependant, on peut lui préférer la courbe des fréquences cumulées.



La visualisation des quantités cumulées est facile lorsque les points sont reliés par des segments de droites car les pentes sont plus lisibles qu'avec des bâtons. Le point de concours des deux courbes renseigne sur la localisation d'un point ayant une certaine centralité (à compléter plus tard par la notion de médiane).



Accommodates

La série quantitative discrète accomodates n'a pas de données manquantes. On dispose de 135 observations. En triant et en rajoutant les modalités non observées, on obtient le tableau :

xi	ni	fi	ni+	fi+	ni-	fi-
2		3	2,22 %	3	2,22 %	135	100,00 %
3		0	0,00 %	3	2,22 %	132	97,78 %
4		13	9,63 %	16	11,85 %	132	97,78 %
5		5	3,70 %	21	15,56 %	119	88,15 %
6		34	25,19 %	55	40,74 %	114	84,44 %
7		12	8,89 %	67	49,63 %	80	59,26 %
8		36	26,67 %	103	76,30 %	68	50,37 %
9		4	2,96 %	107	79,26 %	32	23,70 %
10		17	12,59 %	124	91,85 %	28	20,74 %
11		2	1,48 %	126	93,33 %	11	8,15 %
12		5	3,70 %	131	97,04 %	9	6,67 %
13		2	1,48 %	133	98,52 %	4	2,96 %
14		1	0,74 %	134	99,26 %	2	1,48 %
15		0	0,00 %	134	99,26 %	1	0,74 %
16		1	0,74 %	135	100,00 %	1	0,74 %
total		135	100,00 %				

On déduit de ce tableau

- la capacité la plus fréquemment proposée est de 8 personnes
- 50,37% des logements peuvent accueillir plus de 8 personnes
- 40,74% des logements ne peuvent pas accueillir un groupe de 7 personnes

Pour connaître la taille du groupe qui n'a accès qu'à 30% des offres, il faut lire la courbe des fréquences cumulées décroissantes. Pour la modalité 9, 23,70% des offres sont accessibles. Pour la modalité 8, 50,37% des offres sont accessibles. La modalité 9 est donc la plus petite taille des groupes qui n'ont accès qu'à 30% des offres au maximum. Sur le graphique, il faut tracer une ligne horizontale à 30% et chercher la modalité immédiatement supérieure au point d'intersection.

bathrooms_text, bedrooms et beds

bathrooms_text est une variable quantitative discrète. Le filtre standard "sans doublons" donne les modalités observées. On constate que la progression n'est pas toujours constante, car il manque la valeur 6. Les valeurs non-entières indiquent la présence d'une salle d'eau. Par exemple 2,5 peut signifier 2 salles de bain et 1 salle d'eau, ou 1 salle de bain et 3 salles d'eau, etc ... . Pour les représentations graphiques, il faudra rajouter la modalité 6. Le tableau statistique complet s'écrit :

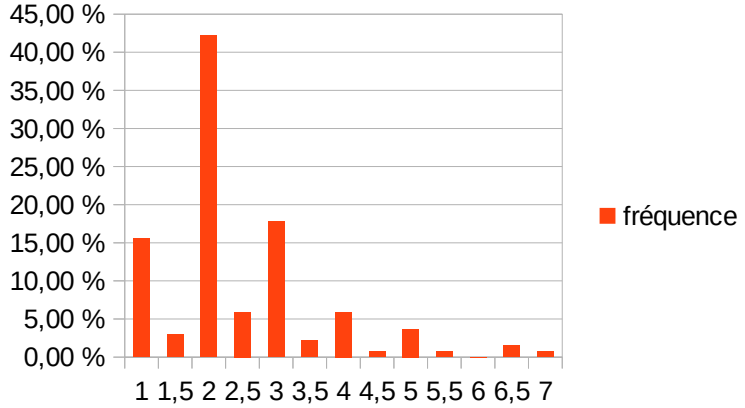
x_i	n_i	f_i	n_i+	f_i+	n_i-	f_i-
1	21	15,56 %	21	15,56 %	135	100,00 %
1,5	4	2,96 %	25	18,52 %	114	84,44 %
2	57	42,22 %	82	60,74 %	110	81,48 %
2,5	8	5,93 %	90	66,67 %	53	39,26 %
3	24	17,78 %	114	84,44 %	45	33,33 %
3,5	3	2,22 %	117	86,67 %	21	15,56 %
4	8	5,93 %	125	92,59 %	18	13,33 %
4,5	1	0,74 %	126	93,33 %	10	7,41 %
5	5	3,70 %	131	97,04 %	9	6,67 %
5,5	1	0,74 %	132	97,78 %	4	2,96 %
6	0	0,00 %	132	97,78 %	1	0,74 %
6,5	2	1,48 %	134	99,26 %	3	2,22 %
7	1	0,74 %	135	100,00 %	1	0,74 %
Total	**135**	**100,00 %**				

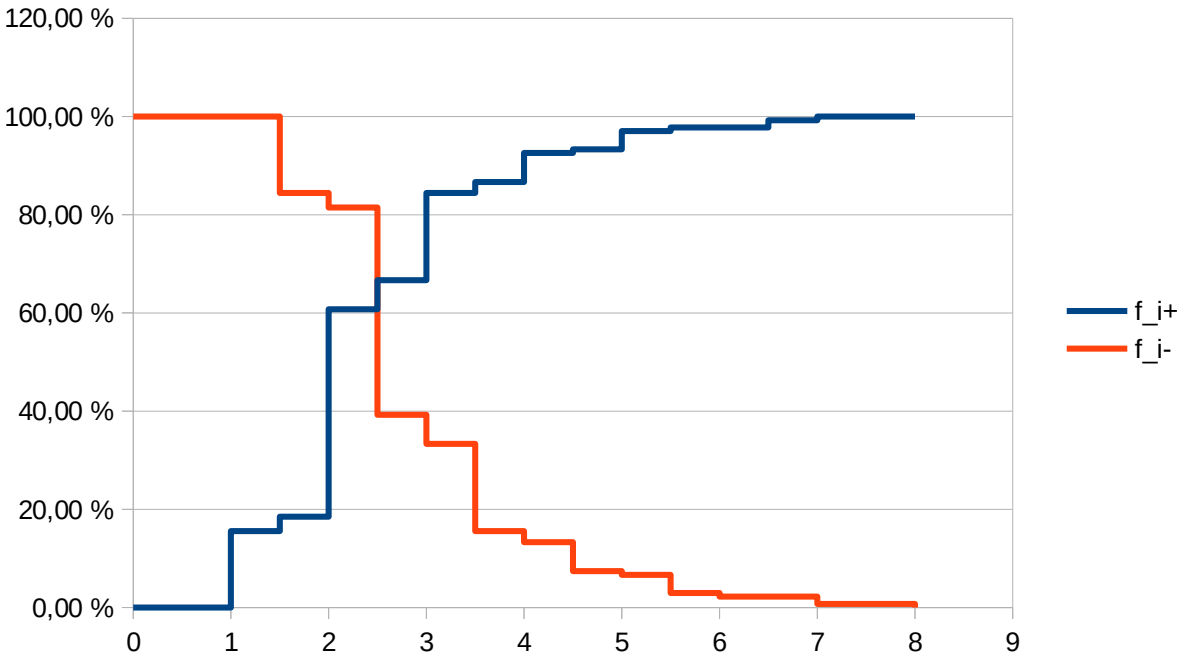
La lecture de ce tableau donne les réponses aux questions du type :

- Quelle est le nombre de logements proposés disposant de 4 salles de bain ? (8);
- Quelle est le nombre de salles de bain le plus fréquemment proposé ? (2);
- Combien de logements disposent d'au moins 3 salles de bain ? (45);
- Quelle proportion de logements ne propose pas plus de 2 salles de bain ? (60,74%).

Les représentations graphiques adaptées sont

- le diagramme en bâtons pour les fréquences;
- le tracé des courbes cumulatives.





Les deux autres variables bedrooms et beds sont aussi quantitatives discrètes et se résument de la même manière.

host_listings_count

D'après le dictionnaire des données, cette variable représente le nombre d'annonces appartenant à l'annonceur. Cette variable quantitative discrète pose plusieurs problèmes lisibles sur le tableau des effectifs :

host_listings_count	0	1	2	3	4	5	8	9	16	19	20	64
n_i	47	58	14	4	2	1	1	1	1	1	4	1

- la modalité 0 semble contradictoire avec le fait que l'annonceur possède au moins une annonce.
- Certains hôtes ont plus d'une annonce dans cet extrait. Il faudrait donc d'abord ne garder qu'une annonce par hôte. Les données brutes de cette série devraient donc être pré-traitées.
- Le nombre de modalités observées est faible (11), mais elles sont réparties très inéquitablement entre 0 et 64. Les 4 premières modalités représentent plus de 90 % des observations sur un intervalle de 3 unités, et les 7 autres moins de 10 % sur un intervalle de 61. Les diagrammes seront vraisemblablement illisibles.

Une suggestion possible est de considérer cette variable comme qualitative ordinale avec les modalités 0, 1, 2 et plus de 3.