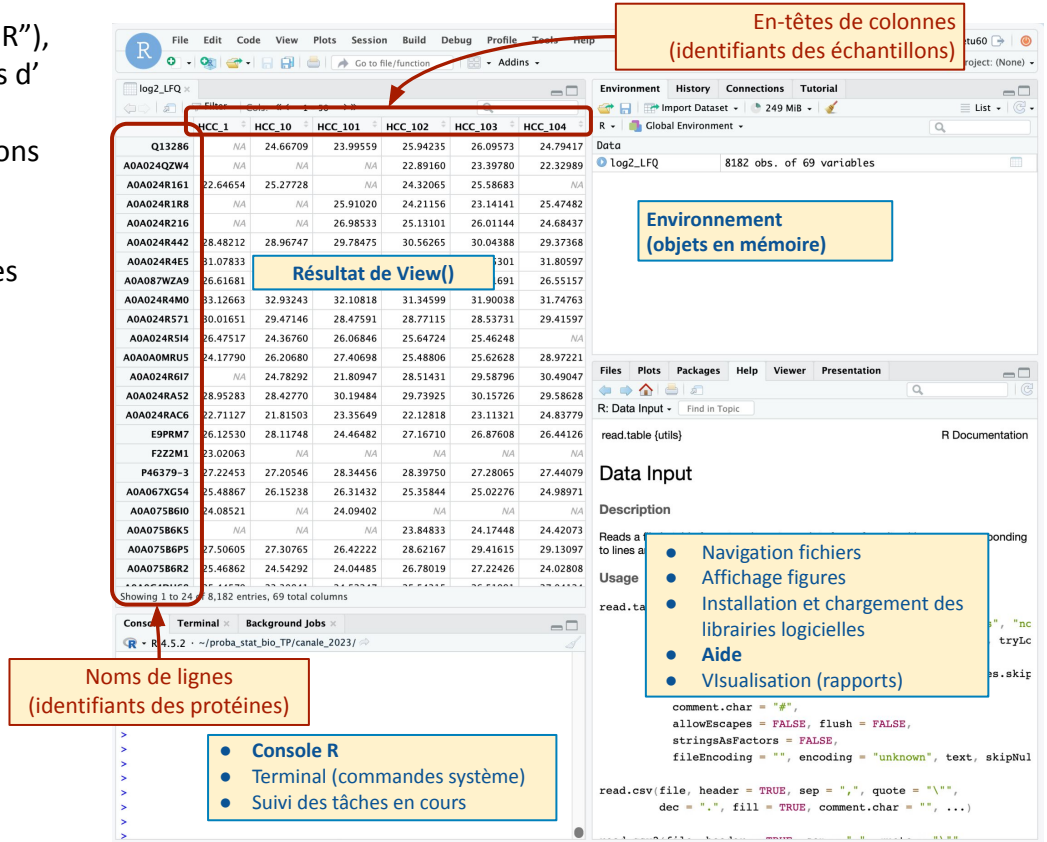
Chapitre 3. Apprivoiser vos données : statistiques descriptives et exploratoires

Andreas Zanzoni (orcid.org/0000-0002-4818-6161)
Jacques van Helden (orcid.org/0000-0002-8799-8584)
Lou BERGOGNE (orcid.org/0009-0004-9409-0154)

Cas d'étude

Dans le chapitre précédent (voir chapitre "Premiers pas avec R"), nous avons chargé dans R un tableau de données de notre cas d' étude : étude protéomique de cellules de foie (échantillons étiquetés "**Liver**") et de carcinome hépatocellulaire (échantillons étiquetés "**HCC**").

Ces données ont été chargées dans une data.frame nommée log2_LFQ, qui contient 69 colonnes (échantillons) x 8182 lignes (protéines).

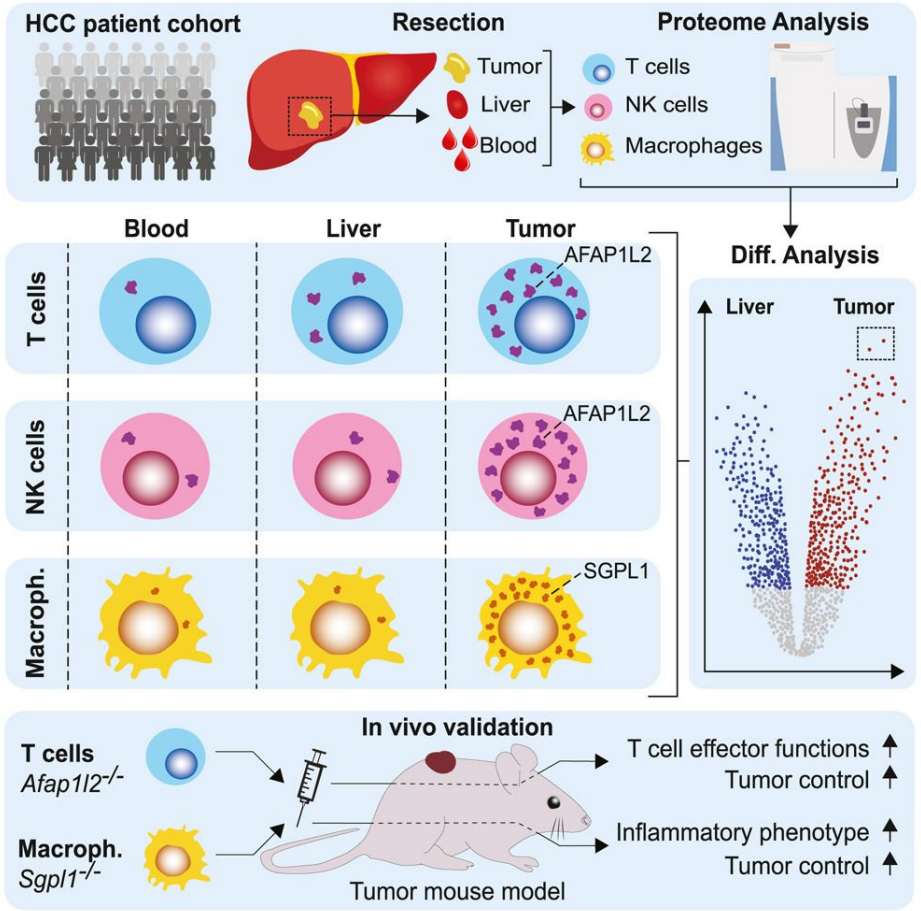


Canale, F. P. et al. Proteomics of immune cells from liver tumors reveals immunotherapy targets. Cell Genomics 3, 100331 (2023). doi.org/10.1016/j.xgen.2023.100331

Objectifs

Dans ce chapitre, nous utiliserons différentes **statistiques descriptives** et **méthodes de visualisation** pour **explorer les données** protéomiques, décrire leurs propriétés et analyser leur composition et leur structuration.

Dans la mesure du possible, nous veillerons à **appréhender les données dans leur ensemble** plutôt que de nous limiter aux paramètres descriptifs habituels.

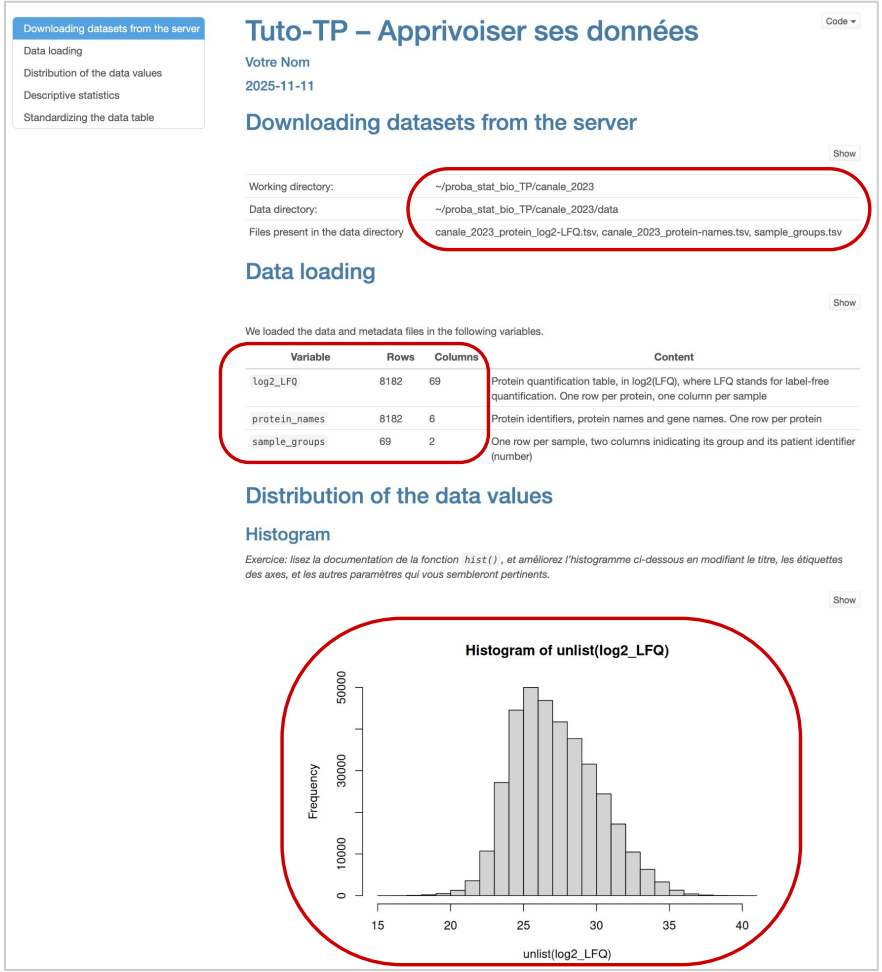


Canale, F. P. et al. Proteomics of immune cells from liver tumors reveals immunotherapy targets. Cell Genomics 3, 100331 (2023). doi.org/10.1016/j.xgen.2023.100331

Trame de rapport R markdown

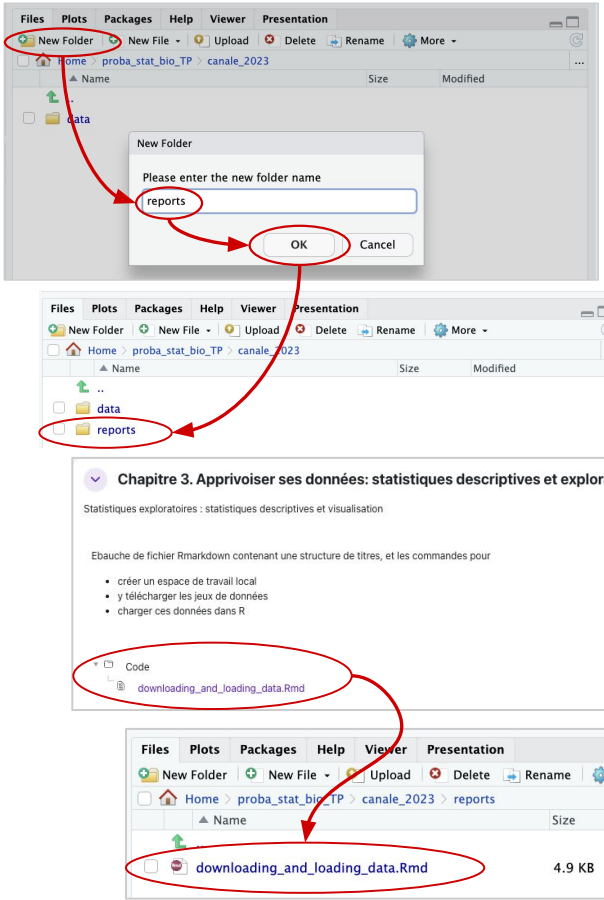
Nous avons préparé une **trame de rapport R markdown** qui effectue pour vous les étapes suivantes :

- **Crée un dossier de travail** (working directory) sur votre ordinateur. Le chemin de ce dossier par défaut est `~/proba_stat_bio_TP/canale_2023/` mais vous pouvez le modifier si vous le désirez. Tous les autres chemins seront définis par rapport à ce dossier de travail.
- Crée un sous-dossier `data` et y **télécharge les données et les métadonnées**. Les fichiers déjà présents ne sont pas re-téléchargés, pour éviter de perdre du temps et de gaspiller des ressources.
- **Charge dans R** les données et métadonnées dans 3 variables de type data.frame:
  - `log2_LFQ`: quantification en log2(LFQ), où LFQ est la mesure dénommée “Label-Free Quantification”
  - `protein_names`: identifiants, noms de protéines et noms de gènes
  - `sample_groups`: groupe et numéro de patient de chaque échantillon
- **Génère un histogramme** de l’ensemble des valeurs.
- **Génère un rapport préliminaire** qui vous permet de vérifier que les données ont bien été téléchargées sur votre ordinateur et chargées dans R (figure ci-contre).



Téléchargement de la trame de rapport R markdown

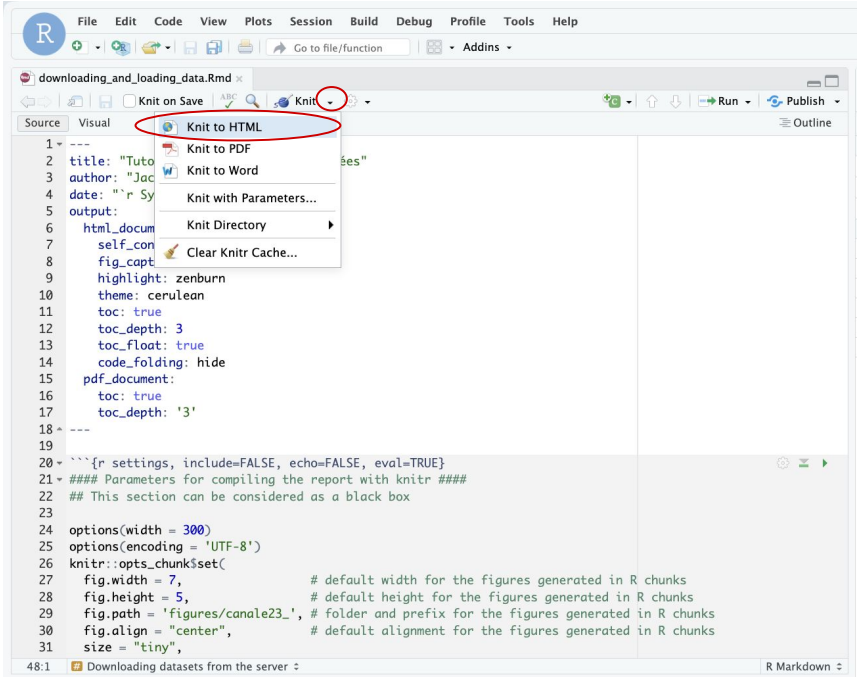
- Sur votre ordinateur, créez un sous-dossier `reports` dans le dossier de travail. Ce sous-dossier devrait avoir le chemin complet suivant.
  `~/proba_stat_bio_TP/canale_2023/reports`
- Téléchargez dans ce sous-dossier `reports` le document `downloading_and_loading_data.Rmd`, soit à partir de votre compte Ametice, soit à partir de github.
  jvanheld.github.io/proba-stat-bio_SBGAU14L/tuto-exo/
- Placez ce document R markdown dans le sous-dossier `reports`.
- Dans l'onglet **Files de RStudio**, ouvrez le sous-dossier `reports` et vérifiez que le document `downloading_and_loading_data.Rmd` s'y trouve bien.



Exécution de la trame de rapport R markdown

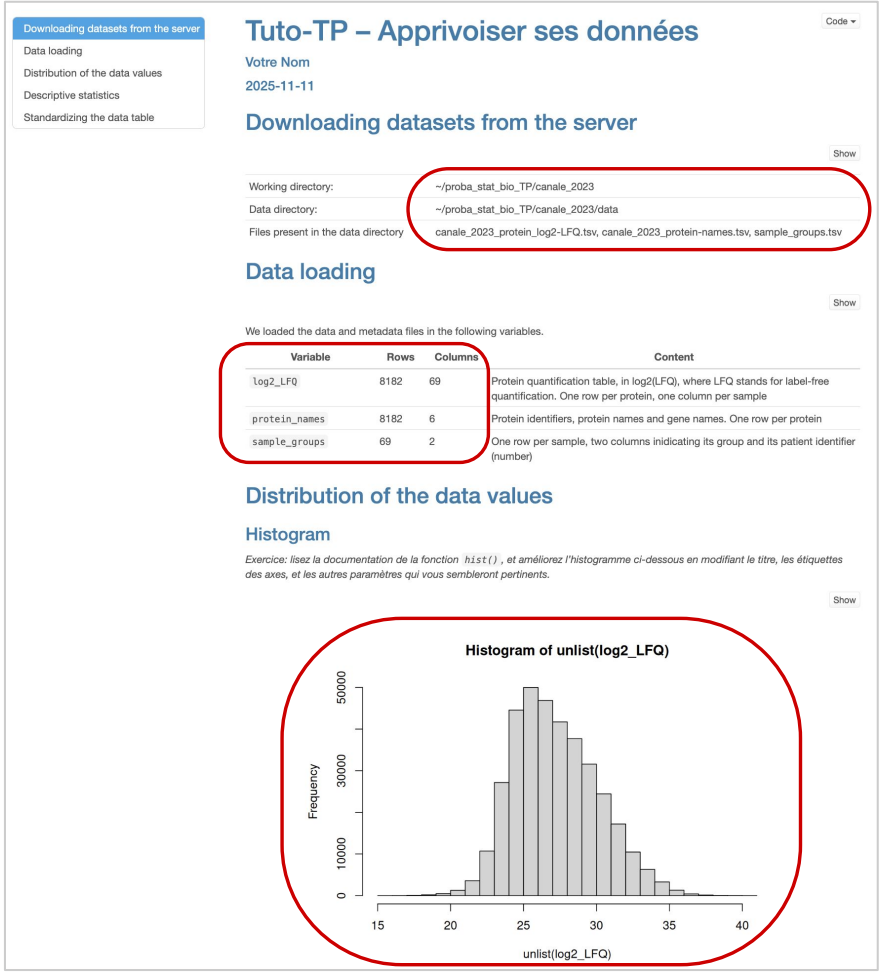
- Dans RStudio, double-cliquez sur le document downloading_and_loading_data.Rmd, pour l'ouvrir dans le panneau d'édition.
- Au-dessus de la fenêtre d'édition, cliquez sur le tout petit triangle à côté de **Knit** et choisissez **Knit to HTML**.
- Après quelques secondes, le rapport devrait s'afficher dans une fenêtre séparée.
- Si vous rencontrez des problèmes, consultez les enseignants.

Astuce: Knit permet de générer le rapport dans différents formats, qu'on peut choisir via ce menu déroulant. Après avoir une première fois choisi un format donné dans le menu déroulant, il vous suffira de cliquer sur Knit pour les prochaines compilations, et RStudio utilisera le même format.



Première ébauche de rapport d'analyse

- Après avoir compilé le rapport HTML avec Knit, vérifiez les éléments suivants
  - Chemin des dossiers
  - Liste des fichiers dans le dossier data
  - Dimensions des tableaux de données (log2_LFQ) et de métadonnées (protein_names et sample_groups).
- Testez les **boutons "Show/Hide"** qui s'affichent au-dessus de chaque segment de code ("chunk").
- Testez le **menu "Code"** au début du rapport.
- Dans l'éditeur de code de RStudio, **adaptez les métadonnées** (format yaml) au début du document Rmd en remplaçant "Votre Nom" par votre nom.



Statistiques descriptives - Réduction des données

Réduction des données

On utilise une série de paramètres pour caractériser certaines propriétés d'un ensemble de données.

- Tendance centrale (moyenne, médiane)
- Dispersion (variance, écart-type, écart interquartiles)
- Dissymétrie
- Aplatissement

Certains paramètres sont basés sur les **moments**, d'autres sur les **quantiles**.

Paramètres basés sur les moments

Moment d'ordre *k* par rapport à *c*

Moyenne des différences entre chaque observation x_i et une constante *c*, élevées à la puissance *k*.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k$$

Moment d'ordre *k* par rapport à l'origine

Moment par rapport à une constante nulle (*c=0*).

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Moment centré d'ordre *k*

Moment par rapport à la moyenne des observations.

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Moyenne

$$\bar{x} = a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Ecart-type (standard deviation)

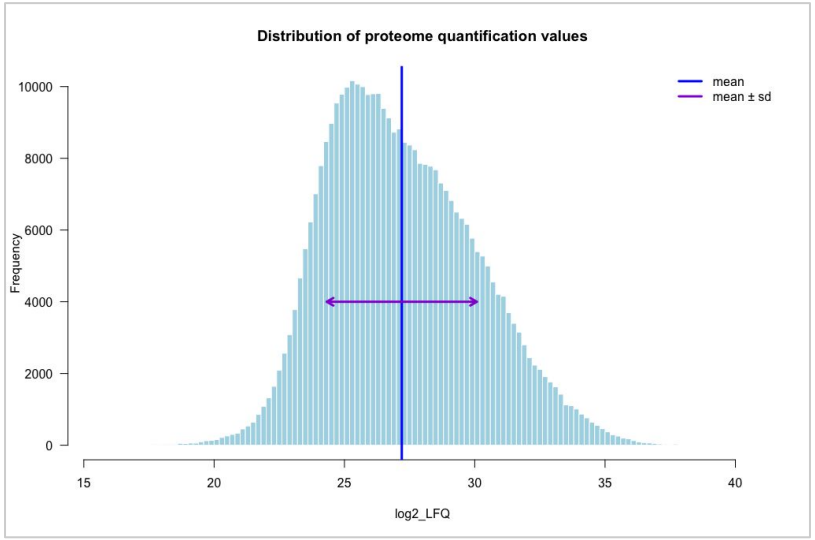
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Dissymétrie (skewness)

$$g_1 = m_3/s^3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$$

Acuité / voussure (*kurtosis, peakedness*)

$$g_2 = m_4/s^4 - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$$



Paramètres basés sur les quantiles

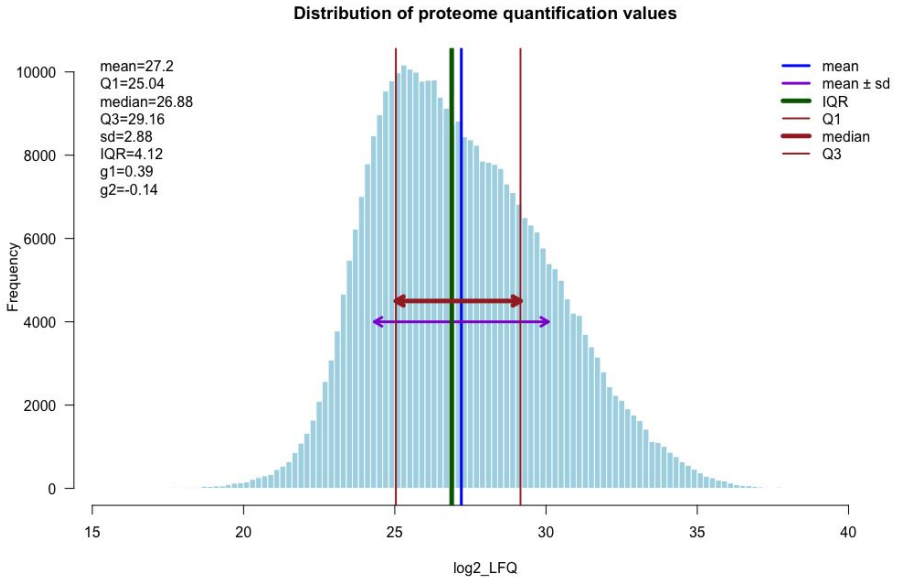
Quantile *Q(q)* : valeur de *x* qui “plafonne” une certaine **proportion *q*** des valeurs observées.

Percentile *P(p)* : valeur de *x* qui “plafonne” un certain **pourcentage *p*** des valeurs observées.

Quartiles Q1, Q2, Q3 : valeurs de *x* qui départagent les valeurs en 4 parts d’effectif égal

Exemples

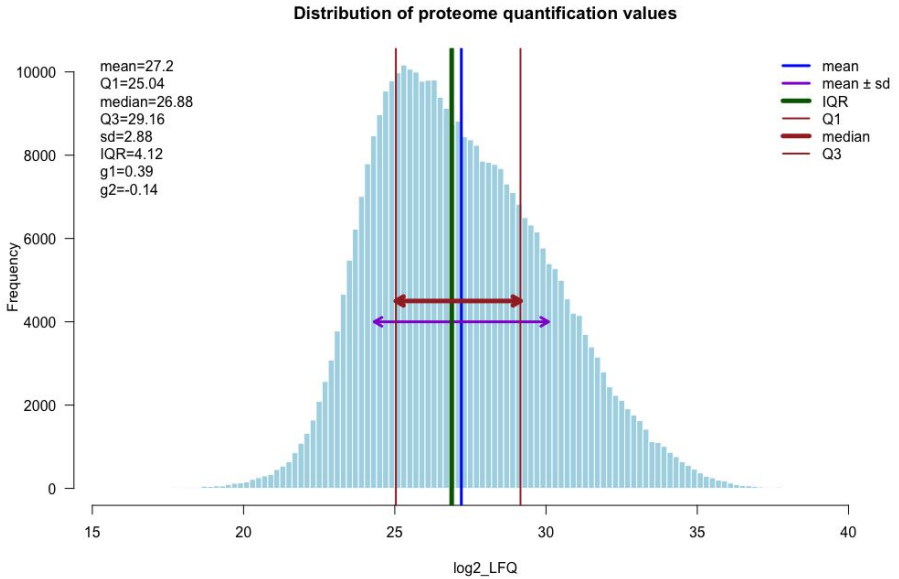
- 5% des valeurs sont ≤ P(05) = Q(0.05)
- 25% des valeurs sont ≤ P(25) = Q(0.25) = Q1
- 50% des valeurs sont ≤ P(50)=Q(0.5) = Q2 = médiane
- 25% des valeurs sont ≤ P(25) = Q(0.75) = Q3



Interprétation des paramètres

- Différence moyenne - médiane: indique généralement une dissymétrie
- g1 > 0 ⟶ dissymétrie vers la droite
- g2 < 0 ⟶ distribution “aplatie” par rapport à une normale.

Pas surprenant, cette distribution n’a aucune raison d’être normale puisqu’elle résulte d’un mélange de mesures d’ échantillons pour >8000 protéines distinctes. A priori, chaque protéine tourne autour d’une concentration spécifique, du fait de la régulation transcriptionnelle et post-transcriptionnelle.

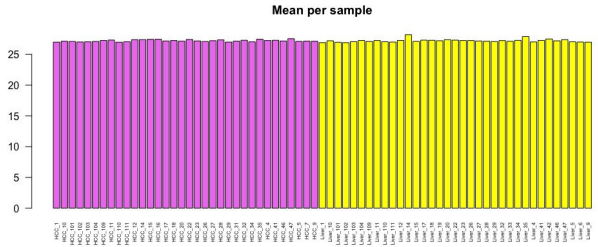


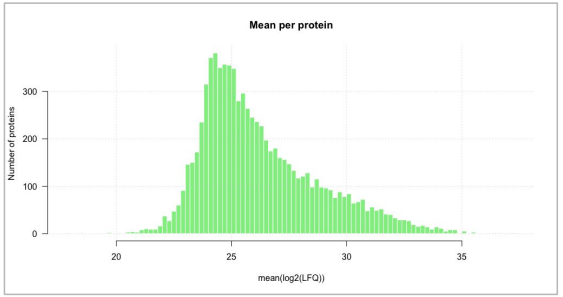
Statistiques marginales

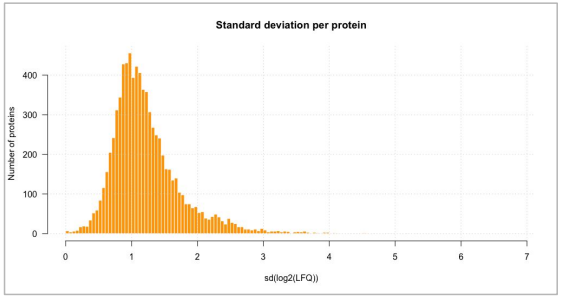
On calcule une statistique (moyenne, écart-type, …) “en marge” de chaque ligne (à droite), ou de chaque colonne (en-dessous).

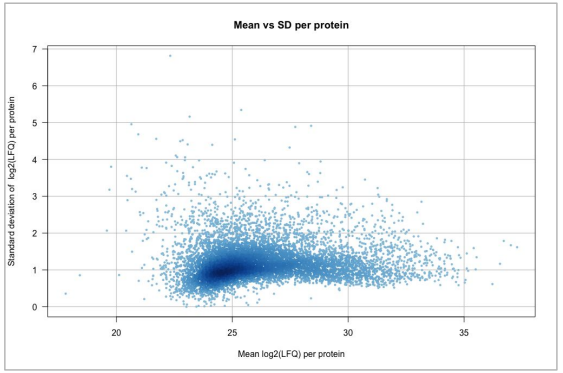
Figure : moyenne (haut) et écart-type (milieu) marginaux par ligne (protéine)des données de Canale (2023). Bas: comparaison entre la moyenne et l’écart-type par ligne. Noter la forte dispersion des points.

Protein	HCC_111	HCC_12	HCC_14	HCC_15	HCC_16	…	mean	sd
P43155-2	31.80	31.80	33.50	27.40	30.10	…	**30.92**	**2.31**
P43304	27.30	28.90	28.10	29.00	29.10	…	**28.48**	**0.77**
P43490	31.70	31.90	32.10	27.20	31.30	…	**30.84**	**2.06**
P45880	32.00	32.70	33.70	31.80	32.50	…	**32.54**	**0.74**
P45954	33.00	33.00	33.90	30.00	32.50	…	**32.48**	**1.48**
P46940	31.10	31.60	31.60	32.10	32.70	…	**31.82**	**0.61**
P47985	30.40	29.60	31.20	29.10	31.00	…	**30.26**	**0.90**
P48047	31.70	31.50	33.50	30.80	31.70	…	**31.84**	**1.00**
P48637	30.70	30.30	29.50	29.60	30.30	…	**30.08**	**0.51**
P48735	33.70	34.40	35.20	31.20	32.50	…	**33.40**	**1.58**
P49189	33.20	32.20	33.90	31.10	32.20	…	**32.52**	**1.07**
…	…	…	…	…	…	…		
mean	**31.51**	**31.63**	**32.38**	**29.94**	**31.45**		**31.38**	
sd	**1.65**	**1.49**	**2.03**	**1.58**	**1.14**			









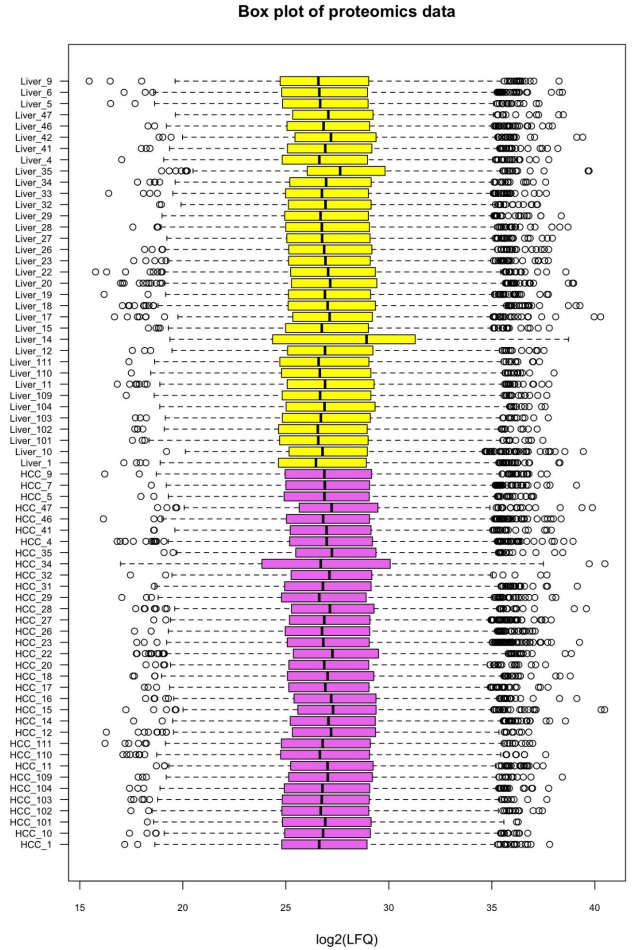
Box plot

Les diagramme de boîtes à moustaches (boxplot) permet d'appréhender d'un seul coup d'oeil la distribution de chaque colonne d'un tableau de données (échantillon dans notre cas).

- Rectangles : espace interquartile
- Barres verticales épaisses : médiane
- Ligne pointillée : intervalle de confiance
- Cercles : valeurs aberrantes ("outliers"), autrement dit observations qui sortent de l'intervalle de confiance

Observations

- Les médianes ne sont pas spécialement alignées
- Les différents échantillons occupent des zones de valeurs similaires
- On note que 2 échantillons ont des espaces inter-quartiles plus étendus que les autres (HCC_34 et Liver_14).
- Les valeurs aberrantes sont plus nombreuses à droite qu'à gauche.

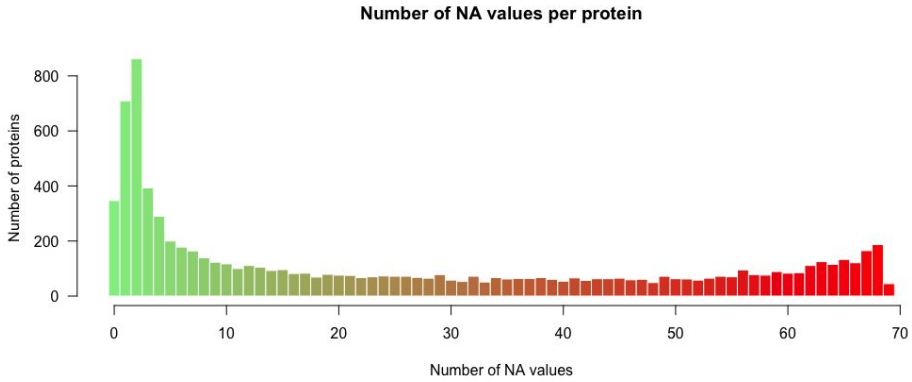


Filtrage des données trop incomplètes

Nombre de valeurs manquantes (*NA*) par protéine (ligne)

Nombre de valeurs NA par protéine (ligne).

Noter que certaines protéines ont des valeurs manquantes pour *tous* les échantillons (en rouge, à l'extrême droite du barplot).

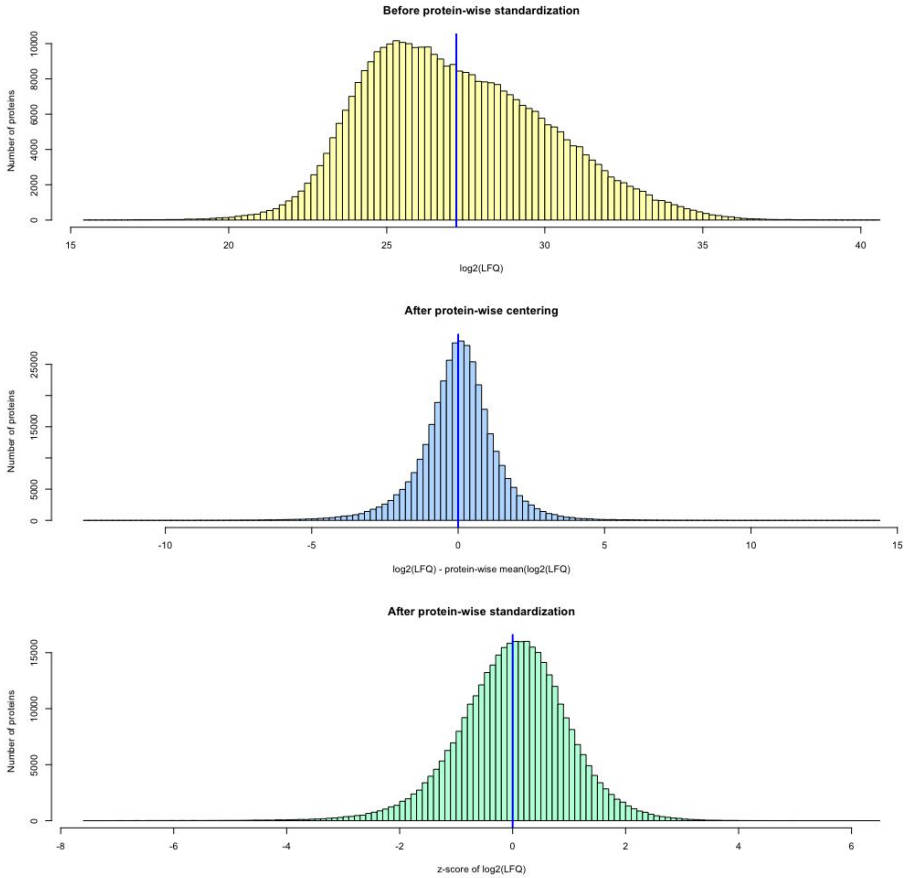


Data standardization

Standardisation par ligne (protéine)

Centrage (centring) : on soustrait de chaque valeur la moyenne de sa ligne

Mise à l'échelle (scaling) : on divise la



Exercices supplémentaires pour ceux qui ont fini tôt

Sélection de protéines d'intérêt

Dans un premier temps nous nous intéresserons aux protéines produites par une poignée de gènes dont l'étude de Canale a révélé l'intérêt.

Voici les noms de gènes correspondant à ces protéines

- PYCR2
- ITGA6
- PBXIP1
- LEPRE1
- UROC1
- UGP2
- AGL
- ASS1
- PYGL

Dans un premier temps, nous analyserons les profils protéomiques de ces gènes ciblés, et nous effectuerons ensuite des analyses de l'ensemble du tableau pour appréhender son contenu.

Exercice

A l'aide de la fonction R `c()`, créez un vecteur qui contient les noms des gènes d'intérêt.

Affichez avec `View()` le contenu de la table `protein_descr`, et utilisez le filtre pour trouver les identifiants protéiques (noms de lignes) des gènes **PYCR2** et **UROC1**.