



Semaine 3 – Du catalogue à l'échantillon analysable

Outils informatiques pour les géosciences II

Sébastien Biass 

sebastien.biasse@unige.ch

1st floor, office 103

23 octobre 2026



Statistiques univariées



Une variable à la fois

En sciences de la Terre, des **variables** peuvent être:

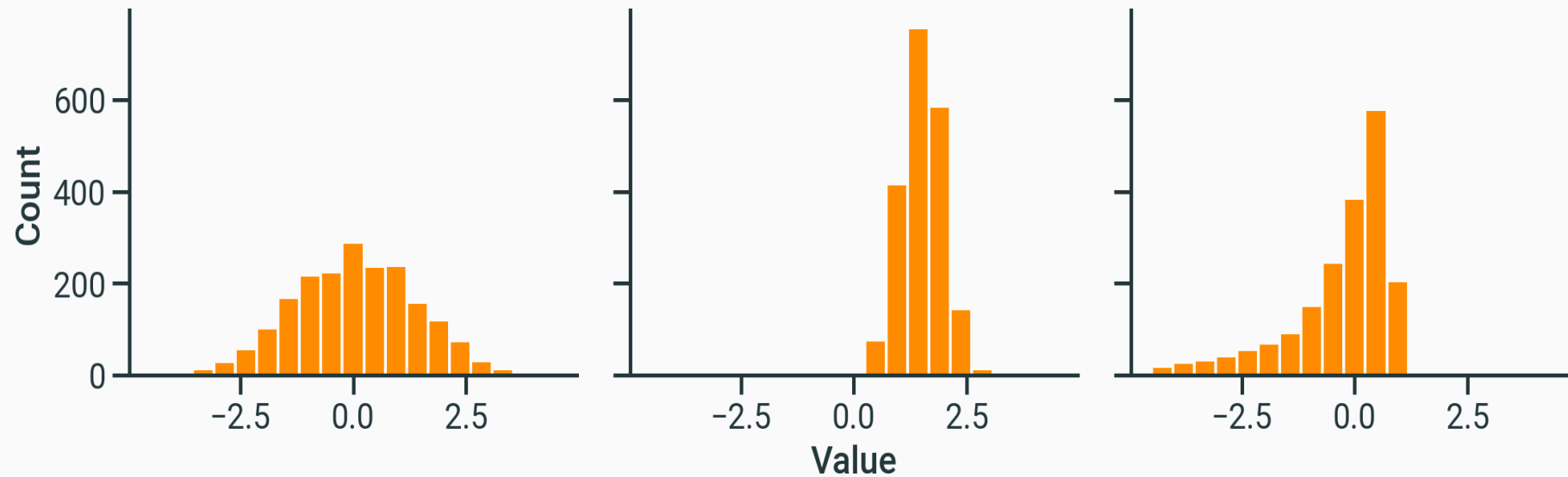
- le volume d'un glissement de terrain
- la taille des grains d'un sédiment
- la concentration d'un élément
- l'âge d'un cristal
- l'épaisseur d'une unité
- le débit d'une rivière

Les **statistiques univariées** décrivent **une seule variable** à la fois. Le but n'est pas encore **d'expliquer les relations entre plusieurs propriétés**, mais de comprendre la **distribution** d'une mesure.



Décrire une distribution

Propriété	Question	Exemples de résumés
Position	Où se situent les valeurs?	moyenne, médiane, quantiles
Dispersion	À quel point diffèrent-elles?	étendue, écart-type, IQR
Forme	Comment se répartissent-elles?	asymétrie, modes, valeurs extrêmes



Échantillon et population

La population

La **population** est l'ensemble complet des objets auxquels la question scientifique s'intéresse.

| *Exemple: tous les glissements de terrain capables de déclencher un tsunami.*

Cette population est généralement inaccessible:

- certains événements n'ont pas été observés
- certains événements anciens n'ont laissé aucune mesure
- certaines régions sont mieux documentées que d'autres

L'échantillon

L'**échantillon** est l'ensemble limité des observations disponibles utilisées dans l'analyse.

Exemple: les glissements ayant déclenché un tsunami qui ont été documentés, validés et dont le volume est disponible.

L'échantillon dépend:

- de la manière dont les événements ont été documentés
- des variables disponibles
- des critères appliqués pendant la préparation

D'autres exemples

Echantillon

Epaisseurs mesurées dans un **nombre limité d'affleurements**, de forages ou de profils géophysiques

Concentrations en arsenic mesurées dans l'eau prélevée dans un **nombre limité de puits**

Granulométrie dans prélèvements réalisés à **quelques endroits**

Population

→ Permet d'estimer **l'épaisseur moyenne** de l'unité entière

→ Estimer la **concentration** typique et sa variabilité dans l'aquifère

→ Estimer la **granulométrie** de l'unité

De l'échantillon à la population

Echantillon

= Observations disponibles

Statistiques *empiriques* calculées:

- moyenne $\bar{x}$
- écart-type s
- médiane et quantiles

Population

= Ensemble visé par la question

Paramètres recherchés:

- moyenne μ
- écart-type σ
- distribution complète

L'**inférence** évalue ce que l'échantillon permet d'estimer sur la population, avec quelle incertitude et sous quelles hypothèses.

Le bloc des semaines 3–5

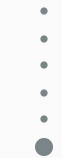
Semaine	Étape	Question principale
3	Préparation des données	Quelles observations pouvons-nous analyser?
4	Description et visualisation	Comment les valeurs sont-elles distribuées?
5	Inférence	Que suggère l'échantillon sur la population?

Le bloc des semaines 3–5

Semaine	Étape	Question principale
3	Préparation des données	Quelles observations pouvons-nous analyser?
4	Description et visualisation	Comment les valeurs sont-elles distribuées?
5	Inférence	Que suggère l'échantillon sur la population?

Volume d'un glissement de terrain

1. **Préparer**: Quels volumes sont disponibles et admissibles?
2. **Décrire**: Comment les volume sont-ils distribués?
3. **Inférer**: Une différence observée entre deux mécanismes peut-elle être généralisée au-delà des observations disponibles?



Le catalogue étudié

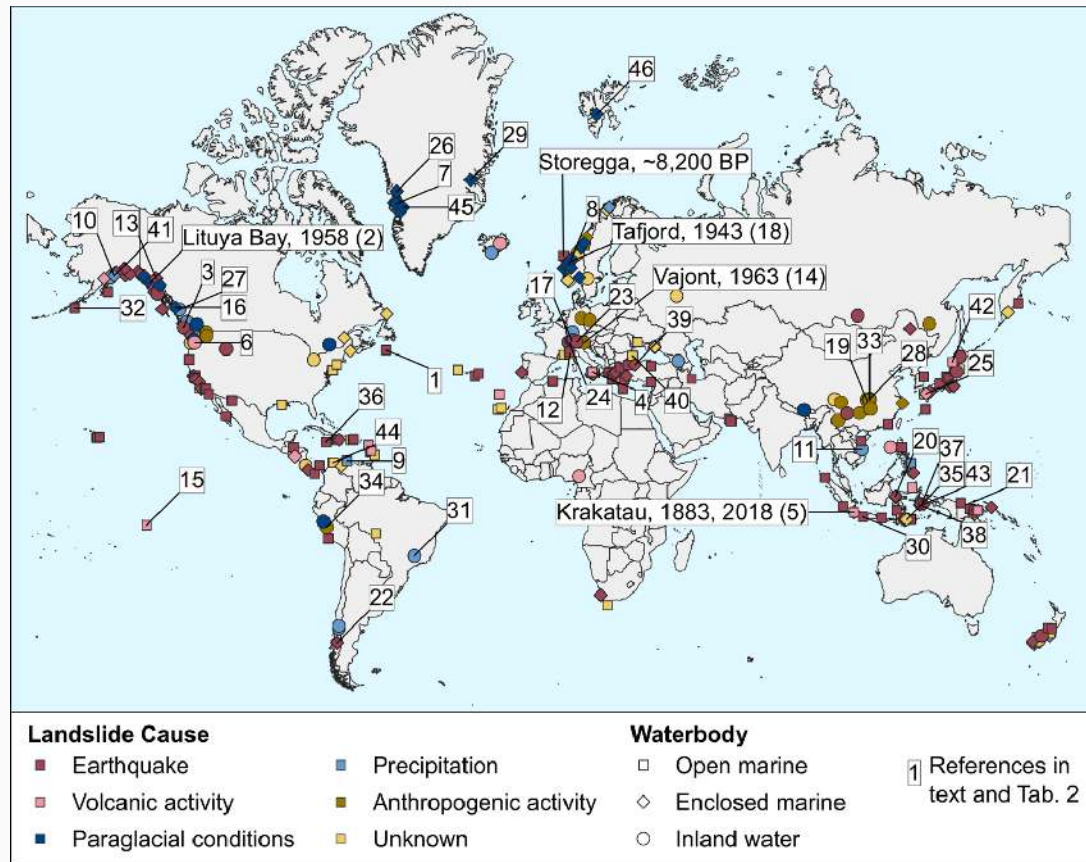


Tsunamis générés par un glissement

Un **glissement** qui entre dans l'eau peut **déplacer brutalement un volume d'eau** et produire un **tsunami**.

- Environ **10 % des tsunamis observés** ont été déclenchés par des glissements, seuls ou combinés à d'autres mécanismes.
- Ils peuvent produire des vagues **locales extrêmement hautes**.
- La proximité entre la source et la côte **limite le temps d'alerte**.
- Leur localisation et leurs caractéristiques restent **incertaines**.

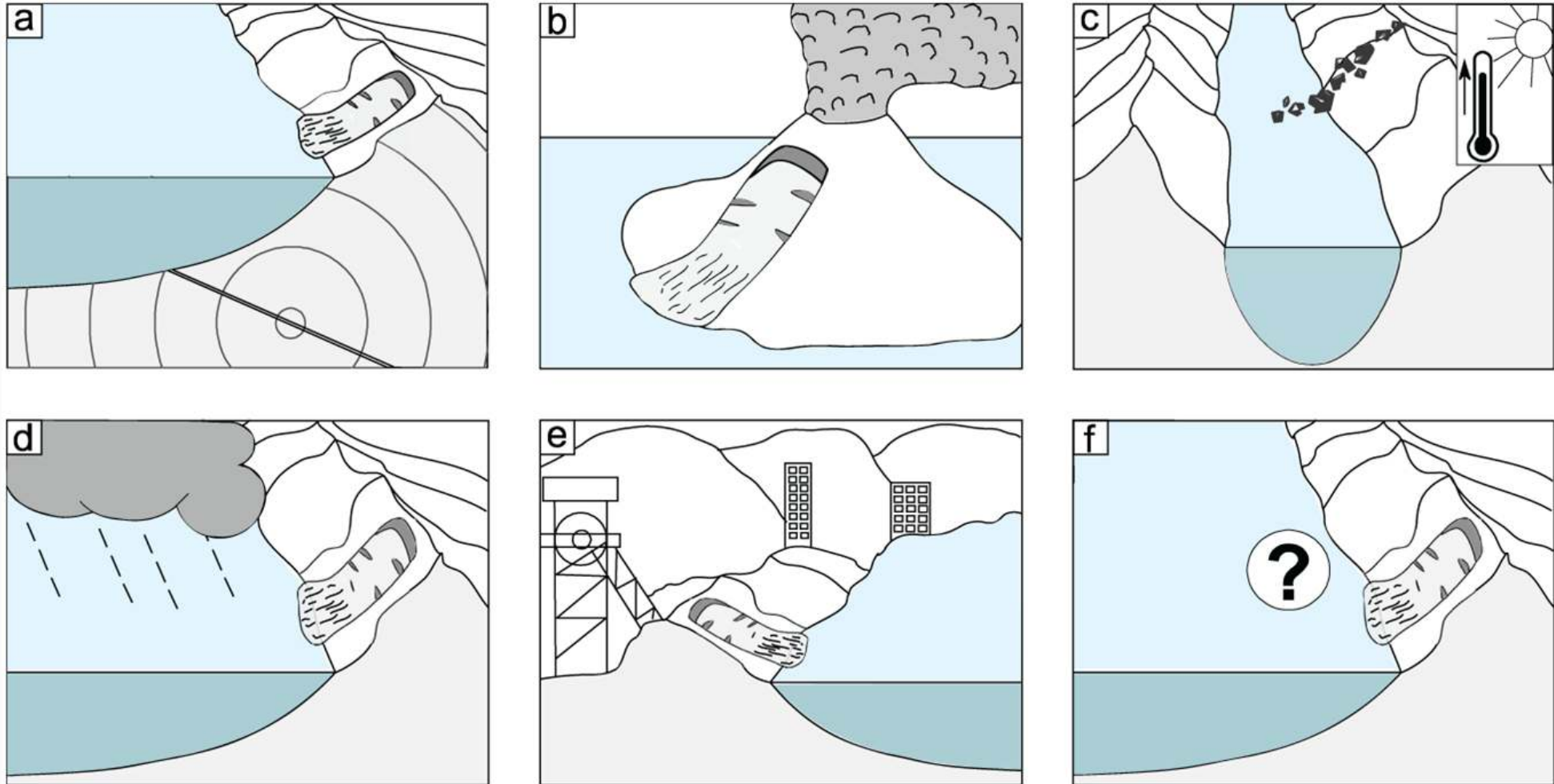
Un catalogue global



Distribution des événements (Dohmen et al. 2025)

- Recensement de **tous les cas documentés** jusqu'au 30 octobre 2024
- **355 entrées** issues de sources historiques et scientifiques
- **38 entrées exclues**: origine incorrecte, preuves insuffisantes ou mécanisme réfuté
- **317 glissements** considérés comme valides
- **297 tsunamis distincts** associés à ces glissements

Six catégories de causes



a) séisme, b) activité volcanique, c) conditions paraglaciaires, d) précipitations, e) activité anthropique, f) cause inconnue ([Dohmen et al. 2025](#))

Structure du catalogue

Données

Famille	Exemples	Rôle
Identifiants	landslide_id, tsunami_id	relier une ligne au phénomène étudié
Catégories	cause, type, plan d'eau	décrire le mécanisme et le contexte
Mesures	volume, hauteur d'eau, distance	quantifier le glissement ou le tsunami
Impacts	décès, dommages	documenter les conséquences



Structure du catalogue

Données

Famille	Exemples	Rôle
Identifiants	landslide_id, tsunami_id	relier une ligne au phénomène étudié
Catégories	cause, type, plan d'eau	décrire le mécanisme et le contexte
Mesures	volume, hauteur d'eau, distance	quantifier le glissement ou le tsunami
Impacts	décès, dommages	documenter les conséquences

Nombre de lignes, de glissements et de tsunamis

Événement	landslide_id	tsunami_id	Pays
Seram Elpaputi Bay (1899)	95	95	Indonésie
Seram Kawa (1899)	96	95	Indonésie
Seram Piru Bay (1899)	97	95	Indonésie
Seram Teluti Bay (1899)	98	95	Indonésie



Disponibilité

Les variables (*colonnes*) varient en **complétude**:

- localisation du glissement: **89 %**
- hauteur d'eau maximale: **84 %**
- type de glissement: **57 %**
- matériau: **41 %**
- volume: **35 %**

! Une **valeur manquante** signifie que l'information **n'est pas disponible**. Elle ne signifie pas que le phénomène est absent ou que sa valeur est nulle.

Disponibilité

Les variables (*colonnes*) varient en **complétude**:

- localisation du glissement: **89 %**
- hauteur d'eau maximale: **84 %**
- type de glissement: **57 %**
- matériau: **41 %**
- volume: **35 %**

! Une **valeur manquante** signifie que l'information **n'est pas disponible**. Elle ne signifie pas que le phénomène est absent ou que sa valeur est nulle.

Echantillon du cours

1. les entrées considérées comme **valides**
2. les glissements dont le **volume est renseigné** et **strictement positifs**

Aujourd'hui

À vous

1. Créez un notebook `semaine_03.ipynb` dans le dossier `outils_info`
2. Téléchargez les données et placez-les dans le dossier `outils_info/data`
3. Suivre [Semaine 3](#) depuis le site
 - Copier/coller le code fournit (tout en expérimentant!)
 - Répondez aux [questions](#)
4. Utilisez les [cheat sheets](#) et la [documentation simplifiée](#) pour expérimenter et découvrir de nouvelles fonctionnalités
5. Répondez à la [Mini investigation](#) et documentez votre code avec Markdown
6. Sauvegardez la table préparée pour la semaine 4

Question fil rouge

La question *technique* est:

- *Comment préparer et clarifier la structure d'un jeu de données pour en préparer l'analyse?*

La question *scientifique* est:

- *Quelle est la complétude d'enregistrement des volumes associés aux glissements de terrain dans le catalogue?*



À la fin du notebook, nous devons pouvoir expliquer:

- ce que représente une ligne / quelle unité répond à chaque question
- quelles observations ont été exclues / combien d'observations restent disponibles
- quelle population nos résultats peuvent décrire

Objectifs

À la fin de l'exercice, vous devriez pouvoir:

- ✓ distinguer population et échantillon
- ✓ identifier ce que représentent les lignes et les colonnes
- ✓ compter des objets distincts et des catégories
- ✓ choisir une table adaptée à l'unité d'analyse
- ✓ évaluer la complétude d'une variable
- ✓ construire et documenter un échantillon analysable



References

Dohmen K, Blum P, Braun A, Fernandez-Steege TM (2025) Landslide-Triggered Tsunamis – a Review. Natural Hazards 121:22341-22373. <https://doi.org/10.1007/s11069-025-07551-2>

