

Analyse EAH

Accès à l'eau, qualité et fonctionnalité des points d'eau au regard des échelles de service du JMP et des indicateurs Sphère, y compris le résiduel de chlore que personne ne consigne régulièrement.

Formateur	Pierre Robentz Cassion
Niveau	Intermédiaire
Heures apprenant	16 h
Enseignée en	Python · R
Mise à jour	28 juillet 2026
Secteurs	EAH
Normes et méthodologies	Standards Sphère · Objectifs de développement durable (ODD) · Norme humanitaire fondamentale (CHS)

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/wash-analysis

Le contenu pédagogique est diffusé sous licence CC BY 4.0. Vous pouvez l'adapter et le rediffuser, y compris à des fins commerciales, en créditant Cassion.

Ce que vous saurez faire

- Placer un ménage sur les trois échelles de service du JMP, et dire pourquoi une source améliorée n'équivaut pas à un service élémentaire
- Calculer quantité, temps de collecte et attente au regard des standards minimaux Sphère, avec les erreurs d'unité que portent ces champs
- Rapporter la qualité de l'eau depuis un sous-échantillon testé sans lui laisser emprunter le dénominateur de l'accès
- Traiter une mesure censurée comme une donnée plutôt que de la supprimer ou de la convertir en zéro
- Produire trois taux de fonctionnalité défendables à partir d'un seul registre de suivi et dire de quoi chacun est l'énoncé
- Distinguer un point d'eau qui se tarit chaque saison sèche d'un point abandonné, à partir de la séquence des visites

Prérequis

Aucun. Cette formation ne suppose aucune expérience préalable de la programmation.

Sommaire

I	L'échelle n'est pas la source	1
1	Quatre cinquièmes améliorés, la moitié servis	2
1.1	Le nombre que donnent la plupart des rapports EAH	2
1.2	L'échelle que l'indicateur ODD emploie réellement	2
1.3	Pourquoi les trente minutes figurent dans la définition	3
1.4	L'échelon que cette enquête ne peut pas calculer	3
1.5	Rapportez la distribution, non l'échelon supérieur	4
1.6	La suite	4
2	Le booléen qui divise une couverture par deux	5
2.1	L'assainissement a la même forme que l'eau	5
2.2	Pourquoi le partage coûte un échelon	6
2.3	La défécation à l'air libre est le nombre à mettre en tête	6
2.4	L'échelle de l'hygiène, et l'observation qui la porte	6
2.5	Les trois échelles ne se moyennent pas	7
2.6	La suite	7
II	Ce que Sphère demande et qu'une échelle ne dit pas	8
3	Quinze litres, trente minutes, cinq cents mètres	9
3.1	Ce que Sphère ajoute à l'échelle	9
3.2	Deux erreurs d'unité, et une seule est visible	9
3.3	L'attente est incluse dans l'aller-retour, et c'est un problème	10
3.4	La distance n'est pas du tout dans le fichier	11
3.5	Rapportez la distribution face au seuil	11
3.6	La suite	11
4	Testé sur un tiers, censuré sur un cinquième	12
4.1	Le dénominateur change, et le rapport généralement non	12
4.2	Les classes de risque du JMP	12
4.3	Amélioré n'est pas sûr, et c'est ici que cela se démontre	13
4.4	La valeur censurée, et pourquoi elle n'est pas manquante	13
4.5	Que faire d'une valeur censurée	14
4.6	L'incohérence entre champs qu'il faut nommer	15
4.7	Rapportez-le comme un bloc à part	15
4.8	La suite	15
III	Le service est une année, non une journée	16
5	74,4 %, 33,9 %, 83,3 %	17
5.1	Ce que rend possible un registre à visites répétées	17
5.2	Taux un : la part des visites ayant trouvé un point en service	17

5.3	Taux deux : la part des points en service à chaque fois	18
5.4	Taux trois : la part des personnes disposant d'un point en service	18
5.5	Lequel rapporter	19
5.6	Les deux points comptés deux fois	19
5.7	La suite	20
6	Tari n'est pas en panne	21
6.1	Un seul état, plusieurs défaillances différentes	21
6.2	Le signal saisonnier, et là où il n'est pas	21
6.3	Classer chaque point à partir de sa propre séquence	22
6.4	La valeur d'état qui ressemble à la réponse sans l'être	23
6.5	La durée d'arrêt est une statistique de gestion	23
6.6	Rapportez les formes, non le seul taux	24
6.7	La suite	24
IV	Ce que le suivi n'a pas vu	25
7	La tournée que personne n'a conduite	26
7.1	Le dénominateur que personne n'a interrogé	26
7.2	La couverture n'est pas répartie uniformément, et c'est tout le problème . . .	26
7.3	Encadrez la réponse	27
7.4	L'absence est-elle aléatoire ?	28
7.5	Ce qu'il ne faut pas faire	28
7.6	Rapportez la couverture à côté de chaque taux	29
7.7	La suite	29
8	Onze nombres et six dénominateurs	30
8.1	Comptez les dénominateurs avant d'écrire quoi que ce soit	30
8.2	Le rapport	31
8.3	Quatre règles auxquelles la mise en page doit obéir	31
8.4	À quoi sert chaque nombre	32
8.5	Ce que ce cours n'a pas couvert	32
8.6	La suite	32

À propos de cette formation

L'EAH est le domaine où l'instruction permanente de cette plateforme — énoncer le numérateur, le dénominateur et la décision — rencontre un secteur qui dispose de définitions publiées pour presque tout et d'une habitude de les rapporter approximativement.

Le cours s'appuie sur deux fichiers. L'**enquête ménage EAH** est apparue dans trois cours antérieurs comme problème de nettoyage puis de dénominateur ; elle est ici enfin analysée pour ce qu'elle est, une enquête de niveau de service. Le **registre de suivi des points d'eau** est nouveau, et c'est le seul fichier de la plateforme à comporter des visites répétées sur le même ouvrage — ce qu'exige un taux de fonctionnalité et ce qu'une enquête ménage ne peut structurellement pas fournir.

Trois résultats structurent le cours, tous calculés à partir de ces fichiers.

Quatre cinquièmes des ménages sont sur une source améliorée et à peine la moitié disposent d'un service élémentaire. 82,3 % contre 56,3 %, et les vingt-six points d'écart sont du temps de collecte : une source améliorée à plus de trente minutes est un service *limité*, et une analyse qui classe sur le seul type de source rapportera le premier nombre en l'appelant l'accès.

La fonctionnalité vaut trois nombres, non un. 74,4 % des visites de suivi trouvent un point en service ; 33,9 % des points fonctionnent à chacune de leurs visites ; 83,3 % des personnes desservies ont un point qui fonctionne. Les trois sont exacts et énoncent des choses différentes.

Le district qui paraît le meilleur porte le nombre le moins fiable. Nord-Ouest affiche 76,4 %, le plus élevé des trois, et n'a réalisé que 82,7 % de ses visites dues parce que les routes se ferment avec les pluies. Si chaque visite manquée avait trouvé un point en panne, il serait à 63,2 % — une fourchette large de 13,2 points, contre 2,6 pour le district qui a presque tout visité.

Python et R tout du long. Vous devriez avoir suivi le module 3 — ce cours suppose que vous savez défendre un dénominateur, lire un chiffre de routine à travers son taux de rapportage, et écrire une fiche de référence d'indicateur avant l'analyse.

Première partie

L'échelle n'est pas la source

CHAPITRE 1

Quatre cinquièmes améliorés, la moitié servis

Leçon 1 sur 8 · 120 min

82,3 % de ces ménages sont sur une source améliorée et 56,3 % disposent au moins d'un service élémentaire d'eau de boisson. Les vingt-six points d'écart sont du temps de collecte, et ils sont toute la raison d'être des échelons.

1.1 Le nombre que donnent la plupart des rapports EAH

Compter les ménages sur une source améliorée, diviser par les ménages enquêtés, et publier le résultat comme l'accès.

```
import pandas as pd

households = pd.read_csv("wash-household-survey-2024.v1.csv")

IMPROVED = {
    # the JMP list, delivered water included
    "piped-into-dwelling", "piped-into-yard", "public-tap",
    "borehole", "protected-well", "protected-spring", "tanker-truck",
}

improved = households["water_source"].isin(IMPROVED)
print(f"improved source: {improved.mean():.1%} of {len(households)} households")

library(dplyr)

improved <- c("piped-into-dwelling", "piped-into-yard", "public-tap",
             "borehole", "protected-well", "protected-spring",
             "tanker-truck")

households |> summarise(improved = mean(water_source %in% improved), n = n())
```

82,3 %. C'est le calcul exact d'une chose que personne n'a demandée.

1.2 L'échelle que l'indicateur ODD emploie réellement

L'échelle du JMP pour l'eau de boisson compte quatre échelons, et le type de source ne décide que des deux entre lesquels vous choisissez.

Échelon	Définition
Géré en toute sécurité	Source améliorée sur place, disponible en cas de besoin, et exempte de contamination
Élémentaire	Source améliorée, aller-retour de 30 minutes ou moins, attente comprise
Limité	Source améliorée, aller-retour de plus de 30 minutes
Non amélioré	Puits ou source non protégés
Eau de surface	Rivière, retenue, lac, mare, canal, canal d'irrigation

L'indicateur ODD 6.1.1 porte sur l'échelon **géré en toute sécurité**. La ligne de rapportage humanitaire, celle sur laquelle la plupart des programmes sont jugés, est **au moins élémentaire** — géré en toute sécurité plus élémentaire.

```
def ladder(row):
    if row["water_source"] == "surface-water":
        return "surface water"
    if row["water_source"] not in IMPROVED:
        return "unimproved"
    minutes = row["round_trip_minutes"]
    return "basic" if minutes <= 30 else "limited"

households["service"] = households.apply(ladder, axis=1)
print(households["service"].value_counts(normalize=True).round(3))
```

```
households |>
  mutate(service = case_when(
    water_source == "surface-water" ~ "surface water",
    !water_source %in% improved ~ "unimproved",
    round_trip_minutes <= 30 ~ "basic",
    TRUE ~ "limited"
  )) |>
  count(service) |>
  mutate(share = n / sum(n))
```

Échelon	Ménages	Part
Élémentaire (sur place compris)	1 352	56,3 %
Limité	626	26,1 %
Non amélioré	326	13,6 %
Eau de surface	99	4,1 %

Service d'eau de boisson au moins élémentaire : 56,3 %. Le chiffre par type de source valait 82,3 %, et chacun des vingt-six points d'écart est un ménage sur une source protégée qu'il faut plus d'une demi-heure pour atteindre et quitter.

1.3 Pourquoi les trente minutes figurent dans la définition

Ce n'est pas de la rigueur administrative. Un aller-retour de plus de trente minutes modifie les comportements de façons mesurables ailleurs dans ce même fichier :

- les ménages collectent **moins** d'eau, si bien que la quantité passe plus souvent sous le minimum Sphère ;
- la collecte est faite par les femmes et les filles, si bien que le temps est pris sur la scolarité et le travail rémunéré ;
- les ménages complètent depuis une source plus proche et de moindre qualité, si bien que la source enregistrée n'est pas la seule.

Une définition qui paraît bureaucratique encode généralement un constat. Avant de soutenir qu'un seuil est arbitraire, cherchez ce qu'il a été fixé pour séparer.

1.4 L'échelon que cette enquête ne peut pas calculer

« Géré en toute sécurité » exige trois conditions et ce fichier n'en documente qu'une.

```
on_premises = (households["round_trip_minutes"] == 0)
tested = households["ecoli_cfu_100ml"].notna()
```

```
print(f"on premises: {on_premises.mean():.1%}")
print(f"water quality tested: {tested.mean():.1%}")
print(f"both: {(on_premises & tested).mean():.1%}")
```

```
households |> summarise(
  on_premises = mean(round_trip_minutes == 0),
  tested = mean(!is.na(ecoli_cfu_100ml))
)
```

27,7 % des ménages ont l'eau sur place. La disponibilité en cas de besoin n'est pas posée. L'absence de contamination est testée sur un tiers de l'échantillon. Donc « **géré en toute sécurité** » **n'est pas calculable ici, et le rapport honnête le dit** plutôt que de rapporter « sur place » comme s'il s'agissait de l'échelon supérieur.

C'est la façon la plus courante dont un chiffre ODD se trouve surestimé : une des trois conditions est mesurée, et l'indicateur est publié sous le nom des trois.

1.5 Rapportez la distribution, non l'échelon supérieur

Drinking water service level, 2,403 households

At least basic	56.3%	1,352
of which on premises	27.7%	665
Limited	26.1%	626
Unimproved	13.6%	326
Surface water	4.1%	99

JMP ladder. Safely managed not computable: availability was not asked and water quality was tested on 33.4% of households.

Chaque échelon, avec les effectifs. Un rapport qui ne donne que « 56,3 % ont un service élémentaire » masque qu'un quart de la population est à un seuil de distance et basculerait avec un point d'eau plus proche — c'est précisément le groupe sur lequel un programme peut agir.

1.6 La suite

L'eau a une échelle et il en reste deux. La leçon suivante traite l'assainissement et l'hygiène, où l'équivalent de la règle des trente minutes est une unique colonne booléenne qui divise discrètement un taux de couverture par deux.

CHAPITRE 2

Le booléen qui divise une couverture par deux

Leçon 2 sur 8 · 120 min

62,0 % des ménages ont un ouvrage d'assainissement amélioré et 43,2 % un service élémentaire. L'écart tient à une colonne — la latrine est-elle partagée — et l'échelle de l'hygiène a un équivalent qui lui coûte encore la moitié.

2.1 L'assainissement a la même forme que l'eau

Un ouvrage amélioré est nécessaire et non suffisant, et le champ qui fait la différence est un booléen que personne ne regarde deux fois.

Échelon	Définition
Géré en toute sécurité	Amélioré, non partagé, et excréta évacués ou traités sans risque
Élémentaire	Amélioré, non partagé avec d'autres ménages
Limité	Amélioré mais partagé
Non amélioré	Latrine à fosse sans dalle, latrine suspendue, seau
Défécation à l'air libre	Aucun ouvrage

```
IMPROVED_SANITATION = {  
  "flush-to-sewer", "flush-to-septic", "vip-latrine", "pit-latrine-with-slab",  
}
```

```
improved = households["sanitation_facility"].isin(IMPROVED_SANITATION)  
shared = households["shared_sanitation"]
```

```
print(f"improved facility: {improved.mean():.1%}")  
print(f"basic (not shared): {(improved & ~shared).mean():.1%}")  
print(f"limited (shared): {(improved & shared).mean():.1%}")
```

```
improved_sanitation <- c("flush-to-sewer", "flush-to-septic",  
  "vip-latrine", "pit-latrine-with-slab")
```

```
households |>  
  mutate(service = case_when(  
    sanitation_facility == "open-defecation" ~ "open defecation",  
    !sanitation_facility %in% improved_sanitation ~ "unimproved",  
    shared_sanitation ~ "limited",  
    TRUE ~ "basic"  
  )) |>  
  count(service) |>  
  mutate(share = n / sum(n))
```

Échelon	Ménages	Part
Élémentaire	1 038	43,2 %
Limité (partagé)	452	18,8 %
Non amélioré	594	24,7 %
Défécation à l'air libre	319	13,3 %

62,0 % ont un ouvrage amélioré et 43,2 % un service élémentaire. Dix-huit points reposent sur une latrine conforme au standard de construction et partagée avec un autre ménage.

2.2 Pourquoi le partage coûte un échelon

Parce que la raison d'être du standard est l'usage, non la construction. Une latrine partagée est moins utilisée la nuit, moins par les femmes et les filles, et moins par les enfants — c'est-à-dire là où se situe le bénéfice sanitaire. La classification suit le comportement, non le béton.

C'est le même argument que les trente minutes. Dans les deux échelles, la seconde condition existe parce qu'un ouvrage techniquement adéquat et pratiquement évité ne produit aucun résultat sanitaire, et qu'un indicateur qui le compterait rapporterait un progrès que personne n'a vécu.

2.3 La défécation à l'air libre est le nombre à mettre en tête

```
od = households["sanitation_facility"].eq("open-defecation")
by_district = households.groupby("district")["sanitation_facility"].apply(
    lambda s: s.eq("open-defecation").mean()
)
print(f"{od.sum()} households, {od.mean():.1%}")
print((by_district * 100).round(1))
```

```
households |>
  summarise(open_defecation = mean(sanitation_facility == "open-defecation"),
            n = n(), .by = district)
```

13,3 %, et c'est le seul chiffre EAH qu'un non-spécialiste lit correctement sans qu'on lui donne la définition. Mettez-le en tête, désagrégez-le, et notez que le champ district de ce fichier est écrit de six façons — le cours sur le nettoyage s'en est occupé, et grouper sans normaliser coupe le pire district en quatre.

2.4 L'échelle de l'hygiène, et l'observation qui la porte

Le service d'hygiène élémentaire exige **un dispositif de lavage des mains avec eau et savon disponibles**. Trois échelons, et l'enquête répond aux trois, parce que l'enquêteur a observé au lieu de demander.

```
has_facility = households["handwashing_facility"].ne("no-facility")
soap = households["soap_observed"]

print(f"any facility: {has_facility.mean():.1%}")
print(f"basic (facility + soap): {(has_facility & soap).mean():.1%}")
print(f"limited (no soap): {(has_facility & ~soap).mean():.1%}")
print(f"no facility: {(~has_facility).mean():.1%}")

households |>
  mutate(hygiene = case_when(
    handwashing_facility == "no-facility" ~ "no facility",
    soap_observed ~ "basic",
```

```

    TRUE ~ "limited"
  )) |>
  count(hygiene) |>
  mutate(share = n / sum(n))

```

Échelon	Ménages	Part
Élémentaire — dispositif et savon	819	34,1 %
Limité — dispositif, sans savon	811	33,7 %
Aucun dispositif	773	32,2 %

67,8 % ont un endroit où se laver les mains et 34,1 % y ont du savon. L'échelle de l'hygiène perd la moitié de son échelon supérieur sur un seul article observé, ce qui est une chute plus forte que dans les deux autres échelles.

C'est l'observation qui rend ce chiffre fiable. Le lavage des mains déclaré dépasse largement le lavage observé partout où les deux ont été mesurés, parce que la question a une réponse manifestement attendue. La colonne employée ici est ce que l'enquêteur a vu, et son nom doit le dire — `soap_observed`, non `soap`.

2.5 Les trois échelles ne se moyennent pas

```

summary = pd.DataFrame({
  "ladder": ["Drinking water", "Sanitation", "Hygiene"],
  "at_least_basic": [0.547, 0.432, 0.341],
  "denominator": ["all households"] * 3,
})
print(summary)

```

```

tibble::tribble(
  ~ladder,      ~at_least_basic,
  "Drinking water",      0.547,
  "Sanitation",          0.432,
  "Hygiene",             0.341
)

```

Trois nombres sur le même dénominateur, ce qui est inhabituel et mérite d'être remarqué — ils peuvent légitimement figurer côte à côte. Ce qu'ils ne peuvent pas être, c'est combinés. Il n'existe pas d'indicateur de « couverture EAH », et une moyenne des trois ne décrirait aucun ménage : un ménage avec eau élémentaire, assainissement partagé et pas de savon n'est pas « couvert à 58 % », c'est un ménage qui a un service sur trois.

Là où un chiffre composite est exigé, publiez les trois et le nombre de ménages qui disposent des trois, qui est une quantité réelle et généralement dégrisante.

2.6 La suite

Les échelles classent. Elles ne disent pas combien d'eau est arrivée, combien de temps a duré l'attente, ni si l'eau était potable. Les deux leçons suivantes portent sur les mesures que Sphère ajoute par-dessus — et la première recèle deux erreurs d'unité.

Deuxième partie

Ce que Sphère demande et qu'une
échelle ne dit pas

CHAPITRE 3

Quinze litres, trente minutes, cinq cents mètres

Leçon 3 sur 8 · 120 min

13,4 % de ces ménages passent sous le minimum Sphère de quantité. Deux des colonnes qui produisent ce nombre portent des erreurs d'unité, et l'une d'elles fait monter un ménage d'un échelon plutôt que sortir d'un graphique.

3.1 Ce que Sphère ajoute à l'échelle

Les échelles du JMP classent un service. Les standards minimaux Sphère y mettent des nombres, et les trois sur lesquels un programme d'eau est jugé sont :

Standard	Seuil	Dans ce fichier
Quantité	Au moins 15 litres par personne et par jour	litres_per_person_day
Temps d'attente	30 minutes au plus	Inclus dans round_trip_minutes
Distance	500 m au plus jusqu'au point le plus proche	Non enregistrée ; le temps sert de proxy

Les seuils sont des minimums pour survivre dans la dignité, non des cibles. **Un programme à 15,2 litres a atteint le standard et ne se porte pas bien**, et un rapport qui n'imprime que la part au-dessus du seuil masque où se situe réellement la distribution.

```
import pandas as pd

households = pd.read_csv("wash-household-survey-2024.v1.csv")
litres = households["litres_per_person_day"]

print(f"median {litres.median():.1f} L/p/d")
print(f"below Sphere minimum: {(litres < 15).mean():.1%}")
print(litres.describe().round(1))

library(dplyr)

households |> summarise(
  median = median(litres_per_person_day),
  below_15 = mean(litres_per_person_day < 15),
  n = n()
)
```

13,4 % sous 15 litres, avec une médiane de 23,7. Les deux nombres ont leur place dans le rapport : le premier est le standard, le second dit que le ménage médian dispose d'environ la moitié en plus de ce qu'il lui faut et que le problème est de répartition plutôt que de volume total.

3.2 Deux erreurs d'unité, et une seule est visible

```
print(litres.nlargest(8).round(1).tolist())
```

```
households |> slice_max(litres_per_person_day, n = 8) |>
  select(household_id, household_size, litres_per_person_day)
```

Onze ménages dépassent 80 litres par personne et par jour, jusqu'à 277,6. C'est implausible dans ce contexte et ce n'est pas une faute de frappe — c'est la consommation **totale** du ménage saisie dans une colonne par personne. Divisez par la taille du ménage et chacun d'eux retombe dans une plage normale.

```
suspect = households[litres > 80]
recovered = suspect["litres_per_person_day"] / suspect["household_size"]
print(recovered.round(1).describe())
```

```
households |>
  filter(litres_per_person_day > 80) |>
  mutate(recovered = litres_per_person_day / household_size) |>
  summarise(min = min(recovered), max = max(recovered))
```

La seconde erreur est la **dangereuse**, parce qu'elle rend une valeur plus petite et non plus grande. Onze ménages ont un temps de collecte saisi en **heures**, si bien qu'un aller-retour de 90 minutes est enregistré comme 2. Il n'y a aucune valeur aberrante à attraper : 2 minutes est une valeur parfaitement ordinaire, et 85 ménages déclarent légitimement moins de 8 minutes.

```
short = households[households["round_trip_minutes"].between(1, 8)]
print(f"{len(short)} households report a round trip of 1 to 8 minutes")
print(short["water_source"].value_counts())
```

```
households |> filter(between(round_trip_minutes, 1, 8)) |> count(water_source)
```

Une erreur qui ramène une valeur dans la plage normale ne se trouve pas en **cherchant des valeurs aberrantes**. Elle se trouve par une règle qui lie deux champs — un aller-retour de moins de dix minutes vers une source qui n'est ni sur place ni une borne publique de la même communauté mérite une vérification — ou elle s'évite à la saisie par une étiquette d'unité sur le formulaire.

Sa conséquence ici n'est pas une moyenne fausse. C'est un **échelon faux** : chacun de ces ménages est classé en service élémentaire alors qu'il est limité.

3.3 L'attente est incluse dans l'aller-retour, et c'est un problème

Sphère fixe l'attente séparément du temps de collecte parce que les deux ont des causes et des remèdes différents. Une longue marche appelle un nouveau point d'eau ; une longue file appelle davantage de robinets au point existant, ou une plage horaire plus large.

Cette enquête n'enregistre que l'aller-retour, qui contient les deux. Le rapport honnête dit donc ce qu'il peut et ne peut pas séparer :

```
print("round_trip_minutes includes walking, queueing and return.")
print(f"over 30 minutes: {(households['round_trip_minutes'] > 30).mean():.1%}")
print("Queue time alone: not collected. Recommend adding it to the next round.")
```

```
# One line in the limitations section beats a queue statistic that is a walk.
```

39,3 % dépassent trente minutes, et rien de ce nombre ne peut être attribué à l'attente ou à la distance sans la ventilation. **Dites laquelle des deux vous ne voyez pas**, car l'action corrective diffère et un programme qui devine construira la mauvaise chose.

3.4 La distance n'est pas du tout dans le fichier

Le standard des 500 mètres exige une coordonnée ou une distance mesurée, et cette enquête n'a ni l'une ni l'autre. Le temps en est un proxy et un mauvais : trente minutes font un kilomètre sur une route et trois cents mètres dans un ravin.

Là où un standard nomme une unité dont vous ne disposez pas, rapportez le proxy sous son propre nom. Rapporter le temps d'aller-retour comme s'il s'agissait du standard de distance est la façon dont une évaluation conclut qu'une communauté est à moins de 500 mètres de l'eau alors qu'elle en est très loin.

3.5 Rapportez la distribution face au seuil

Water quantity, 2,403 households

Median	23.7 L/person/day
Below Sphere minimum (15 L)	13.4% 323 households
Below 7.5 L	2.2% 54 households

Collection time

Over 30 minutes	39.3% 944 households
-----------------	---------------------------

Queue time not separated from walking time in this round.

Distance not recorded; time is the proxy and is not equivalent.

11 records held household total litres in the per-person column and were corrected; 11 held collection time in hours and were corrected.

Le seuil, la distribution, et les corrections énoncées avec leurs effectifs. Un journal de nettoyage n'est pas un aveu — c'est ce qui permet au lecteur de voir que 13,4 % vient après correction de deux erreurs connues et non avant.

3.6 La suite

La quantité et le temps disent si l'eau est arrivée et à quel prix d'effort. Ils ne disent rien de sa potabilité, et la leçon suivante est celle où le dénominateur change sous vos pieds.

CHAPITRE 4

Testé sur un tiers, censuré sur un cinquième

Leçon 4 sur 8 · 120 min

E. coli a été testé sur 33,4 % des ménages, le chlore sur 40,0 %, et les deux sur 13,3 %. Dans le registre des points d'eau, 271 lectures de chlore sont la chaîne "<0.1" — et les supprimer relève la conformité apparente de seize points.

4.1 Le dénominateur change, et le rapport généralement non

La qualité de l'eau coûte cher à mesurer. Une enquête qui interroge 2 403 ménages en teste une fraction, et tout indicateur de qualité tourne donc sur un dénominateur différent de tout indicateur d'accès.

```
import pandas as pd

households = pd.read_csv("wash-household-survey-2024.v1.csv")

ecoli = households["ecoli_cfu_100ml"].notna()
chlorine = households["free_residual_chlorine_mgl"].notna()

print(f"E. coli tested: {ecoli.sum():>5} = {ecoli.mean():.1%}")
print(f"chlorine tested: {chlorine.sum():>5} = {chlorine.mean():.1%}")
print(f"both: {(ecoli & chlorine).sum():>5} = {(ecoli & chlorine).mean():.1%}"
      )

library(dplyr)

households |> summarise(
  ecoli = mean(!is.na(ecoli_cfu_100ml)),
  chlorine = mean(!is.na(free_residual_chlorine_mgl)),
  both = mean(!is.na(ecoli_cfu_100ml) & !is.na(free_residual_chlorine_mgl))
)
```

802 ménages pour E. coli, 961 pour le chlore, et 320 pour les deux. Le dernier nombre est celui qui compte : toute question liant les deux porte sur 13,3 % de l'enquête, et son intervalle sera large.

Imprimez le n analysable pour chaque chiffre de qualité, dans le tableau et non en note de bas de page. Un pourcentage de qualité voisin d'un pourcentage d'accès sur la même ligne se lit comme le même dénominateur, et ici ce n'est jamais le cas.

4.2 Les classes de risque du JMP

```
tested = households[ecoli]
classes = pd.cut(
    tested["ecoli_cfu_100ml"],
    [-1, 0, 10, 100, 10**9],
    labels=["safe <1", "low 1-10", "moderate 11-100", "high >100"],
)
print(classes.value_counts(normalize=True).round(3))
```

```
households |>
  filter(!is.na(ecoli_cfu_100ml)) |>
  mutate(risk = cut(ecoli_cfu_100ml, c(-1, 0, 10, 100, Inf),
    labels = c("safe", "low", "moderate", "high"))) |>
  count(risk) |> mutate(share = n / sum(n))
```

Classe de risque	Ménages	Part des testés
Sûre (<1 UFC/100 mL)	430	53,6 %
Faible (1–10)	173	21,6 %
Modérée (11–100)	139	17,3 %
Élevée (>100)	60	7,5 %

Rapportez les classes, non une moyenne. Les dénombrements d'E. coli sont fortement asymétriques et une moyenne de 14 UFC/100 mL ne décrit rien — les classes sont ce dans quoi la directive est écrite et ce sur quoi une décision se prend.

4.3 Amélioré n'est pas sûr, et c'est ici que cela se démontre

```
IMPROVED = {"piped-into-dwelling", "piped-into-yard", "public-tap",
  "borehole", "protected-well", "protected-spring"}

tested = tested.assign(improved=tested["water_source"].isin(IMPROVED))
print(tested.groupby("improved")["ecoli_cfu_100ml"].agg(
  n="size", safe=lambda s: (s < 1).mean(), high=lambda s: (s > 100).mean())
).round(3))
```

```
households |>
  filter(!is.na(ecoli_cfu_100ml)) |>
  mutate(improved = water_source %in% improved) |>
  summarise(n = n(), safe = mean(ecoli_cfu_100ml < 1),
    high = mean(ecoli_cfu_100ml > 100), .by = improved)
```

Source	Testés	Sûre	Risque élevé
Améliorée	625	61,1 %	3,8 %
Non améliorée ou de surface	177	27,1 %	20,3 %

Le type de source prédit fortement la qualité, et **38,9 % des ménages sur une source améliorée ont malgré tout de l'E. coli détectable**. La leçon 1 disait que l'échelon supérieur exige l'absence de contamination en plus d'une source améliorée; voici le nombre qui montre pourquoi c'est une condition distincte et non une formalité.

4.4 La valeur censurée, et pourquoi elle n'est pas manquante

Le registre des points d'eau mesure le chlore sur le terrain, et le kit a une limite de détection.

```
points = pd.read_csv("water-point-monitoring-2024.v1.csv")
readings = points["free_residual_chlorine_mgl"].dropna()

censored = readings.astype(str).str.startswith("<")
print(f"tested {len(readings)}, censored {censored.sum()}")
print(readings[censored].unique())
```

```
points |>
  filter(!is.na(free_residual_chlorine_mgl)) |>
  count(censored = startsWith(free_residual_chlorine_mgl, "<"))
```

811 visites portent une lecture et **271 d'entre elles sont la chaîne <0.1**. C'est une mesure : la vraie valeur se situe entre zéro et la limite de détection. Ce n'est pas une valeur manquante, et les deux choses en lesquelles on la transforme habituellement sont fausses toutes les deux.

```
numeric = pd.to_numeric(readings, errors="coerce") # <-- the silent one
print(f"after coercion: {numeric.isna().sum()} became NaN")

in_range = readings[~censored].astype(float).between(0.2, 0.5)
print(f"in target range, over all tested: {in_range.sum() / len(readings):.1%}")
print(f"in target range, dropping censored: {in_range.mean():.1%}")
```

```
points |>
  filter(!is.na(free_residual_chlorine_mgl),
         !startsWith(free_residual_chlorine_mgl, "<")) |>
  summarise(in_range = mean(between(as.numeric(free_residual_chlorine_mgl), 0.2, 0.5)))
```

30,9 % contre 46,5 %. Supprimer les lectures censurées relève la conformité apparente à la cible de chloration de près de seize points, et le fait dans une seule direction — toute valeur censurée est un échec, si bien que les retirer ne retire que des échecs.

`pd.to_numeric(..., errors="coerce")` le fait en silence et ressemble à une correction de type.

4.5 Que faire d'une valeur censurée

Trois options, par ordre de préférence pour cet indicateur.

Classer plutôt que moyennner. La cible est une plage, si bien qu'une lecture censurée répond parfaitement à la question : <0.1 est en dessous de 0,2 et donc non conforme. La conformité se calcule sur les 811 lectures sans aucune substitution.

Substituer la limite, ou sa moitié, s'il vous faut vraiment une moyenne. Dites laquelle vous avez employée — LD/2 est la convention courante — et rapportez à combien de valeurs elle s'est appliquée.

Rapporter la part sous la limite comme un nombre à part entière. 33,4 % de ces lectures sont sous le seuil de détection, ce qui est un constat sur la pratique de chloration et non une gêne dans les données.

```
compliant = pd.to_numeric(readings.where(~censored), errors="coerce").between(0.2, 0.5)
result = pd.DataFrame({
  "readings": [len(readings)],
  "below detection": [censored.sum()],
  "in 0.2-0.5 range": [compliant.sum()],
  "compliance": [f"{compliant.sum() / len(readings):.1%}"],
})
print(result)
```

```
# The denominator is every reading, including the ones below the limit.
```

Ne laissez jamais une limite de détection rétrécir un dénominateur. C'est la façon la plus courante dont un rapport de qualité de l'eau devient optimiste, et elle survit à

la relecture parce que l'arithmétique sur les valeurs restantes est juste.

4.6 L'incohérence entre champs qu'il faut nommer

```
treats = households["water_treated_at_home"]
print(f"treat at home: {treats.sum()}")
print(f" of which no chlorine reading: {(treats & ~chlorine).sum()}")
```

```
households |> filter(water_treated_at_home) |>
  count(no_reading = is.na(free_residual_chlorine_mgl))
```

533 ménages déclarent traiter leur eau et n'ont aucune mesure de chlore. Le cours sur le nettoyage a établi de quoi il s'agit — une absence corrélée à la chose mesurée — et la conséquence ici est précise : **supprimer ces lignes retire des ménages qui traitent leur eau plus souvent que la moyenne**, si bien que le sous-échantillon testé n'est pas représentatif de l'enquête dont il vient.

4.7 Rapportez-le comme un bloc à part

Water quality

E. coli, point of collection	802 households tested (33.4% of survey)
Safe <1 CFU/100 mL	53.6%
Low 1-10	21.6%
Moderate 11-100	17.3%
High >100	7.5%
Improved sources: 61.1% safe (n=625)	
Unimproved and surface: 27.1% safe (n=177)	

Free residual chlorine, water points	811 visits tested (30.8% of visits)
Below detection limit (<0.1 mg/L)	33.4% counted as non-compliant
Within 0.2-0.5 mg/L target	30.9%

Quality was tested on a subsample and does not share the denominator of the access indicators above. 533 households reporting home treatment have no chlorine reading, so the tested subsample over-represents untreated water.

Son propre bloc, ses propres dénominateurs, sa propre limite. **Un chiffre de qualité placé dans un tableau de chiffres d'accès sera lu contre leur dénominateur**, et rien dans la mise en page ne l'en empêchera.

4.8 La suite

Tout ce qui précède est une coupe transversale : ce qui était vrai le jour où quelqu'un est passé. L'unité suivante porte sur le registre qui visite les mêmes points d'eau douze fois, et sa première conséquence est de transformer un taux de fonctionnalité en trois.

Troisième partie

**Le service est une année, non une
journée**

CHAPITRE 5

74,4 %, 33,9 %, 83,3 %

Leçon 5 sur 8 · 120 min

Trois taux de fonctionnalité issus d'un seul registre, tous exacts. Ils diffèrent parce qu'ils comptent des visites, des points et des personnes, et un seul répond à la question qu'on pose réellement à un programme d'eau.

5.1 Ce que rend possible un registre à visites répétées

Toute la première unité est une coupe transversale. Une enquête ménage dit quel service existait le jour où l'enquêteur est passé, et elle ne peut structurellement pas dire combien de temps dans l'année l'eau a été là.

Le registre de suivi le peut, parce qu'il revisite les mêmes points.

```
import pandas as pd

points = pd.read_csv("water-point-monitoring-2024.v1.csv", parse_dates=["visit_date"])

print(f"visits: {len(points):,}")
print(f"water points: {points['water_point_id'].nunique()}")
print(points["functional_status"].value_counts())

library(dplyr)

points |> summarise(visits = n(), water_points = n_distinct(water_point_id))
points |> count(functional_status)
```

2 629 visites sur 242 points, en douze tournées mensuelles. Cette structure permet trois taux de fonctionnalité différents, et l'écart entre eux n'est pas une chicane de méthode — ce sont trois questions différentes.

5.2 Taux un : la part des visites ayant trouvé un point en service

```
WORKING = {"functional", "partially-functional"}
working = points["functional_status"].isin(WORKING)

print(f"visit-level functionality: {working.mean():.1%} of {len(points):,} visits")

points |> summarise(functionality = mean(functional_status %in%
                                     c("functional", "partially-functional")), n = n())
```

74,4 %. C'est ce que publie presque tout rapport sur les points d'eau, et cela répond à « *si je visite un point au hasard, sera-t-il en service ?* »

C'est le plus facile à calculer et le plus facile à biaiser, parce que le dénominateur est les visites faites et non les visites dues. La leçon 7 y est entièrement consacrée.

Partiellement fonctionnel compte comme en service. Un point qui tourne à débit réduit fournit de l'eau, et le classer en panne mettrait un problème d'attente dans la même

catégorie qu'un forage tari. Dites de quel côté de la ligne vous le placez, car des analystes raisonnables divergent et les deux réponses sont écartées de trois points.

5.3 Taux deux : la part des points en service à chaque fois

```
by_point = points.assign(ok=working).groupby("water_point_id")["ok"]
always = by_point.all()

print(f"point-level functionality: {always.mean():.1%} of {len(always)} points")
print(f"points that failed at least once: {(~always).sum()}")
```

```
points |>
  summarise(always = all(functional_status %in%
    c("functional", "partially-functional")), .by = water_point_id) |>
  summarise(share = mean(always), n = n())
```

33,9 % — 82 points sur 242. Deux tiers du parc ont défailli au moins une fois dans l'année.

Cela répond à une autre question : « *combien de ces ouvrages sont fiables ?* » Et c'est le nombre dont un gestionnaire de patrimoine a besoin, car un point qui fonctionne trois visites sur quatre n'est pas les trois quarts d'une desserte en eau — c'est une desserte trouée qu'un ménage doit combler autrement.

Les deux taux ne se contredisent pas. 74,4 % des visites et 33,9 % des points sont vrais tous les deux, du même fichier, en même temps. Un point peut fonctionner la plupart du temps et défaillir quand même dans l'année, et un parc peut être majoritairement en service et majoritairement peu fiable.

5.4 Taux trois : la part des personnes disposant d'un point en service

```
served = points.dropna(subset=["users_estimated"])
population_weighted = (
  served.loc[served["functional_status"].isin(WORKING), "users_estimated"].sum()
  / served["users_estimated"].sum()
)
print(f"population-weighted functionality: {population_weighted:.1%}")
```

```
points |>
  filter(!is.na(users_estimated)) |>
  summarise(weighted = sum(users_estimated[functional_status %in%
    c("functional", "partially-functional")]) / sum(users_estimated))
```

83,3 %, près de neuf points au-dessus du chiffre par visite. La raison est dans les types de source.

```
print(points.groupby("source_type").agg(
  points=("water_point_id", "nunique"),
  median_users=("users_estimated", "median"),
  functionality=("functional_status", lambda s: s.isin(WORKING).mean()),
).round(3))
```

```
points |>
  summarise(n = n_distinct(water_point_id),
            median_users = median(users_estimated, na.rm = TRUE),
            functionality = mean(functional_status %in%
                                c("functional", "partially-functional")), .by = source_type)
```

Type de source	Fonctionnalité	Usagers typiques
Borne de réseau	97,2 %	environ 1 290
Forage à pompe manuelle	75,5 %	environ 280
Source protégée	66,6 %	environ 200
Puits protégé	58,1 %	environ 185

Les points qui desservent le plus de monde tombent le moins en panne. Pondérer par la population fait donc monter le chiffre, et l'écart entre 74,4 % et 83,3 % est un énoncé réel sur qui supporte les pannes : les usagers des puits protégés, qui sont les moins nombreux par point et les plus mal servis.

5.5 Lequel rapporter

Les trois, et c'est un des cas où trois nombres sont vraiment la bonne réponse.

Water point functionality, 2024

```
Visit-level      74.4%    1,957 of 2,629 monitoring visits
Point-level      33.9%      82 of 242 points working at every visit
Population-weighted 83.3% of an estimated 96,700 users
```

Partially functional counted as working (134 visits).
Two points were re-registered after a handover and are counted twice;
removing them moves the visit-level figure to 74.7%.

S'il faut n'en donner qu'un, donnez le taux pondéré par la population et dites-le, car le standard porte sur des personnes et non sur des ouvrages. Mais publiez le taux par point à côté, car c'est lui qui dit que le parc n'est pas fiable, et le taux pondéré le masque derrière une poignée de réseaux bien tenus.

5.6 Les deux points comptés deux fois

```
duplicates = points[points["water_point_id"].str.startswith("WP09")]
print(duplicates.groupby("water_point_id")["community"].agg(["nunique", "first"]))
```

```
points |> filter(startsWith(water_point_id, "WP09")) |> count(water_point_id)
```

Deux points ont été transférés d'un programme à un autre et réenregistrés sous de nouveaux identifiants, si bien que le registre les détient deux fois et que chaque dénominateur est trop grand de deux points. L'effet est faible — 74,4 % devient 74,7 % — et **la taille de l'effet n'est pas la raison de corriger**. Un registre patrimonial qui compte double se trompe sur ce qui existe, et la première chose que l'on en fait ensuite est de planifier un budget de maintenance.

5.7 La suite

Deux tiers de ces points ont défailli au moins une fois. Ce nombre traite un forage tari en février exactement comme un point abandonné depuis mars, et la leçon suivante les sépare — au moyen de la séquence des visites plutôt que du champ d'état, car le champ d'état ne peut pas les distinguer.

CHAPITRE 6

Tari n’est pas en panne

Leçon 6 sur 8 · 120 min

Les puits protégés tournent à 74,7 % en saison humide et à 36,4 % en saison sèche. Les forages à pompe manuelle ne bougent presque pas. Le champ d’état ne distingue pas ces deux défaillances, la séquence des visites le peut.

6.1 Un seul état, plusieurs défaillances différentes

`functional_status` dit qu’un point ne fonctionne pas. Il ne dit pas si la pompe est cassée, si la nappe a baissé, si le comité a cessé de percevoir les redevances, ou si la communauté est partie. Ces cas appellent quatre réponses différentes et l’une d’elles n’est pas une réparation. La séquence des visites peut les séparer, parce qu’ils ont des formes différentes dans le temps.

Forme sur l’année	Ce que c’est	Que
À l’arrêt les mêmes mois chaque année, en service entre-temps	Saisonnier — la source se tarit	Appro
À l’arrêt une fois, puis en service après un intervalle	Panne — réparée	Regar
À l’arrêt et le reste jusqu’à la fin	Abandonné ou en attente de pièces	Déter
En service et à l’arrêt à répétition	Chronique — sous-entretenu ou surchargé	Probl

6.2 Le signal saisonnier, et là où il n’est pas

```
import pandas as pd

points = pd.read_csv("water-point-monitoring-2024.v1.csv", parse_dates=["visit_date"])
WORKING = {"functional", "partially-functional"}
DRY_MONTHS = {1, 2, 3, 11, 12}

points["dry_season"] = points["visit_date"].dt.month.isin(DRY_MONTHS)
points["ok"] = points["functional_status"].isin(WORKING)

print(points.groupby("dry_season")["ok"].agg(["mean", "size"]).round(3))

library(dplyr)

points |>
  mutate(dry = lubridate::month(visit_date) %in% c(1, 2, 3, 11, 12),
         ok = functional_status %in% c("functional", "partially-functional")) |>
  summarise(functionality = mean(ok), n = n(), .by = dry)
```

67,8 % en saison sèche contre 79,4 % en saison humide. Onze points d’écart, qu’un rapport comparant une évaluation de mars à une de juillet lirait comme un effondrement ou comme un redressement selon le sens de la soustraction. Ventilez maintenant par type de source, où se trouve réellement le constat.

```
seasonal = (
    points.groupby(["source_type", "dry_season"])["ok"].mean().unstack()
)
seasonal.columns = ["wet", "dry"]
seasonal["swing"] = seasonal["wet"] - seasonal["dry"]
print((seasonal * 100).round(1).sort_values("swing", ascending=False))
```

```
points |>
  summarise(functionality = mean(ok), .by = c(source_type, dry)) |>
  tidyr::pivot_wider(names_from = dry, values_from = functionality)
```

Type de source	Humide	Sèche	Écart
Puits protégé	74,7 %	36,4 %	38,3
Source protégée	84,8 %	41,0 %	43,8
Forage à pompe manuelle	75,0 %	76,1 %	−1,1
Borne de réseau	95,9 %	98,8 %	−2,9

Deux types de source portent tout l'effet saisonnier et deux ne bougent pas du tout. Les puits peu profonds et les sources puisent dans une nappe qui baisse ; un forage descend en dessous et un réseau dispose d'un réservoir. Les onze points de moyenne étaient un mélange d'un effet de quarante points et d'aucun effet.

C'est là le constat. **Les forages de ce parc n'ont pas de problème saisonnier et les puits en ont un**, si bien qu'un programme de saison sèche qui réhabilite des forages réhabilite le mauvais ouvrage.

6.3 Classer chaque point à partir de sa propre séquence

```
def classify(group):
    group = group.sort_values("visit_date")
    failures = group.loc[~group["ok"]]
    if failures.empty:
        return "never failed"
    if not group["ok"].tail(3).any() and len(group) >= 3:
        return "down at year end"
    if failures["dry_season"].all():
        return "seasonal"
    return "intermittent"

shape = points.groupby("water_point_id").apply(classify, include_groups=False)
print(shape.value_counts())
```

```
points |>
  arrange(visit_date) |>
  summarise(
    shape = case_when(
      all(ok) ~ "never failed",
      !any(tail(ok, 3)) & n() >= 3 ~ "down at year end",
      all(dry[!ok]) ~ "seasonal",
      TRUE ~ "intermittent"
    ), .by = water_point_id) |>
  count(shape)
```

Forme	Points
Jamais défailli	82
Intermittent	80
Saisonnier	46
À l'arrêt en fin d'année	34

L'ordre des tests compte et c'est un jugement, non un détail. Un point qui ne défaille qu'en mois secs et qui est encore à l'arrêt en décembre satisfait les deux règles. Tester la permanence d'abord l'appelle abandonné ; tester la saisonnalité d'abord l'appelle saisonnier. Ce code teste la permanence d'abord, parce qu'un point qui n'a pas tourné depuis trois visites est un problème opérationnel maintenant quelle qu'en soit la cause — et parce que la lecture de décembre est celle sur laquelle un plan de janvier s'écrit.

Notez l'ordre employé. Deux analystes avec le même fichier et des ordres différents produiront des effectifs différents et auront tous deux raison.

6.4 La valeur d'état qui ressemble à la réponse sans l'être

Le registre comporte un état `abandoned`, et il est tentant de l'employer directement.

```
ever_abandoned = points.loc[points["functional_status"].eq("abandoned"),
                             "water_point_id"].nunique()
print(f"points ever recorded as abandoned: {ever_abandoned}")
print(f"points down at the end of the year: 34")
```

```
points |> filter(functional_status == "abandoned") |>
  summarise(points = n_distinct(water_point_id))
```

Dix-sept points sont enregistrés comme abandonnés lors d'une visite ; trente-quatre sont encore à l'arrêt en fin d'année. **L'état est le jugement d'un enquêteur porté à une visite, et la séquence est une preuve.** La moitié des points qui ne reviennent jamais n'ont jamais été étiquetés abandonnés, parce que l'enquêteur qui les a vus en juillet n'avait aucun moyen de savoir qu'ils seraient encore à l'arrêt en décembre.

Dérivez l'état de l'historique là où vous le pouvez, et employez l'état enregistré comme recoupement. Un champ qu'un enquêteur remplit à partir d'une observation unique ne peut pas porter une information qui n'existe qu'à travers plusieurs observations.

6.5 La durée d'arrêt est une statistique de gestion

```
down = points.loc[~points["ok"] & points["days_since_breakdown"].notna()]
print(down.groupby("management")["days_since_breakdown"].agg(
    median="median", visits="size"
).sort_values("median"))
```

```
points |>
  filter(!ok, !is.na(days_since_breakdown)) |>
  summarise(median_days = median(days_since_breakdown), n = n(), .by = management)
```

Gestion	Jours d'arrêt médians à la visite
Opérateur privé	29
Service public	32
Géré par une ONG	48
Comité communautaire	68

Un point en panne sous comité communautaire est à l'arrêt depuis plus de deux fois plus longtemps qu'un point sous opérateur privé. **C'est le seul nombre du registre qui désigne une action plutôt qu'un ouvrage** — la différence tient à une chaîne d'approvisionnement en pièces et à un fonds d'entretien, non à la pompe.

Méfiez-vous des trente-sept lignes où un point est enregistré comme fonctionnel et porte malgré tout un `days_since_breakdown` : c'est la réponse du mois précédent reportée. Filtrez sur l'état, non sur la présence du champ.

6.6 Rapportez les formes, non le seul taux

Water point failure, 2024, 242 points

Never failed	82	33.9%
Intermittent	80	33.1%
Seasonal (dry months only)	46	19.0%
Down at year end	34	14.0%

Seasonality is concentrated in protected wells (74.7% wet, 36.4% dry) and protected springs (84.8% wet, 41.0% dry). Handpump boreholes and piped schemes show no seasonal effect.

Points classified as down at year end if not working at any of the last three visits; seasonal if every recorded failure fell in a dry month. Permanence tested before seasonality.

Quatre effectifs et la règle qui les a produits. **Un pourcentage de fonctionnalité unique enverrait un budget de réhabilitation vers les mauvais quarante-six points**, car ils n'ont pas besoin d'être réhabilités — il leur faut une alternative de saison sèche, et leurs pompes vont bien.

6.7 La suite

Chaque nombre de cette leçon est calculé sur les visites qui ont été faites. Un district a manqué un tiers de ses tournées en saison des pluies, et les points manqués n'étaient pas un tiers au hasard — c'est la leçon suivante, et la raison pour laquelle sa fonctionnalité semble s'améliorer quand les routes se ferment.

Quatrième partie

Ce que le suivi n'a pas vu

CHAPITRE 7

La tournée que personne n'a conduite

Leçon 7 sur 8 · 120 min

2 629 visites faites sur 2 904 dues. Nord-Ouest se lit comme le meilleur des trois districts à 76,4 % et sa réponse est incertaine à treize points près, contre deux et demi pour le district qui a presque tout visité.

7.1 Le dénominateur que personne n'a interrogé

Tous les chiffres de fonctionnalité des deux dernières leçons divisaient par les **visites faites**. Personne n'a choisi ce dénominateur ; c'est ce qu'un `groupby` produit à partir d'un fichier dont les lignes sont des visites.

Le dénominateur que vous vouliez est celui des **visites dues**, et le registre sait ce que c'est : 242 points, douze tournées prévues.

```
import pandas as pd

points = pd.read_csv("water-point-monitoring-2024.v1.csv", parse_dates=["visit_date"])

due = points["water_point_id"].nunique() * 12
made = len(points)
print(f"due {due:}, made {made:}, coverage {made / due:.1%}")

library(dplyr)

points |> summarise(
  due = n_distinct(water_point_id) * 12,
  made = n(),
  coverage = n() / (n_distinct(water_point_id) * 12)
)
```

2 629 sur 2 904 — une couverture de 90,5 %. Deux cent soixante-quinze tournées n'ont jamais été conduites, et le registre ne contient aucune ligne pour le dire. C'est le taux de rapportage du cours DHIS2 qui revient sous un autre format de fichier : **une absence n'est pas une valeur**, et rien dans les données ne vous rappellera qu'elle a eu lieu.

7.2 La couverture n'est pas répartie uniformément, et c'est tout le problème

```
coverage = points.groupby("district").agg(
  points=("water_point_id", "nunique"),
  visits=("water_point_id", "size"),
)
coverage["due"] = coverage["points"] * 12
coverage["coverage"] = coverage["visits"] / coverage["due"]
print(coverage.round(3))

points |>
```

```
summarise(points = n_distinct(water_point_id), visits = n(), .by = district) |>
mutate(due = points * 12, coverage = visits / due)
```

District	Points	Visites faites	Dues	Couverture
Sud-Est	79	913	948	96,3 %
Centre	79	882	948	93,0 %
Nord-Ouest	84	834	1 008	82,7 %

Les tournées de Nord-Ouest s'effondrent quand les routes se ferment.

```
monthly = points.pivot_table(
  index=points["visit_date"].dt.month, columns="district",
  values="water_point_id", aggfunc="size",
)
print(monthly)
```

```
points |> count(district, month = lubridate::month(visit_date)) |>
tidyr::pivot_wider(names_from = district, values_from = n)
```

D'environ 77 points par mois à entre 51 et 60 en juin, juillet, août et septembre. Centre et Sud-Est bougent à peine.

7.3 Encadrez la réponse

La réponse honnête à un dénominateur que vous ne pouvez pas observer entièrement n'est pas un facteur de correction. C'est une **fourchette**, calculée à partir des deux hypothèses qui encadrent la vérité.

- **Borne haute** : chaque visite manquée aurait trouvé un point en service. C'est le taux observé, qui suppose déjà que les visites manquées ressemblent aux visites faites.
- **Borne basse** : chaque visite manquée aurait trouvé un point en panne.

```
WORKING = {"functional", "partially-functional"}
ok = points["functional_status"].isin(WORKING)

bounds = points.assign(ok=ok).groupby("district").agg(
  ok=("ok", "sum"), made=("ok", "size"), points=("water_point_id", "nunique"),
)
bounds["due"] = bounds["points"] * 12
bounds["observed"] = bounds["ok"] / bounds["made"]
bounds["lower"] = bounds["ok"] / bounds["due"]
bounds["band"] = bounds["observed"] - bounds["lower"]
print((bounds[["observed", "lower", "band"]] * 100).round(1))
```

```
points |>
mutate(ok = functional_status %in% c("functional", "partially-functional")) |>
summarise(ok = sum(ok), made = n(), pts = n_distinct(water_point_id),
  .by = district) |>
mutate(due = pts * 12, observed = ok / made, lower = ok / due)
```

District	Observé	Borne basse	Fourchette
Nord-Ouest	76,4 %	63,2 %	13,2
Centre	75,7 %	70,5 %	5,3
Sud-Est	71,4 %	68,8 %	2,6

Nord-Ouest est le meilleur district sur le chiffre observé et le seul dont le classement n'est pas sûr. Sa fourchette recouvre entièrement celles des deux autres. Sud-Est paraît le plus mauvais et c'est le seul dont on puisse être sûr, parce qu'il a visité presque tout.

Un nombre calculé sur 96 % de son dénominateur et un nombre calculé sur 83 % ne sont pas comparables, et rien dans le tableau ne le dira si vous ne l'y mettez pas.

7.4 L'absence est-elle aléatoire ?

Les bornes sont larges parce qu'elles ne supposent rien. Les resserrer suppose de soutenir que les visites manquées ressemblent aux visites faites, et cet argument est testable.

```
observed = points.pivot_table(
    index="water_point_id", columns="round", values="functional_status",
    aggfunc="first",
)
followed_by_gap = []
for point, row in observed.iterrows():
    for r in range(1, 12):
        if pd.isna(row.get(r)):
            continue
        broken = row[r] not in WORKING
        followed_by_gap.append((pd.isna(row.get(r + 1)), broken))

check = pd.DataFrame(followed_by_gap, columns=["gap_next", "broken"])
print(check.groupby("gap_next")["broken"].agg(["mean", "size"]).round(3))
```

Same shape: was the point broken at the visit immediately before a missed round?

26,7 % des visites suivies d'une tournée manquée ont trouvé un point en panne, contre 24,7 % des visites suivies d'une tournée faite. L'écart va dans le sens que l'on redoutait — un point que l'on sait en panne est un peu plus souvent sauté — et il est assez faible pour ne ni prouver ni écarter le biais.

C'est un constat légitime et il doit être rapporté comme tel. **Un test dont la réponse est ambiguë n'est pas un test raté** ; il vous dit que les bornes ne peuvent pas honnêtement être resserrées, ce qui est ce qu'il fallait savoir.

7.5 Ce qu'il ne faut pas faire

Trois réparations qui paraissent toutes raisonnables.

Reporter la dernière observation. Remplir chaque tournée manquante par l'état précédent du point. Cela donne 74,2 % au total, presque identique au taux observé, et c'est l'hypothèse qu'un point en panne le reste et qu'un point en service continue — c'est-à-dire exactement ce qui est en question. Cela fabrique de la précision sans ajouter d'information.

Imputer par la moyenne du district. Suppose que les points manqués ressemblent aux points visités du même district, ce qui est l'hypothèse dont toute la leçon doute.

Retirer le district. Ôte du rapport le district le moins bien couvert, qui est presque toujours aussi le plus difficile d'accès et le plus mal servi.

```
print("Reported: 76.4% (observed), 63.2% to 76.4% (bounds), coverage 82.7%")
print("Not reported: a single imputed number that hides which of these it is")
```

```
# The band and the coverage are the finding. Neither is a caveat.
```

7.6 Rapportez la couverture à côté de chaque taux

Water point functionality by district, 2024

District	Functionality	Coverage	Range if unvisited were broken
Sud-Est	71.4%	96.3%	68.8 - 71.4
Centre	75.7%	93.0%	70.5 - 75.7
Nord-Ouest	76.4%	82.7%	63.2 - 76.4

275 of 2,904 scheduled visits were not made, concentrated in Nord-Ouest during June to September when roads are impassable. Districts are not ranked here: Nord-Ouest's range overlaps both others.

Le taux, la couverture, la fourchette, et une phrase qui refuse de classer. **Refuser de classer est un résultat**, et c'est plus utile qu'un palmarès qui s'inverse dès que quelqu'un conduit les tournées manquantes.

7.7 La suite

Vous disposez maintenant de tous les nombres EAH que ce cours peut produire et d'un jeu de réserves attaché à chacun. La dernière leçon porte sur le rapport où ils vont — quels chiffres tiennent sur la même page, quels dénominateurs doivent être énoncés, et le tableau unique qui dit au lecteur ce qui peut être comparé avec quoi.

CHAPITRE 8

Onze nombres et six dénominateurs

Leçon 8 sur 8 · 120 min

Tout ce que ce cours a calculé, sur une page, avec le tableau qui dit lesquels de ces chiffres peuvent être comparés entre eux. Six dénominateurs pour onze indicateurs, et aucune mise en page n'en avertira un lecteur d'elle-même.

8.1 Comptez les dénominateurs avant d'écrire quoi que ce soit

Le cours a produit onze indicateurs. Ils viennent de deux fichiers et tournent sur six dénominateurs différents.

Indicateur	Dénominateur	n
Eau de boisson au moins élémentaire	Tous les ménages	2 403
Assainissement au moins élémentaire	Tous les ménages	2 403
Service d'hygiène élémentaire	Tous les ménages	2 403
Défécation à l'air libre	Tous les ménages	2 403
Moins de 15 L/personne/jour	Tous les ménages	2 403
Aller-retour de plus de 30 minutes	Tous les ménages	2 403
Classe de risque E. coli	Ménages testés	802
Chlore au ménage	Ménages testés	961
Fonctionnalité des points d'eau	Visites de suivi faites	2 629
Points en service à chaque visite	Points d'eau	242
Fonctionnalité pondérée par la population	Personnes desservies	~96 700
Chlore au point d'eau	Visites testées	811

Six dénominateurs, et seules les six premières lignes en partagent un. Toute comparaison horizontale franchissant une frontière de ce tableau exige une phrase, et toute comparaison faite sans cette phrase sera faite de travers.

```
import pandas as pd

denominators = pd.DataFrame({
    "indicator": ["basic water", "basic sanitation", "basic hygiene",
                  "E. coli safe", "point functionality", "reliable points"],
    "denominator": ["households", "households", "households",
                    "households tested", "visits made", "water points"],
    "n": [2403, 2403, 2403, 802, 2629, 242],
})
print(denominators)
```

Build this table first, in code, and print it above the results.

Écrivez-le comme une donnée, non comme de la prose. Un tableau de dénominateurs produit par le script qui a calculé les nombres ne peut pas s'en écarter ; un paragraphe qui décrit les dénominateurs le peut et le fera.

8.2 Le rapport

WASH situation, three districts, 2024

ACCESS		n = 2,403 households
At least basic drinking water	56.3%	(82.3% on an improved source; 25.0 points sit on limited service for collection time alone)
At least basic sanitation	43.2%	(62.0% have an improved facility; 18.8 points are shared)
Basic hygiene service	34.1%	(67.8% have a facility; soap observed at half)
Open defecation	13.3%	

Safely managed not computable: availability was not asked and quality was tested on a subsample.

QUANTITY AND TIME		n = 2,403 households
Median	23.7 L/person/day	
Below Sphere minimum of 15 L	13.4%	323 households
Round trip over 30 minutes	39.3%	944 households
Queue time not separated from walking time. Distance not recorded.		

WATER QUALITY		n = 802 tested (33.4%)
Free from E. coli	53.6%	
High risk (>100 CFU/100 mL)	7.5%	
Improved sources	61.1% safe	n = 625
Unimproved and surface	27.1% safe	n = 177

Tested subsample over-represents untreated water: 533 households reporting home treatment have no chlorine reading.

WATER POINT FUNCTIONALITY		242 points, 2,904 visits due
Visits finding a working point	74.4%	2,629 visits made (90.5%)
Points working at every visit	33.9%	82 of 242
Population-weighted	83.3%	~96,700 users
Range if every missed visit found a failure: 67.4% - 74.4%		

Failure shapes: 82 never failed, 80 intermittent, 46 seasonal, 34 down at year end. Seasonality is confined to protected wells and springs.

Median days down at the visit: 68 under community committees, 29 under private operators.

NOT COMPARABLE

Water quality figures are on a tested subsample and do not share the access denominator. Functionality figures count visits, points and people and are three different quantities. Districts are not ranked: Nord-Ouest visited 82.7% of its due rounds against Sud-Est's 96.3%.

8.3 Quatre règles auxquelles la mise en page doit obéir

Grouper par dénominateur, non par thème. Les intitulés de blocs ci-dessus sont des dénominateurs portant des noms de sujets. Un lecteur qui prend un nombre dans chaque bloc a pris quatre nombres à quatre bases, et c'est la ligne blanche entre les blocs qui le lui dit.

Mettre le n dans l'intitulé, non en note. Une note se lit après que la comparaison a déjà été faite.

Énoncer ce qui n'est pas calculable. L'eau gérée en toute sécurité, le temps d'attente et la distance sont tous absents, et chacun est une chose qu'un lecteur s'attend à trouver dans un rapport EAH. Un indicateur absent que l'on ne nomme pas se lit comme un indicateur à zéro ou comme un oubli.

Placer la non-comparabilité dans son propre bloc à la fin. Non comme réserve sous chaque nombre, où elle devient du bruit, et non dans une annexe méthodologique, où elle n'est pas lue.

8.4 À quoi sert chaque nombre

Un rapport qui ne fait que classer est une évaluation. Un rapport qui dit aussi quoi faire est un livrable, et les chiffres EAH se répartissent nettement selon la décision qu'ils éclairent.

Constat	Décision éclairée
25 points en service limité pour le temps de collecte	Où implanter de nouveaux points
18,8 points sur assainissement partagé	Le programme construit-il ou subventionne-t-il
Du savon dans la moitié des dispositifs existants	Promotion de l'hygiène contre matériel
46 points saisonniers	Alternative de saison sèche, non réhabilitation
34 points à l'arrêt en fin d'année	Pièces et gestion, non construction neuve
68 jours d'arrêt médians sous comité	Le modèle d'entretien lui-même
Nord-Ouest à 82,7 % de couverture	Le plan de suivi avant le prochain rapport

La dernière ligne est celle que la plupart des rapports omettent et la plus susceptible de changer les chiffres de l'an prochain. Un système de suivi incapable d'atteindre un district pendant les pluies produit un rapport incapable de le décrire, et y remédier coûte moins cher que tout le reste de la liste.

8.5 Ce que ce cours n'a pas couvert

Nommé clairement, car on demandera tout cela à un analyste EAH.

- **La gestion de l'hygiène menstruelle** exige des questions que cette enquête ne pose pas — espace privé, matériel, élimination — et ses indicateurs ne se déduisent pas de ce qui est ici.
- **L'EAH en milieu scolaire et sanitaire** relève d'instruments distincts avec leurs propres échelles, et l'échelle ménage ne se transpose pas à une institution.
- **La gestion des boues de vidange** est la seconde moitié de l'assainissement géré en toute sécurité et exige une enquête sur le confinement et la vidange.
- **Le coût par bénéficiaire et l'analyse tarifaire** exigent des données financières ; le champ `fee_collected` dit qu'une redevance a été perçue, non combien ni où elle est allée.

Un rapport qui nomme ses lacunes est plus crédible qu'un rapport qui paraît complet, et une lacune nommée est la première ligne des prochains termes de référence.

8.6 La suite

Le module 4 se poursuit dans d'autres secteurs — sécurité alimentaire et IPC, gestion de cas en protection, assiduité scolaire — et chacun apporte les mêmes trois questions dans un

vocabulaire nouveau. Vous les avez maintenant rencontrées trois fois : qu'ai-je compté, sur quoi, et que faut-il changer à cause de cela.

À propos de cette édition

Ce PDF est produit à partir de la même source que la version web de la formation ; les deux ne peuvent donc pas diverger. L'édition en ligne comporte le code exécutable, les jeux de données téléchargeables et les corrections publiées depuis la production de ce fichier.

Tous les jeux de données cités dans cette formation sont synthétiques ou entièrement anonymisés. Aucun enregistrement ne renvoie à une personne réelle.

Cette formation est publiée en anglais et en français. L'édition française est une adaptation professionnelle reprenant la terminologie employée par l'UNICEF, l'OMS, le PAM, OCHA et les ministères nationaux — et non une traduction littérale.

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/wash-analysis

Le contenu pédagogique est diffusé sous licence CC BY 4.0. Vous pouvez l'adapter et le rediffuser, y compris à des fins commerciales, en créditant Cassion.

Généré le 28 juillet 2026.