

Deux écarts, deux réponses

Un écart de 5,5 points avec un intervalle de $-0,4$ à $+11,3$, et un écart de 19,4 points avec un intervalle de $-26,1$ à $-12,8$. Même test, même code, conclusions opposées — et le module 4 avait laissé les deux ouverts exprès.

Pierre Robentz Cassion

Leçon 3 sur 8

Statistiques appliquées aux programmes — Comparer deux groupes

Ce que couvre cette leçon

- Les deux questions que le module 4 a reportées
- Écart numéro un : celui qui n'y est pas
- Écart numéro deux : celui qui y est
- Les deux résultats côte à côte
- Choisir le test
- Vérifiez les hypothèses, et dites que vous l'avez fait
- Rapportez la différence, non les deux nombres
- La suite

Les deux questions que le module 4 a reportées — En Python

```
import pandas as pd
import numpy as np
from statsmodels.stats.proportion import proportions_ztest, confint_proportions_2indep

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
clean = enrolment[(enrolment["age_years"] <= 20)
                  & (enrolment["school_year"] == 2024)
                  & enrolment["grade"].between(1, 6)]
clean = clean.assign(over_age=clean["age_years"] > clean["grade"] + 5)

boys = clean[clean["sex"] == "m"]["over_age"]
girls = clean[clean["sex"] == "f"]["over_age"]
print(f"boys {boys.sum()}/{len(boys)} = {boys.mean():.1%}")
print(f"girls {girls.sum()}/{len(girls)} = {girls.mean():.1%}")
```

Les deux questions que le module 4 a reportées — En R

```
library(dplyr)

enrolment |>
  filter(age_years <= 20, school_year == 2024, between(grade, 1, 6)) |>
  mutate(over_age = age_years > grade + 5) |>
  summarise(k = sum(over_age), n = n(), .by = sex)
```

Écart numéro un : celui qui n'y est pas

Groupe	Retard d'âge	n	Intervalle à 95 %
Garçons	39,2 %	510	35,1–43,5 %
Filles	33,8 %	542	29,9–37,8 %

Écart numéro un : celui qui n'y est pas — En Python

```
count = np.array([boys.sum(), girls.sum()])
nobs = np.array([len(boys), len(girls)])

stat, pvalue = proportions_ztest(count, nobs)
low, high = confint_proportions_2indep(count[0], nobs[0], count[1], nobs[1])
print(f"difference {boys.mean() - girls.mean():+.1%}")
print(f"95% CI [{low:+.1%}, {high:+.1%}]")
print(f"z = {stat:.2f}, p = {pvalue:.3f}")
```

Écart numéro un : celui qui n'y est pas — En R

```
prop.test(c(200, 183), c(510, 542))
```

Écart numéro un : celui qui n'y est pas

- Différence +5,5 points, IC à 95 % $-0,4$ à $+11,3$, $p = 0,066$

Écart numéro deux : celui qui y est — En Python

```
protection = pd.read_csv("protection-referrals-2024.v1.csv")
consenting = protection[protection["consent_to_refer"]]
reached = (consenting["referral_accepted"]
           & consenting["days_to_first_service"].notna())

disability = consenting["disability_reported"].isin([True, "true", "Yes"])
k = np.array([reached[disability].sum(), reached[~disability].sum()])
n = np.array([disability.sum(), (~disability).sum()])

stat, pvalue = proportions_ztest(k, n)
low, high = confint_proportions_2indep(k[0], n[0], k[1], n[1])
print(f"{k[0]}/{n[0]} = {k[0]/n[0]:.1%}    {k[1]}/{n[1]} = {k[1]/n[1]:.1%}")
print(f"difference {k[0]/n[0] - k[1]/n[1]:+.1%}, CI [{low:+.1%}, {high:+.1%}]")
print(f"z = {stat:.2f}, p = {pvalue:.2e}")
```

Écart numéro deux : celui qui y est — En R

```
prop.test(c(54, 663), c(202, 1436))
```

Écart numéro deux : celui qui y est

Groupe	Aboutissement	n
Handicap signalé	26,7 %	202
Non signalé	46,2 %	1 436

Écart numéro deux : celui qui y est

- Différence $-19,4$ points, IC à 95 % $-26,1$ à $-12,8$, $p < 0,001$
- Toute valeur de cet intervalle est un écart substantiel — Le plus petit écart compatible avec les données vaut $12,8 \dots$

Les deux résultats côte à côte — En Python

```
results = pd.DataFrame([
    {"comparison": "over-age, boys vs girls", "diff": 5.5,
     "ci": "-0.4 to +11.3", "n": "510 vs 542", "p": 0.066},
    {"comparison": "completion, disability", "diff": -19.4,
     "ci": "-26.1 to -12.8", "n": "202 vs 1436", "p": 0.0000002},
])
print(results)
```

Les deux résultats côte à côte — En R

Two rows. The conclusion column is the analyst's, not the test's.

- Remarquez que le résultat significatif a le plus petit échantillon dans l'un des bras
— 202 cas ont produit une réponse. . .

Choisir le test

Comparaison	Test	Dans ce cours
Deux proportions	Test z à deux échantillons, ou <code>prop.test</code>	Les deux écarts ci-dessus
Deux moyennes	Test t de Welch	Présence et cantine, leçon 6
Une variable catégorielle contre une autre	Khi-deux	Motif de fermeture par district

Choisir le test — En Python

```
from scipy import stats

points = pd.read_csv("water-point-monitoring-2024.v1.csv")
table = pd.crosstab(points["admin2"], points["functional_status"])
chi2, p, dof, expected = stats.chi2_contingency(table)
print(f"chi-square {chi2:.1f}, dof {dof}, p = {p:.4f}")
print(f"smallest expected cell: {expected.min():.1f}")
```

```
chisq.test(table(points$admin2, points$functional_status))
```

Vérifiez les hypothèses, et dites que vous l'avez fait

- Indépendance — Tous les tests ci-dessus supposent que chaque ligne est une observation indépendante
- Effectifs attendus pour le khi-deux — Le test devient peu fiable quand une cellule attendue passe sous cinq environ
- Égalité des variances pour le test t — Ne la vérifiez pas — employez toujours le test t de Welch, qui ne la suppose pas

Vérifiez les hypothèses, et dites que vous l'avez fait — En Python

```
girls_rate, boys_rate = girls.mean(), boys.mean()
print("assumptions checked:")
print(f"  independence: one row per student, no student in two rows -- verified")
print(f"  sample sizes: {len(boys)} and {len(girls)}, both well above 30")
print(f"  expected counts: smallest is {min(len(boys), len(girls)) * min(girls_rate, 1 -
    boys_rate):.0f}")
```

Vérifiez les hypothèses, et dites que vous l'avez fait — En R

Write the check into the script, not into your memory of having done it.

Rapportez la différence, non les deux nombres — Exemple (suite)

Over-age enrolment by sex, primary, 2024

Boys 39.2% n = 510

Girls 33.8% n = 542

Difference +5.5 points, 95% CI -0.4 to +11.3, $p = 0.066$

The interval includes zero. This survey cannot establish whether over-age enrolment differs by sex; if it does, the gap is between 0 and 11 points and boys are the disadvantaged group. Not reported as a finding.

Referral completion by disability status, 1,638 consenting cases

Disability reported 26.7% n = 202

Not reported 46.2% n = 1,436

Difference -19.4 points, 95% CI -26.1 to -12.8, $p < 0.001$

Rapportez la différence, non les deux nombres — Exemple (suite)

The smallest gap consistent with the data is 12.8 points. Reported as a finding, and the pathway lesson locates it at referral-making and acceptance rather than at consent.

Rapportez la différence, non les deux nombres

- Les deux blocs mettent la différence et son intervalle en titre — avec les deux valeurs de groupe en dessous

- L'un de ces écarts est statistiquement significatif.

Lire la leçon complète, avec le code exécutable

`https://data-analysis.cassion.dev/fr/cours/
applied-statistics-for-programmes/stat-03-two-gaps-two-answers`