
Bloc Tranversal (T). Recollida de dades

NOTA: Abans de llegir aquest document, has d'haver llegit el document: ***00_Elecció del tema.pdf***

La recollida de dades és una part essencial del procés científic i s'ha de fer de forma apropiada.

Quines característiques han de tenir les dades recollides?

Han de ser una mostra aleatòria simple MAS de la població d'estudi. Això vol dir que cada element o combinació d'elements de la població que estudieu ha de tenir la mateixa probabilitat de ser escollit.

Bons exemples:

- Si es fa un estudi sobre alguna característica dels alumnes de la FIB, tenir un cens dels alumnes permetria escollir els alumnes de la mostra de forma aleatòria amb un generador de nombres aleatoris
- Si es fa un estudi de velocitat de càrrega de pàgines web, aquestes s'haurien d'escollir de forma aleatòria per exemple emprant: https://random-website.com/

Mals exemples:

- Si es fa un estudi sobre alguna característica dels alumnes de la FIB, escollir els meus amics o els estudiants de la meva classe.
- Si es fa un estudi de velocitat de càrrega de pàgines web, escollir les darreres planes web que he visitat.

Quin format ha de tenir el fitxer on emmagatzemi les dades?

Qualsevol format de fitxer que R pugui llegir. Podeu consultar aquí alguns dels formats més habituals: Excel, text, csv...

Quin format/estructura han de tenir les dades?

Qualsevol format/estructura que R pugui llegir. En principi, recomanem que tinguin format de taula plana amb valors organitzats en files (observacions/registres) i columnes (variables/característiques). Procureu evitar l'ús de símbols ("*","?","!", accents, etc...) en les dades per evitar problemes de lectura.

Quants registres/observacions s'han de recollir?

No hi ha un nombre definit. Es recomana un **mínim de 30** i el màxim només ve limitat pel temps que li heu de dedicar a la recollida de dades.

Quant temps hauríem d'invertir en recollir les dades?

No hi ha un temps fixat, però el temps no hauria de ser excessiu per deixar temps per l'anàlisi de les dades. **La programació de l'activitat, en general, no permet més de 3 setmanes**.

Què vol dir que les dades són independents o aparellades?

Les dades són **aparellades** quan la unitat experimental és la mateixa sota les categories de la variable X. P.ex: mesuro la velocitat de càrrega cada pàgina web X amb Chrome i també amb Firefox.

Les dades són **independents** quan les unitat experimentals són diferents sota les categories de la variable X. P.ex: mesuro la velocitat de càrrega de unes planes web determinades amb Chrome i després mesuro la velocitat de càrrega de unes altres planes web diferents amb Firefox.

Sempre que sigui viable, es farà un disseny amb dades aparellades, perquè és més eficient des del punt de vista estadístic, però no sempre és viable: p.ex, al comparar hores d'estudi entre homes i dones.

Com podria organitzar l'estructura de les dades recollides?

Si són aparellades, les dades hauran de tenir l'estructura de l'esquerra i si són independents la de la dreta. NOTA: en les dades de la dreta, no necessàriament, les observacions han d'estar endreçades segons els valors de la variable X, poden alternar-se o estar en qualsevol ordre.

Aparellades

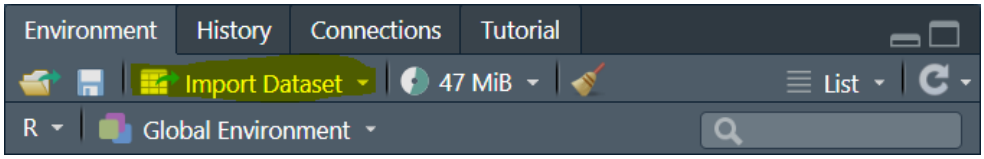
Firefox	Chrome	Z_1	Z_2
0.8	1.44	0.92	2.08
0.8	4	3.52	1.4
0.48	3.72	0.28	0.04
3.4	3.84	2.04	1.56
3.4	0.36	0.88	1
1.88	1.68	2.88	2.72
2.32	0.12	0.16	2.04
1.24	2	0.32	0.04
0.8	3.2	1.96	2.88

Independents

X	Y	Z_1	Z_2
Firefox	2.84	0.88	1.24
Firefox	0.12	3.04	2.72
Firefox	1.12	0.6	1.64
Firefox	2.28	2.84	2.64
Firefox	0.96	2.4	1.8
Firefox	0.04	0.32	3.2
Firefox	1.48	2.2	1.88
Firefox	1.4	3.2	2.36
Firefox	3.36	3.12	1.28
Chrome	3.32	3.44	2.12
Chrome	1.56	2.6	3.56
Chrome	1.12	0.2	2.52
Chrome	0.96	0.08	3.04
Chrome	0.68	2.16	3.84
Chrome	2.32	3.2	1.32
Chrome	1	2	3.68
Chrome	0.84	1.76	3.48
Chrome	2.76	2.6	2.52

Com importo les dades en la memòria del R?

Emprant R Studio és senzill importar les dades per menú:



No obstant, et recomanem que les **importis emprant comandes** perquè estalviaràs temps si les has d'importar moltes vegades. Hi ha molts tutorials de com fer-ho.

Quins són els errors més comuns a l'hora de recollir les dades?

- No fer cas dels consells mencionats en les preguntes
- Deixar files o columnes en blanc en el fitxer
- Incloure símbols estranys o espais en el nom de les variables
- Considerar dades aparellades quan són independents o viceversa
- No tenir una MAS o una mostra representativa de la meva població