

Tema V: Representación tabular.

Principios, errores comunes y herramientas en R para análisis reproducibles

Dr. Maicel Monzón, MSc. | ENSAP & CECMED

Pre-procesamiento de datos: El 80% del tiempo en análisis

- Aprendimos un conjunto de bibliotecas del “universo ordenado” pueden ser útiles para el pre-procesamiento de datos.
 - readr (importar)
 - tidyr (ordenar)
 - dplyr (transformar)

La Crisis de Reproducibilidad

- La **reproducibilidad** es crucial en la investigación.
- `gtsummary` facilita la generación de **informes reproducibles**.

Introducción (Qué aprenderemos en esta clase ?)

En esta conferencia los estudiantes aprenderán a **seleccionar, confeccionar en R y analizar** un cuadro o tabla estadística.

Sumario

- Concepto de **tabla estadística** y sus **partes**.
- **Clasificación** de las tabla
- Elección del **tipos** de tablas en función del **número** y **tipo** de variables.
- **Errores** más comunes en la confección de una tabla estadística.
- Biblioteca **gtsummary** y sus principales **funciones**.
- Casos de uso de visualización de datos de fortificación de alimentos a gran escala.

Cuadro o tabla estadística

Una tabla estadística es un recurso que emplea la **Estadística** con el fin de **presentar información resumida**, organizada por **filas y columnas**.

Partes de una tabla estadística

- Presentación (Identificación y Título)
- Cuerpo
- Fuente
- Notas explicativas

Presentación (Identificación y Título)

- **Identificación** (Número consecutivo según orden en el trabajo 1..n)
- **Título:** completo y conciso
 - Qué?
 - Cómo?
 - Dónde?
 - Cuándo?

Qué?

Dónde?

Tabla 1. Distribución de las defunciones de Cuba ocurridas durante el año 2012 clasificadas según sexo y edad de los fallecidos.

Cómo?

Cuándo?

Tabla 1. Defunciones por edad y sexo. Cuba 2012.

Cuerpo de la tabla

Nombre de la variable (Columna Matriz)	Fila de los encabezamientos		Total
	Celdas		
A			
B			
C			
D			
E			
Total			TOTAL

Fuente y Notas aclaratorias o explicativas.

- Las **notas** se colocan al pie de la tabla si es necesario

Ej. Inclusión, omisión, tipo de información (definitiva o provisional)

- **Fuente:** de dónde se obtuvo la información

INCIDENCIA DE CÁNCER SEGÚN SEXO. PRINCIPALES LOCALIZACIONES CUBA 2013

LOCALIZACIÓN	SEXO MASCULINO			SEXO FEMENINO			
	No.	Tasa* Bruta	Tasa** Ajustada		No.	Tasa* Bruta	Tasa** Ajustada
PULMON	3042	54.1	41.3	MAMA FEMENINA	2573	45.8	33.9
PIEL	2994	53.2	39.4	PIEL	2443	43.5	30.5
PROSTATA	2422	43.1	28.3	CUELLO DE UTERO	1512	26.9	19.2
LARINGE	765	13.6	10.8	PULMON	1403	25.0	18.1
COLON	757	13.5	9.7	COLON	1035	18.4	11.8
VEJIGA	586	10.4	7.5	CUERPO DE UTERO	489	8.7	6.5
BOCA	581	10.3	8.0	SISTEMA HEMATOPOY	478	8.5	7.0
ESTOMAGO	540	9.6	7.3	OVARIO	383	6.8	5.4
SISTEMA HEMATOPOY	493	8.8	7.5	PANCREAS	333	5.9	3.9
GANGLIOS LINFATICOS	431	7.7	6.5	HIGADO	308	5.5	3.7
TOTAL***	15535	276.2	206.7	TOTAL**	14301	254.4	183.7

* Tasa por 100 000 habitantes

** A la población mundial

*** Todas las localizaciones

Fuente: Registro Nacional del Cáncer. INOR

Las tablas se clasifican según el número de variables que representan

- Unidimensionales
- Bidimensionales
- Multidimensionales

Tabla Unidimensional

Tabla Unidimensional

**Pacientes de dengue según temperatura corporal.
Sala A. Hospital X. Enero – Junio 2012**

Temp	Enfermos	Porcentaje
37	3	7,30
38	18	43,90
39	16	39,02
40	4	9,80
Total	41	100,00

Tabla Bidimensional

Tabla Bidimensional

**Tabla 1. Distribución de niños según raza y sexo.
Escuela X. Municipio Z. 2012.**

Raza	Sexo			
	Masculino		Femenino	
	No.	%	No.	%
Blanca	79	42.9	60	51.7
Negra	63	34.2	34	29.3
Mestiza	42	22.8	22	19.0
Total	184	100.0	116	100.0

Fuente: Libro de matrícula de la escuela X.

Nota: Se excluyen 6 niños en los que no se clasificó la raza.

Tabla Tridimensional

Tabla Tridimensional

INCIDENCIA DE CÁNCER DE 15 A 44 AÑOS.

PRINCIPALES LOCALIZACIONES SEGUN GRUPOS DE EDAD Y SEXO. 2013

LOCALIZACIÓN	15 - 19		20 - 24		25 - 29		30 - 34		35 - 39		40 - 44	
	F	M	F	M	F	M	F	M	F	M	F	M
PULMON												
PIEL												
PROSTATA												
LARINGE												
COLON												
VEJIGA												
BOCA												
ESTOMAGO												
SISTEMA HEMATOPOY.												
GANGLIOS LINFATICOS												
TOTAL*												

* Se excluyen 108 casos con edad desconocida

Fuente: Registro Nacional del Cáncer. INOR

Tres errores que invalidan tus tablas (y cómo evitarlos)

1. Errores en la **presentación**.
2. Errores en el **cuerpo** de la tabla.
3. Errores en la **fuentes**.

Errores en la presentación.

- Cuadros **sin identificación**.
- Título o encabezamiento incorrecto o inadecuado: (ej: Telegráfico, no claro, ampuloso, demasiado extenso)

Errores en el cuerpo de la tabla

- Errores de **cálculo**.
- **Disposición incorrecta** de los datos.
- Se muestran **solamente medidas relativas ó de resumen**.
- Cuadros **sobrecargados**.
- No se especifican **unidades de medida**.

Errores en la fuente y las notas explicativas

- **Citar** la fuente **incorrectamente**. (No debe citarse la encuesta o ficha de vaciamiento del autor)
- **No citar** la fuente
- Consignar como fuente aquello que **no es un documento**. (oficinas, departamentos, centros, otros)
- No utilizar notas **cuando son necesarias**.

Consejos para la lectura y análisis de una tabla:

- Leer cuidadosamente el **título**.
- Leer las **notas aclaratorias**.
- Informarse sobre las **unidades de medidas** utilizadas.
- Prestar atención a los **totales**
- **Relacionar los totales** con los valores de las categorías de las variables en cada celda.
- Relacionar entre sí los valores de las variables estudiadas.

Biblioteca gtsummary

Es una **biblioteca** que está diseñada para **generar tablas con medidas resúmenes a partir de dataframe**, realizar **asociaciones** entre variables y **otros análisis** estadísticos (como regresiones) y formatearlos **listo para la publicación**.

Funciones básicas de la biblioteca gtsummary

- Resumir **estadísticas descriptivas** con `(tbl_summary)`
- **Tablas de contingencia** con `(tbl_cross)`
- Resumir **modelos de regresión** con `(tbl_regression)`
- ...

tbl_summary (argumento *include* para seleccionar variables)

```
library(gtsummary)
trial %>%
tbl_summary(include = c("age", "grade", "response"))
```

Characteristic	N = 200 ¹
Age	47 (38, 57)
Unknown	11
Grade	
I	68 (34%)
II	68 (34%)
III	64 (32%)
Tumor Response	61 (32%)
Unknown	7

¹Median (Q1, Q3); n (%)

tbl_summary (argumento *by* para estratificar)

```
library(gtsummary)
trial %>%
tbl_summary(by = "trt", include = c("age", "grade", "response"))
```

Characteristic	Drug A N = 98 ¹	Drug B N = 102 ¹
Age	46 (37, 60)	48 (39, 56)
Unknown	7	4
Grade		
I	35 (36%)	33 (32%)
II	32 (33%)	36 (35%)
III	31 (32%)	33 (32%)
Tumor Response	28 (29%)	33 (34%)
Unknown	3	4

¹Median (Q1, Q3); n (%)

tbl_summary (funcion *add_p()* para comparaciones con valores de p)

```
trial %>%  
tbl_summary(by = "trt", include = c("age", "grade", "response")) %>%  
add_p()
```

Characteristic	Drug A N = 98 ¹	Drug B N = 102 ¹	p-value ²
Age	46 (37, 60)	48 (39, 56)	0.7
Unknown	7	4	
Grade			0.9
I	35 (36%)	33 (32%)	
II	32 (33%)	36 (35%)	
III	31 (32%)	33 (32%)	
Tumor Response	28 (29%)	33 (34%)	0.5
Unknown	3	4	

¹Median (Q1, Q3); n (%)

tbl_summary (funcion *add_ci()*) para estimación por intervalos de confianza

```
trial %>%  
tbl_summary(by = "trt", include = c("age", "grade", "response")) %>%  
add_ci()
```

Characteristic	Drug A N = 98 ¹	95% CI	Drug B N = 102 ¹	95% CI
Age	46 (37, 60)	44, 50	48 (39, 56)	45, 50
Unknown	7		4	
Grade				
I	35 (36%)	26%, 46%	33 (32%)	24%, 42%
II	32 (33%)	24%, 43%	36 (35%)	26%, 45%
III	31 (32%)	23%, 42%	33 (32%)	24%, 42%
Tumor Response	28 (29%)	21%, 40%	33 (34%)	25%, 44%
Unknown	3		4	

¹Median (Q1, Q3): n (%)

tbl_summary (funcion *add_overall()* para añadir totales por columna

```
trial %>%  
tbl_summary(by = "trt", include = c("age", "grade", "response")) %>%  
add_overall()
```

Characteristic	Overall N = 200 ¹	Drug A N = 98 ¹	Drug B N = 102 ¹
Age	47 (38, 57)	46 (37, 60)	48 (39, 56)
Unknown	11	7	4
Grade			
I	68 (34%)	35 (36%)	33 (32%)
II	68 (34%)	32 (33%)	36 (35%)
III	64 (32%)	31 (32%)	33 (32%)
Tumor Response	61 (32%)	28 (29%)	33 (34%)
Unknown	7	3	4

¹Median (Q1, Q3): n (%)

Etiquetas personalizadas con argumento label

```
# Etiquetas personalizadas con argumento label
trial %>%
tbl_summary(by = "trt",
            include = c("age", "grade", "response"),
            label = list(age = "edad, años")
            ) %>%
add_overall()
```

Characteristic	Overall N = 200 ¹	Drug A N = 98 ¹	Drug B N = 102 ¹
edad, años	47 (38, 57)	46 (37, 60)	48 (39, 56)
Unknown	11	7	4
Grade			
I	68 (34%)	35 (36%)	33 (32%)
II	68 (34%)	32 (33%)	36 (35%)

Tablas de contingencia

Estas tablas permiten:

- Analizar la **asociación o independencia** entre variables
- **Calcular probabilidades** conjuntas, marginales o condicionales
- Evaluar si una variable explica o **influye** en otra (análisis bivariado)

Tablas de contingencia (tbl_cross)

```
trial %>%  
  tbl_cross(  
    row = stage, # variable en filas  
    col = trt,   # variable en columnas  
    percent = "cell" # cálculo del porcentaje  
  ) %>%  
  add_p()
```

	Chemotherapy Treatment		Total	p-value ¹
	Drug A	Drug B		
T Stage				0.9
T1	28 (14%)	25 (13%)	53 (27%)	
T2	25 (13%)	29 (15%)	54 (27%)	
T3	22 (11%)	31 (16%)	53 (28%)	

Resumir modelos de regresión

```
# ajusto el modelo de regresión
m1 <- glm(response ~ trt + grade + age, # formula
          data = trial, # datos
          family = "binomial" # regresión logística
        )
```


Resumir modelos de regresión (tbl_regression)

```
# Presenta los resultados de la regresión  
m1 %>%  
  tbl_regression(exponentiate = TRUE)
```

Characteristic	OR	95% CI	p-value
Chemotherapy Treatment			
Drug A	—	—	
Drug B	1.13	0.60, 2.13	0.7
Grade			
I	—	—	
II	0.85	0.39, 1.85	0.7
III	1.01	0.47, 2.15	>0.9
Age	1.02	1.00, 1.04	0.10

Abbreviations: CI = Confidence Interval OR = Odds Ratio

Conclusiones

En la conferencia estudiamos:

- Principios de tablas estadísticas: estructura, clasificación y errores comunes.
- Herramientas en R (gtsummary, dplyr, tidyr) para análisis reproducibles.

FIN

“Sin datos, solo eres otra persona con una opinión”

– W. Edwards Deming