Descriptive tool methodology

1. Crude rate (CR)[1–3]

- X_T: Number of events
- N_T: Population at risk

$$CR = \frac{X_T}{N_T} \cdot 100000 \text{ person-year}$$

CR confidence intervals[4]:

Two methods are used depending on the number of cases, one when the number of cases is less or equal to 30 and another when it is greater than 30.

In the first case, it is used the "exact" method by the factors of the 95% confidence levels for the estimations of a Poisson distribution variable (Table 2)

Table 2. Factors of the 95% confidence levels for the estimations of a Poisson distribution variable

Observed cases	Lower limit factor	Upper limit factor
1	0.0253	5.57
2	0.121	3.61
3	0.206	2.92
4	0.272	2.56
5	0.324	2.33
6	0.367	2.18
7	0.401	2.06
8	0.431	1.97
9	0.458	1.90
10	0.480	1.84
11	0.499	1.79
12	0.517	1.75
13	0.532	1.71
14	0.546	1.68
15	0.560	1.65
16	0.572	1.62
17	0.583	1.60
18	0.593	1.58
19	0.602	1.56
20	0.611	1.54
21	0.619	1.53
22	0.627	1.51
23	0.634	1.50
24	0.641	1.48
25	0.647	1.48
26	0.653	1.47
27	0.659	1.46
28	0.665	1.45
29	0.670	1.44
30	0.675	1.43

Then, the 95% confidence interval is constructed with the lower and upper limits as follows:

$$\text{Lower limit} = \text{CR x lower limit factor}$$
$$\text{Upper limit} = \text{CR x upper limit factor}$$

In the second case, when the number of cases is greater than 30, a normal approximation is used to calculate the CR confidence interval as follows:

$$CR \pm z_{\alpha/2}\sqrt{X_T}/N_T$$

Where $Z_{\alpha/2}$ is the Normal Distribution value with a confidence level of $\alpha/2$ (in our case, this value is 1.96 because a confidence level of 95% is used).

2. Specific Rate (SR) by age group[1–3]

- X_i: Number of events for the i-th age group where $\sum_{i=1}^{k} X_i = X_T$, i = 1, ..., k
- Ni: Population at risk for the i-th age group where $\sum_{i=1}^{k} N_i = N_T$, i = 1, ..., k

$$SR_i = \frac{X_i}{N_i} \cdot 100000 \text{ person-year}$$

3. Age Standardized Rate (ASR) [1–3,5,6]

- wi: weight of the reference population for the i-th age group for i=1,…k where $\sum_{i=1}^{k} w_i = 1$

$$ASR = \sum_{i=1}^{k} \frac{X_i w_i}{N_i} \cdot 100000 = \sum_{i=1}^{k} SR_i w_i \cdot 100000 \text{ person-year}$$

ASR confidence intervals[7]:

The method used is based on the Gamma distribution (Fay and Feuer, 1997).

The variance of ASR is:

$$v = \sum_{i}^{k} X_i \left(\frac{w_i}{N_i}\right)^2$$

The lower limit of the confidence interval is defined as:

$$\text{Lower limit} = \frac{v}{2ASR}(\chi^2)^{-1}_{\frac{2ASR^2}{v}}(\alpha/2)$$

And the upper limit as:

$$\text{Upper limit} = \frac{v + w_M^2}{2(ASR + w_M)}(\chi^2)^{-1}_{\frac{2(ASR+w_M)^2}{v+w_M^2}}(1-\alpha/2)$$

Where, $(\chi^2)_g^{-1}(p)$ is the p-quartile of a $\chi 2$ distribution with g degrees of freedom and $w_M = \max\left\{\frac{w_i}{N_i}\right\}$ is the maximum population reference weight.

4. Truncate Rate (TR) [1–3,8]

Let's suppose the age groups: i=1,...,k and G is a subgroup of these age groups: $G \subset \{1,2,\dots k\}$.

New weights are created based on G as: $w_i^* = \frac{w_i}{\sum_{i\in G} w_i}$ where $\sum_{i\in G} w_i^*=1$.

Then,

$$TR = \sum_i SR_i w_i^* \text{, on } i \in G$$

5. Cumulative Rate (CumulR)[1,6]

- i: i-th age group.
- a: Age group to accumulate, $a \in \{1,2,\dots k\}$
- g_i: Size of the i-th age group
- SR_i: Specific Rate of the i-th age group x 100000 person-year

$$CumulR = \sum_{i=1}^{a} \frac{g_i \cdot SR_i \cdot 100}{100000}$$

6. Cumulative Risk (CRisk)[1,3,6]

Once you have the CumulR, the CRisk is:

$$CRisk = 100 \cdot \left(1 - e^{\frac{-CumulR}{100}}\right)$$

7. Ratio between men and women or vice versa

ASR_{men} is defined as the ASR for men and ASR_{women} as the ASR for women, then the Ratio is calculated as follows:

$$Ratio = \frac{ASR_{men}}{ASR_{women}}$$

8. Estimated mean age in the interval (Age mean)[9]

To estimate the mean age of the individuals which are grouped by age, the following calculations are performed:

$$\overline{X} \approx \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

Where:

- x_i: class mark of the age group (middle point)
- k: number of age groups
- n_i: number of cases with x_i class mark
- n: total number of cases, $n = \sum_{i=1}^{k} n_i$

9. Estimated median age in the interval (Age median)[9]

To estimate the median age of the individuals which are grouped by age the following calculations are performed:

$$Md = LL(I_i) + \frac{a_i \cdot \left(\frac{n}{2} - N_{i-1}\right)}{n_i}$$

Where:

- I_i: the first interval which have a cumulative relative frequency greater than 50%
- LL(Ii) : lower limit of I_i
- a_i: I_i length
- n: number of cases
- N_{i-1}: Cumulative frequency in the interval before I_i interval. This frequency will be zero if I_i is the first interval
- Ni: absolute frequency in I_i

However, if N_{i-1} do not exist, we would be in the event that there are only values in the first age group, and then we should assume $N_{i-1} = 0$.

The previous formula is based on the assumption of a uniform distribution of frequencies throughout the interval. Although you can always calculate a median, you must be careful in the following two situations: i) the total number of cases is even and, ii) when an age group presents a cumulative relative frequency of 50% and the other groups (one or more) are null, therefore, the relative frequency accumulated continue to be 50%. In these cases we don't use the formula because it could lead to confusion. An example of this case is showed in Table 1.

Taula 1. Calculating the estimated median in an interval

I_i	n_i	N_i	f_i (%)	F_i (%)
0-4	2	2	50	50
5-9	0	2	0	50
10-14	0	2	0	50
15-19	0	2	0	50
20-24	0	2	0	50
25-29	0	2	0	50
30-34	0	2	0	50
35-39	0	2	0	50
40-44	0	2	0	50
45-49	0	2	0	50
50-54	0	2	0	50
55-59	0	2	0	50
60-64	0	2	0	50
65-69	1	3	25	75
70-74	1	4	25	100
75-79	0	4	0	100
80-84	0	4	0	100
85-89	0	4	0	100

I_i: age interval; **n_i:** number of cases in I_i; **N_i:** cumulated number of cases until I_i; **f_i:** relative frequency in I_i; **Fi:** cumulative relative frequency until I_i.

If we use the previous formula, the median calculation will be:

$$\mathrm{Md} = 65 + \frac{5 \cdot \left(\frac{4}{2} - 2\right)}{1} = 65$$

If we suppose ages of 2, 3, 66 and 74 years, the median is 34.5. It demonstrates that a correction in the formula must be applied.

The correction calculates the median by calculating the arithmetic mean between the upper limit of the age group that contains the first cumulative relative frequency of 50% and the lower limit of the first major strict accumulated relative frequency of 50%. Thus, following the example gets the average between 4 and 65 and then Md = 34.5, which is closer to the real median.

REFERENCES

1. Jensen O, Parkin D, MacLennan R, Muir C, Skeet R, editors. Cancer Registration: Principles and Methods. IARC Sci Publ. Lyon; 1991;95:1–288.

2. Esteve J, Benhamou E, Raymond L. Statistical Methods in Cancer Research. Volume. IV. Descriptive Epidemiology. IARC Sci Publ. Lyon; 1994;128:1–302.

3. Breslow NE, Day NE. Statistical methods in cancer research. Vol. II. The design and analysis of cohort studies. IARC Sci Publ. Lyon: International Agency for Research on Cancer; 1987;82:1–406.

4. Dos Santos Silva I. Cancer Epidemiology: Principles and Methods [Internet]. Lyon: International Agency for Research on Cancer; 1999. 442 p. Available from: http://www.iarc.fr/en/publications/pdfs-online/epi/cancerepi/

5. Rué M, Borrell C. Los métodos de estandarización de tasas. In: Porta M, Álvarez-Dardet C, Ruiz MT, Guardiola E, editors. Revisiones en salud pública, 3. Barcelona: Masson, S.A.; 1993. p. 263–95.

6. Bray F. Chapter 8. Age-standardization. In: Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB, editors. Cancer incidence in Five Continents Vol VIII IARC Scientific Publications No 155 [Internet]. International Agency for Research on Cancer; 2002. p. 87–9. Available from: http://www.iarc.fr/en/publications/pdfs-online/epi/sp155/ci5v8-chap8.pdf

7. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: A method based on the gamma distribution. Stat Med. 1997;16(7):791–801.

8. Doll R, Cook F. Summarizing indices for comparison of cancer incidence data. Int J Cancer. 1967;2(3):269–79.

9. Casa Aruta E. 200 problemas de estadística descriptiva. 5a ed. Barcelona: Vicens Vives; 1982. 206 p.