Interpreting and Presenting Statistical Analyses Using R

9/10/2020

Contents

This document indicates how the results we send to the researcher are displayed and interpreted. Guidance is also given regarding how to construct tables for reporting purposes. Firstly, the basic description of categorical and numerical variables will be discussed, thereafter comparisons/associations between two sets of variables.

Basic description of results

Categorical variables:

In the output of a frequency table below *agecat* indicates the age category of participants in the study — this variable was originally collected as a numerical variable but we have categorized it here for purposes of explanation. For each category, the number who belong to a specific category *freq* is indicated, and what percentage that is of those who had a response. All the 39 participants in the study provided an answer to this question (see last value in the column for cumulative frequency) and therefore no missing were recorded in this study (table 1)

Table 1: Frequency distribution for agecat

agecat	freq	percentage	cumulative frequency	cumulative percentage
1	25	64.1	25	64.1
2	7	17.9	32	82.0
3	3	7.7	35	89.7
4	1	2.6	36	92.3
5	1	2.6	37	94.9
6	2	5.1	39	100.0

Majority, 64.1% were $\leq$ 23 months (agecat = 1) and only 5.1% were $\geq$ 120 months (age_cat = 6). All percentages should be rounded to one decimal point. The cumulative frequency and cumulative percentage are usually not used except when one intends to group responses into categories.

When creating a table to report the results, you need to indicate what the codes stand for (see table 2):

The information of various categorical variables can be reported in one table, with a more generic title, and, at the end of each heading, an indication of the total number of responses for that variable (n = 39).

Table 2: Frequency distribution for age in months at different categories(n = 39)

agecat	n	percentage
0-23	25	64.1
24-47	7	17.9
48-71	3	7.7
72-95	1	2.6
96-119	1	2.6
120+	2	5.1

The results of sex of the respondents (f = Female, m = Male) in table 3, can be reported in a sentence and without creating a table in the report i.e. 64.1% of the participants in the study were males and 35.9% were females.

Table 3: Distribution of age in months by sex of the respondents

gender	freq	percentage	cumulative frequency	cumulative percentage
f	14	35.9	14	35.9
m	25	64.1	39	100.0

Numerical variables:

The distribution of any numerical variable (age in months) can be summarised as follows (table 4):

Table 4: Distribution of age in months of the respondents

N	Average	standard devaition	median value	lower quartile	upper quartile	minimum	maximum
39	23.74	36.3	4	1	33	0.9	156

where N is total number of observations analyzed. We note that age is skewed (i.e. mean is quantitatively different from median) and therefore we report about its distribution using the median and inter-quartile range, i.e. 4 months (IQR: 1, 33). Otherwise, if age was not skewed then we report the mean and standard deviation, i.e. 23.7 months (SD: 36.3). You don't have to include, in the results, table 4. The minimum and maximum are also shown, please check whether these extreme values are plausible and inclusion criteria are met. The frequency table of a numerical variable such as age in months, in its numerical form, can also be obtained (table 5):

Comparisons (Associations):

Categorical variables

Contingency tables

In the 2x2 contingency table 6 (cell totals) and table 7 (row percentages) below, Serum IgE test results (1 = Positive, 2 = Negative) is given in the columns, sex of the respondent (f = Female, m = Male) in the rows. 9 out of 14 females in the study (64.3%) had a positive Serum IgE test compared to 13 out of 25 males (52.0%). You can combine both table 6 and 7 into one table by including, in each cell, the cell total from table 6 and the corresponding total in table 7. For example, if you want to use row totals, you can put 9 (64.3%) in the cell corresponding to females with a positive serum IgE or if you want to use column totals, you can put 9 (40.9%) in the cell corresponding to positive serum IgE patients who are females. Remember to round percentages to one decimal. *See table 8 for corresponding column percentages for table 6*

Table 5: Frequency distribution for age in months

age_months	freq	percentage	cumulative frequency	cumulative percentage
0.9	2	5.13	2	5.1
1.0	11	28.21	13	33.3
2.0	4	10.26	17	43.6
3.0	2	5.13	19	48.7
4.0	1	2.56	20	51.3
5.0	1	2.56	21	53.9
6.0	1	2.56	22	56.4
11.0	1	2.56	23	59.0
16.0	1	2.56	24	61.5
23.0	1	2.56	25	64.1
24.0	3	7.69	28	71.8
30.0	1	2.56	29	74.3
36.0	3	7.69	32	82.0
48.0	1	2.56	33	84.6
60.0	2	5.13	35	89.7
84.0	1	2.56	36	92.3
96.0	1	2.56	37	94.8
120.0	1	2.56	38	97.4
156.0	1	2.56	39	100.0

Table 6: Sex of respondent by serum IgE

	1	2	Sum
f	9	5	14
m	13	12	25
Sum	22	17	39

Table 7: Row percentages for sex of respondent by serum IgE (for table 6)

	1	2
f	64.3	35.7
m	52.0	48.0

Table 8: column percentages for sex of respondent by serum IgE (for table 6)

	1	2
f	40.9	29.4
m	59.1	70.6

Whether the row or column percentage is used will not affect the results of the statistical test for association i.e. Chi-square test or Fisher's Exact test. The Chi-Square test is commonly used to compare two categorical variables in a 2 by 2 table. It is based on the null hypothesis that the two variable are independent (i.e. you can't predict the second variable if you know the results of the first one). From table 9, there was no evidence (p-value = 0.685) to suggest that there is a gender-effect on the serum IgE test results in the study (i.e., a patient's gender is independent or has no bearing on serum IgE results). Alternatively, you can report that the p-value > 0.05 therefore there was no statistically significant difference.

The Chi-square test is however not used if the expected counts/frequency (see table 10) < 5, and, in such a

Table 9: Statistical signifiance tests for sex of respondent by serum IgE

statistic	DF	Value	P.value
Chi-square	1	0.165	0.6850
Fisher's Exact		1.640	0.5178

case, we use the Fisher's exact test. However, in our example, the expected counts are all > 5 and so we can draw our conclusions about the gender vs serum IgE relationship using Chi-square. *Note that the fishers test is also interpreted the same as the Chi-square test in terms of statistical significance.*

Table 10: Expected cell counts for sex of respondent by serum IgE

	X1	X2
f	7.897436	6.102564
m	14.102564	10.897436

The results regarding the association between sex of the respondent and serum IgE can be represented as follows in a table with other variables as well (table 11).

Table 11: Results table

Variable	Positive.serum.IgE	Negative.serum.IgE	P_value
Sex of respondent (Female)	9 (64.3)	5 (35.7)	0.685

Numerical variables

Numerical variables compared in different categories

If the numerical variable has a skew distribution results will be summarised and compared (table 12)

Table 12: Distribution of age in months by sex of the respondent

gender	N	Median	Lower Quartile	Upper Quartile	Minimum	Maximum
f	14	14.5	1	54	0.9	156
m	25	3.0	1	24	0.9	84

We noted in table 4 that age is not normally distributed and for this reason we cannot use the mean or average to compare groups, but we use the median. We use the Kruskal-Wallis Test, which is a non parametric test, to compare the medians of the two groups (table 13).

Table 13: Statistical signifiance tests for sex of respondent by age in months

Statistic	Chi.Square	DF	P.value
Kruskal-Wallis Test	60.2716	1	0

Computed was the p-value ($P < 0.0001$) from the Kruskal-Wallis test (table 13). Also, the 95% Confidence interval for the difference in median age between the two gender groups (Male = m and Female = f) was calculated as [-1, 31] (table 14). These results can also be represented with the results of other numerical variables in a table as shown in table 14.

Table 14: The relatioship between sex of respondent by age in months

variable	Males.n.25.	Females.n.14.	X95..CI.for.median.diff.	P.value.Kruskal
Age in months	14.5 (1, 54)	3 (1, 24)	-1, 31	0

If the numerical variable has a symmetric distribution, results will be summarised and compared as shown in table 15.

Table 15: Age in months by gender distribution, if Age in month was symmetric

gender	N	Average	Std. deviation	Minimum	Maximum
f	14	38.21	51.06	0.9	156
m	25	15.64	22.00	0.9	84

Using the pooled variance estimate method, the variances between the two groups (Females vs males) were different (F Value = 5.38, p-value = 3.8×10^{-4}). A significant p-value for the F-test here shows that respondents age in months is, as expected, nonzero. The p-value (from the two sample t-test) and 95% confidence interval was calculated. The p-value from the t-test was not significant, which implies while that the mean age in months between the two groups is the same (note that the age in months of Males overlaps that of the Females). These results can also be represented with other numerical variables in a table as follows.

Table 16: The relationship sex of respondent by age in months (for symmetric age)

variable	Males.n.25.	Females.n.14.	X95..CI.for.mean.diff.	P.value
Age in months	38.21 (51.06, 0.9)	15.64 (22, 0.9)	-7.86, 53	0.1353