Practice on some Epidemiology toolboxes: pubh package

Yang Ge

2021 - February - 21, 11:02

Contents

1 Introduction

Package name: pubh package

Big thanks to Josie Athens [aut, cre], Frank Harell [ctb], John Fox [ctb], R-Core [ctb]. I read the vignettes and made notes for myself.

"In the case of epidemiology, there are already some good packages available for R, including: Epi, epibasix, epiDisplay, epiR and epitools. The pubh package does not intend to replace any of them, but to only provide a common syntax for the most frequent statistical analysis in epidemiology."

2 Syntax: the use of formulas

"The following table shows the most common names used in the literature to characterise variables in a cause-effect relationships"

Response variable	Explanatory variable(s)
Outcome	Exposure and confounders
Outcome	Predictors
Dependent variable	Independent variable(s)
y	x

In general, a epidemiological model would denoted as:

$$\text{Outcome} = f(\text{Exposure})$$

When writing articles, it is good to start with simple analysis before we go into multivariate models. Because "If you can't explain it simply, you don't understand it well enough. - Albert Einstein". One way to help other have a good understanding or easier start is to provide some plain results. For example, Table 1 (There is another perfect package called tableone), and stratification tables.

"One way to control for confounders is the use of stratification. In the ggformula package, one way of doing stratification is with a formula like:"

$$\text{Outcome} = f(\text{Exposure}|z) = y \sim x|z, data = data$$

To control on z, we can use multivariate analysis putting z as a covariable, or we can put z as a stratification variable. The difference usually related to sample size, because the stratification will be very mess if the z is unbalanced on sample size.

3 Some used packages

pubh

huxtable, amazing works on word.

jtools, easy peasy lemon squeezy on exploring

4 Descriptive statistics

mytable from the moonBook package was used in **pubh** here.

```
library(pubh)
library(sjlabelled)
library(tidyverse)
library(huxtable)
library(jtools)

data(Oncho)
Oncho %>% head()
```

id	mf	area	agegrp	sex	mfload	lesions
1	Infected	Savannah	20-39	Female	1	No
2	Infected	Rainforest	40+	Male	3	No
3	Infected	Savannah	40+	Female	1	No
4	Not-infected	Rainforest	20-39	Female	0	No
5	Not-infected	Savannah	40+	Female	0	No
6	Not-infected	Rainforest	20-39	Female	0	No

4.1 Two-by-two contingency tables

```
Oncho %>% mutate(mf = relevel(mf, ref = "Infected")) %>% # copy_labels(Oncho) %>%
cross_tab(mf ~ area) %>% theme_pubh() %>% add_footnote("Hello, footnote",
    font_size = 5)
```

	mf		
	Infected	**Not-infected**	**Total**
	(N=822)	(N=480)	(N=1302)
Residence			
- Savannah	281 (34.2%)	267 (55.6%)	548 (42.1%)
- Rainforest	541 (65.8%)	213 (44.4%)	754 (57.9%)

Hello, footnote

```
Oncho %>% select(-c(id, mfload)) %>% mutate(mf = relevel(mf,
    ref = "Infected")) %>% # copy_labels(Oncho) %>%
cross_tab(mf ~ area + .) %>% theme_pubh()
```

	mf		
	Infected	**Not-infected**	**Total**
	(N=822)	(N=480)	(N=1302)
Residence			
- Savannah	281 (34.2%)	267 (55.6%)	548 (42.1%)
- Rainforest	541 (65.8%)	213 (44.4%)	754 (57.9%)
Age group (years)			
- 5-9	46 (5.6%)	156 (32.5%)	202 (15.5%)
- 10-19	99 (12.0%)	119 (24.8%)	218 (16.7%)
- 20-39	299 (36.4%)	125 (26.0%)	424 (32.6%)
- 40+	378 (46.0%)	80 (16.7%)	458 (35.2%)
Sex			
- Male	426 (51.8%)	190 (39.6%)	616 (47.3%)
- Female	396 (48.2%)	290 (60.4%)	686 (52.7%)
Severe eye lesions?			
- No	640 (77.9%)	461 (96.0%)	1101 (84.6%)
- Yes	182 (22.1%)	19 (4.0%)	201 (15.4%)

4.2 Add stratification

```
data(Hodgkin)
Hodgkin <- Hodgkin %>% mutate(Ratio = CD4/CD8) %>% var_labels(Ratio = "CD4+ / CD8+ T-cells ratio")

Hodgkin %>% head()
```

```
Hodgkin %>% estat(~Ratio | Group) %>% as_hux() %>% theme_pubh()
```

```
Hodgkin %>% mutate(Group = relevel(Group, ref = "Hodgkin")) %>%
    copy_labels(Hodgkin) %>% cross_tab(Group ~ CD4 + ., method = 2,
    p_val = TRUE) %>% theme_pubh() %>% add_footnote("Values are medians with interquartile range.")
```

4.3 More

Because the **cross_tab** depended on **mytable_sub moonBook**

CD4	CD8	Group	Ratio
396	836	Hodgkin	0.474
568	978	Hodgkin	0.581
1212	1678	Hodgkin	0.722
171	212	Hodgkin	0.807
554	670	Hodgkin	0.827
1104	1335	Hodgkin	0.827

	Disease	N	Min.	Max.	Mean	Median	SD	CV
CD4+ / CD8+ T-cells ratio	Non-Hodgkin	20	1.1	3.49	2.12	2.15	0.73	0.34
	Hodgkin	20	0.47	3.82	1.5	1.19	0.91	0.61

5 Analysis on contingency tables

"The pubh package offers two wrappers to epiR functions".

1. "contingency calls epi.2by2 and it's used to analyse two by two contingency tables."
2. "diag_test calls epi.tests to compute statistics related with screening tests."

```
data(Bernard)
Bernard %>% head()
```

```
Bernard %>% mutate(fate = relevel(fate, ref = "Dead"), treat = relevel(treat,
    ref = "Ibuprofen")) %>% copy_labels(Bernard) %>% cross_tab(fate ~
    treat) %>% theme_pubh()
```

5.1 By epiR::epi.2by2

```
tab <- Bernard %>% mutate(fate = relevel(fate, ref = "Dead"),
    treat = relevel(treat, ref = "Ibuprofen"))

tab <- table(tab$treat, tab$fate)
tab
```

```
##
##             Dead Alive
##   Ibuprofen   84   140
##   Placebo     92   139
```

```
epiR::epi.2by2(tab)
```

```
##              Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +           84          140        224              37.5       0.600
## Exposed -           92          139        231              39.8       0.662
## Total              176          279        455              38.7       0.631
##
## Point estimates and 95% CIs:
```

	Disease			
	Hodgkin	**Non-Hodgkin**	**Total**	**p**
	(N=20)	(N=20)	(N=40)	
CD4+ T-cells	681.5 [396.5;1158.0]	433.0 [345.0;718.0]	528.5 [375.0;930.0]	0.081
CD8+ T-cells	447.5 [298.5;823.5]	231.5 [146.5;325.0]	319.0 [206.0;601.0]	0.001
CD4+ / CD8+ T-cells ratio	1.2 [0.8; 2.0]	2.2 [1.6; 2.7]	1.7 [1.1; 2.4]	0.007

Values are medians with interquartile range.

id	treat	race	fate	apache	o2del	followup	temp0	temp10
1	Placebo	White	Dead	27	539	50	35.2	36.6
2	Ibuprofen	African American	Alive	14		720	38.7	37.6
3	Placebo	African American	Dead	33	551	33	38.3	
4	Ibuprofen	White	Alive	3	1.38e+03	720	38.3	36.4
5	Placebo	White	Alive	5		720	38.6	37.6
6	Ibuprofen	White	Alive	13	1.52e+03	720	38.2	38.2

```
## ------------------------------------------------------------------
## Inc risk ratio                                 0.94 (0.75, 1.19)
## Odds ratio                                     0.91 (0.62, 1.32)
## Attrib risk *                                  -2.33 (-11.27, 6.62)
## Attrib risk in population *                    -1.15 (-8.88, 6.59)
## Attrib fraction in exposed (%)                 -6.20 (-33.90, 15.76)
## Attrib fraction in population (%)              -2.96 (-15.01, 7.82)
## ------------------------------------------------------------------
##  Test that OR = 1: chi2(1) = 0.260 Pr>chi2 = 0.61
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

Little explanations

1. Risk of death in Ibuprofen group: 84/224 = 0.375, similarly, risk of death in Placebo group: 92/231 = 0.3982684
2. Odds in Ibuprofen group: 84/140 = 0.6, similarly, in Placebo group: 92/139 = 0.6618705
3. RR = $\frac{84/224}{92/231}$ = 0.9415761, OR = $\frac{84/140}{92/139}$ = 0.9065217
4. Attrib risk = 92/231 - 84/224 = 0.0232684
5. Attrib risk in population = 176/455 - 92/231 = 0.0118132
6. Attrib fraction in exposed (%) = (92/231 - 84/224)/(84/224) = 0.0620491
7. Attrib fraction in population (%) = (176/455 - 92/231)/(176/455) = -0.0296143

	Mortality status		
	Dead	**Alive**	**Total**
	(N=176)	(N=279)	(N=455)
Treatment			
- Ibuprofen	84 (47.7%)	140 (50.2%)	224 (49.2%)
- Placebo	92 (52.3%)	139 (49.8%)	231 (50.8%)

5.2 By pubh::contingency

Same results but less code

```
Bernard %>% contingency(fate ~ treat)
```

```
##            Outcome
## Predictor   Dead Alive
##   Ibuprofen   84   140
##   Placebo     92   139
##
##              Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +            84          140        224              37.5       0.600
## Exposed -            92          139        231              39.8       0.662
## Total               176          279        455              38.7       0.631
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                0.94 (0.75, 1.19)
## Odds ratio                                    0.91 (0.62, 1.32)
## Attrib risk *                                 -2.33 (-11.27, 6.62)
## Attrib risk in population *                   -1.15 (-8.88, 6.59)
## Attrib fraction in exposed (%)                -6.20 (-33.90, 15.76)
## Attrib fraction in population (%)             -2.96 (-15.01, 7.82)
## -------------------------------------------------------------------
##  Test that OR = 1: chi2(1) = 0.260 Pr>chi2 = 0.61
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
##
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dat
## X-squared = 0.17076, df = 1, p-value = 0.6794
```

"Advantages of contingency:"

1. "Easier input without the need to create the table."

2. "Displays the standard epidemiological table at the start of the output. This aids to check what are the reference levels on each category."

3. "In the case that the χ^2-test is not appropriate, contingency would show the results of the Fisher exact

test at the end of the output."

5.3 Repeated by GLM

The contingency table results should be same to GLM univariate model, here I proved as follow (Ref[1])

```
mod_logit <- glm(fate ~ treat, data = Bernard, family = binomial(link = "logit"))
# summary(mod_logit)
```

Robust standard errors: ranging from "HC0" to "HC5". The authors of the sandwich package recommend "HC1" (if you set robust = TRUE). In Stata, the default is "HC1".

```
ftlogit <- jtools::summ(mod_logit, robust = "HC1", exp = TRUE,
    confint = TRUE, digits = 3)
# print(ftlogit)
```

```
Bernard2 <- Bernard %>% mutate(fate2 = if_else(as.character(fate) ==
    "Alive", 0, 1))
mod_log <- glm(fate2 ~ treat, data = Bernard2, family = poisson(link = "log"))
# summary(mod_log)
```

```
ftlog <- summ(mod_log, robust = "HC1", exp = TRUE, confint = TRUE,
    digits = 3)
# print(ftlog)
```

```
export_summs(mod_logit, mod_log, model.names = c("Logit-OR",
    "Log-RR"), coefs = c(Ibuprofen = "treatIbuprofen"), robust = "HC1",
    exp = TRUE, confint = TRUE, ci_level = 0.95, error_pos = c("right"),
    statistics = c(N = "nobs", AIC = "AIC", BIC = "BIC", Deviance = "deviance",
        D.F. = "df.residual"), error_format = "CI({conf.low}, {conf.high}), p = {p.value}")
```

	Logit-OR		Log-RR	
Ibuprofen	0.91	CI(0.62, 1.32), p = 0.61	0.94	CI(0.75, 1.19), p = 0.61
N	455		455	
AIC	610.98		690.18	
BIC	619.23		698.42	
Deviance	606.98		334.18	
D.F.	453.00		453.00	

Standard errors are heteroskedasticity robust. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

[1] https://cran.r-project.org/web/packages/jtools/vignettes/summ.html#Table_output_for_Word_and_RMarkdown_documents

6 Diagnostic tests

```
Freq <- c(1739, 8, 51, 22)
BCG <- gl(2, 1, 4, labels = c("Negative", "Positive"))
Xray <- gl(2, 2, labels = c("Negative", "Positive"))
tb <- data.frame(Freq, BCG, Xray)
tb
```

Freq	BCG	Xray
1.74e+03	Negative	Negative
8	Positive	Negative
51	Negative	Positive
22	Positive	Positive

```
tb <- expand_df(tb)
head(tb)
```

BCG	Xray
Negative	Negative
Negative	Negative
Negative	Negative
Negative	Negative
Negative	Negative
Negative	Negative

```
diag_test(BCG ~ Xray, data = tb)
```

```
##           Outcome +    Outcome -      Total
## Test +           22           51         73
## Test -            8         1739       1747
## Total            30         1790       1820
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------
## Apparent prevalence                   0.04 (0.03, 0.05)
## True prevalence                       0.02 (0.01, 0.02)
## Sensitivity                           0.73 (0.54, 0.88)
## Specificity                           0.97 (0.96, 0.98)
## Positive predictive value             0.30 (0.20, 0.42)
## Negative predictive value             1.00 (0.99, 1.00)
## Positive likelihood ratio             25.74 (18.21, 36.38)
## Negative likelihood ratio             0.27 (0.15, 0.50)
## -------------------------------------------------------
```

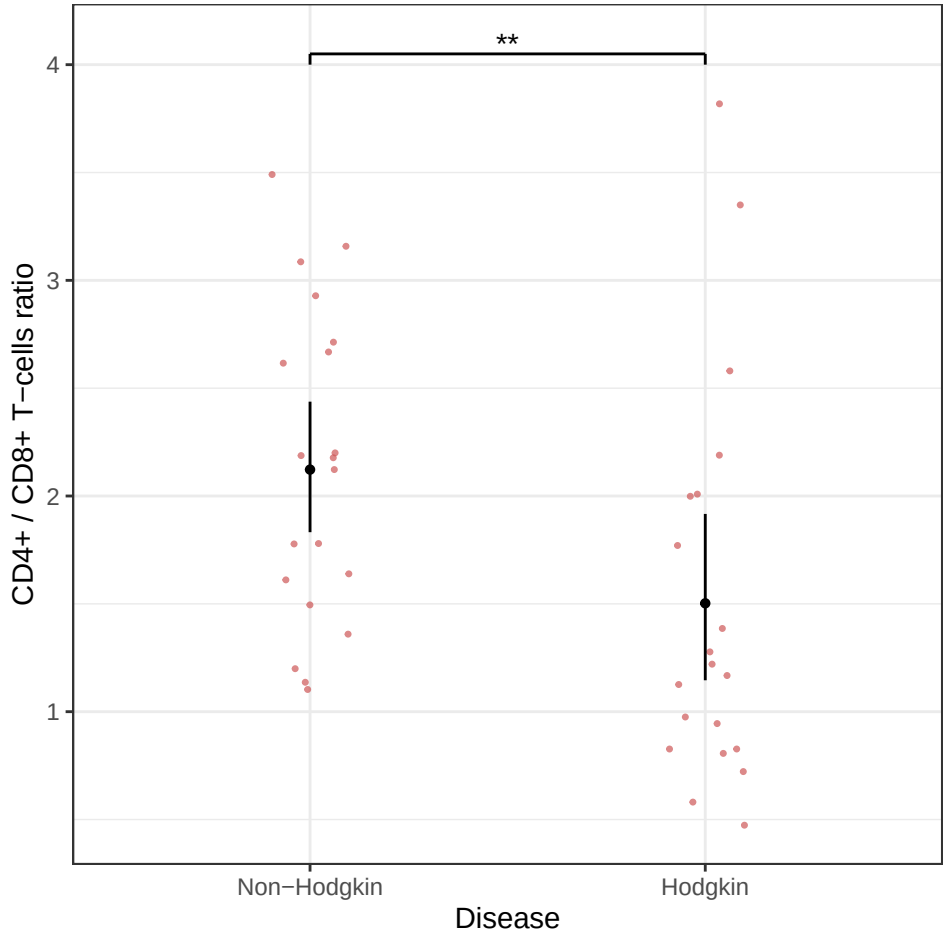
Little explanations

1. Apparent prevalence: 73/1820 = 0.0401099
2. True prevalence: 30/1820 = 0.0164835
3. Sensitivity = 22/30 = 0.7333333
4. Specificity = 1739/1790 = 0.9715084
5. Positive predictive value = 22/73 = 0.3013699
6. Negative predictive value = 1739/1747 = 0.9954207
7. Positive likelihood ratio = (22/30)/(51/1790) 25.7385621
8. Negative likelihood ratio = (8/30)/(1739/1790) 0.2744873

7 Little on graphical output

There are many kinds of function in **pubh**, but I generally prefer write my own ggplot codes which would be much more flexibility. However, some function like **gf_star** are interesting, and very useful when doing exploration no need perfect pretty plots.

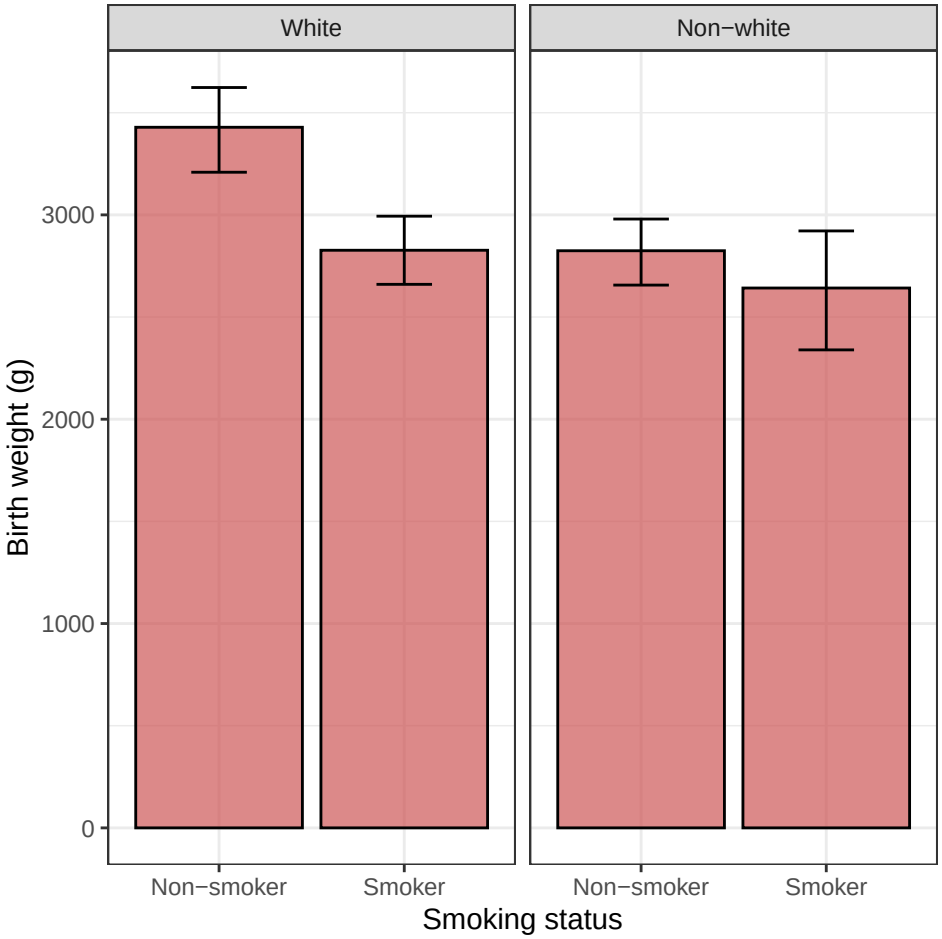
```
Hodgkin %>% strip_error(Ratio ~ Group) %>% axis_labs() %>% gf_star(x1 = 1,
    y1 = 4, x2 = 2, y2 = 4.05, y3 = 4.1, "**") + theme_bw()
```




```
data(birthwt, package = "MASS")

birthwt <- birthwt %>% mutate(smoke = factor(smoke, labels = c("Non-smoker",
    "Smoker")), Race = factor(race > 1, labels = c("White", "Non-white")),
    race = factor(race, labels = c("White", "Afican American",
        "Other"))) %>% var_labels(bwt = "Birth weight (g)", smoke = "Smoking status",
    race = "Race")
```

```
birthwt %>% bar_error(bwt ~ smoke | Race) %>% axis_labs() + theme_bw()
```


White
Non−white
3000
2000
1000
0
Birth weight (g)
Non−smoker
Smoker
Non−smoker
Smoker
Smoking status