MSCR 530 Analysis Exercise 1

Anish Shah

November 13, 2019

# Assignment Objectives

This analysis exercise is to use sample data (from the research study of the work of this author) to create a bivariate analysis and present this in the creation of tables.

The assignment objectives are as follows:

1. Select one variable to be the primary exposure of interest. Select one variable to be the primary outcome of interest.
2. Create a Table 1, which presents the distribution of the primary exposure variable by other participant characteristics in the dataset.

* For Table 1, the exposure variable should be a dichotomous variable (or transformed into a dichotomous variable).
* Table 1 should have five columns (participant characteristics, exposed, not exposed, total, p-value); the exposure variable being in the columns and the participant characteristics in the rows.
* There must be at least 5 participant characteristics (rows) in Table 1 including a minimum of 1 continuous variable. The primary outcome variable should also be one of the participant characteristics included as a row.

1. Create a Table 2, which presents the distribution of the primary outcome variable by other participant characteristics in the dataset.

* For Table 2, the outcome variable should be a dichotomous variable (or transformed into a dichotomous variable).
* Table 2 should have five columns; the outcome variable being in the columns and the participant characteristics in the rows.
* There must be at least 5 participant characteristics in Table 2 including a minimum of 1 continuous variable and the participant characteristics in Table 1. The primary exposure variable should also be one of the participant characteristics included as a row.

1. Answer questions on page 3 of this instruction document which relate to the tables and interpretation of the data. Be sure to read these questions prior to creating the tables. Answers should be typed.

# Table 1

## Using the compareGroup package  
## Demographic data and baseline characteristics  
df <- inner\_join(df\_twins, df\_wide, by = 'patid') %>%  
 subset(., hour == 7)  
  
# Format appropriate data from numerical into categories  
# Adding labels using the Hmisc package to the variables  
tmp <- within(df, {  
 patid <- factor(patid)  
 pair <- factor(pair)  
 Age <- age  
 BMI <- bmi\_number  
 Physical\_Activity <- baecke\_tot  
 FRS <- FRS\_T  
 Depression <- beck\_total  
  
 Smoking <- factor(smoking)  
 levels(Smoking) <- c("No", "Yes")  
  
 Hypertension <- factor(hptn)  
 levels(Hypertension) <- c("No", "Yes")  
  
 Diabetes <- factor(dm)  
 levels(Diabetes) <- c("No", "Yes")  
  
 Antidepressant <- factor(med\_antidepr)  
 levels(Antidepressant) <- c("No", "Yes")  
  
 Antilipidemic <- factor(med\_lipid\_abs)  
 levels(Antilipidemic) <- c("No", "Yes")  
  
 Antihyperglycemic <- factor(med\_diabetes\_mellitus)  
 levels(Antihyperglycemic) <- c("No", "Yes")  
  
 PTSD <- factor(ptsd)  
 levels(PTSD) <- c("No", "Yes")  
  
 Abnormal\_DYX <- factor(CP)  
 levels(Abnormal\_DYX) <- c("No", "Yes")  
  
 Abnormal\_Stress <- factor(PETdiff\_2)  
 levels(Abnormal\_Stress) <- c("No", "Yes")  
})  
  
tmp <- tmp[!is.na(tmp$CP),]  
  
# Comparison group of excluded patients  
cmpExp <- compareGroups(Abnormal\_DYX ~ Age + BMI + Smoking + Hypertension + Diabetes + PTSD + Depression + Antidepressant + Antilipidemic + Antihyperglycemic + Physical\_Activity + FRS + Abnormal\_Stress, data = as.data.frame(tmp))  
  
# Table  
cmpTableExposure <- createTable(cmpExp, show.n = TRUE, show.ratio = FALSE, show.p.overall = TRUE)  
  
# Export to markdown  
export2md(cmpTableExposure)

*Table 1. Description of patient characteristics by low and normal HRV*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Normal DYX  N=216 | Low DYX  N=38 | P-value | N |
| Age (SD) | 55.1 (3.17) | 55.6 (3.58) | 0.397 | 254 |
| BMI (SD) | 29.5 (4.73) | 28.9 (5.19) | 0.507 | 253 |
| Smoking: (%) |  |  | 0.213 | 254 |
| No | 89 (41.2%) | 11 (28.9%) |  |  |
| Yes | 127 (58.8%) | 27 (71.1%) |  |  |
| Hypertension: (%) |  |  | 0.075 | 254 |
| No | 159 (73.6%) | 22 (57.9%) |  |  |
| Yes | 57 (26.4%) | 16 (42.1%) |  |  |
| Diabetes: (%) |  |  | 0.127 | 254 |
| No | 199 (92.1%) | 32 (84.2%) |  |  |
| Yes | 17 (7.87%) | 6 (15.8%) |  |  |
| PTSD: (%) |  |  | 1.000 | 254 |
| No | 190 (88.0%) | 34 (89.5%) |  |  |
| Yes | 26 (12.0%) | 4 (10.5%) |  |  |
| Depression (SD) | 4.48 (5.89) | 7.47 (8.02) | 0.033 | 253 |
| Antidepressant: (%) |  |  | 0.074 | 254 |
| No | 187 (86.6%) | 28 (73.7%) |  |  |
| Yes | 29 (13.4%) | 10 (26.3%) |  |  |
| Antilipidemic: (%) |  |  | 0.277 | 254 |
| No | 215 (99.5%) | 37 (97.4%) |  |  |
| Yes | 1 (0.46%) | 1 (2.63%) |  |  |
| Antihyperglycemic: (%) |  |  | 0.715 | 254 |
| No | 203 (94.0%) | 35 (92.1%) |  |  |
| Yes | 13 (6.02%) | 3 (7.89%) |  |  |
| Physical Activity (SD) | 7.49 (1.62) | 7.14 (2.07) | 0.347 | 251 |
| FRS (SD) | 6.05 (2.27) | 6.21 (2.56) | 0.720 | 254 |
| Abnormal Stress: (%) |  |  | 0.372 | 254 |
| No | 176 (81.5%) | 28 (73.7%) |  |  |
| Yes | 40 (18.5%) | 10 (26.3%) |  |  |

*Table 1. We compared abnormal DYX using a cutpoint < 2.0, and compared the demographic and clinical covariates, as well as the outcome variable of stress myocardial perfusion imaging. FRS = Framingham risk score, PTSD = post-traumatic stress disorder, BMI = body mass index. Continuous normal variables were tested using t-test, and chi-square test (Fisher test when frequency < 5) for categorical variables.*

# Table 2

## Using the compareGroup package  
## Demographic data and baseline characteristics  
df <- inner\_join(df\_twins, df\_wide, by = 'patid') %>%  
 subset(., hour == 7)  
  
# Format appropriate data from numerical into categories  
# Adding labels using the Hmisc package to the variables  
tmp <- within(df, {  
 patid <- factor(patid)  
 pair <- factor(pair)  
 Age <- age  
 BMI <- bmi\_number  
 Physical\_Activity <- baecke\_tot  
 FRS <- FRS\_T  
 Depression <- beck\_total  
  
 Smoking <- factor(smoking)  
 levels(Smoking) <- c("No", "Yes")  
  
 Hypertension <- factor(hptn)  
 levels(Hypertension) <- c("No", "Yes")  
  
 Diabetes <- factor(dm)  
 levels(Diabetes) <- c("No", "Yes")  
  
 Antidepressant <- factor(med\_antidepr)  
 levels(Antidepressant) <- c("No", "Yes")  
   
 Antilipidemic <- factor(med\_lipid\_abs)  
 levels(Antilipidemic) <- c("No", "Yes")  
   
 Antihyperglycemic <- factor(med\_diabetes\_mellitus)  
 levels(Antihyperglycemic) <- c("No", "Yes")  
  
 PTSD <- factor(ptsd)  
 levels(PTSD) <- c("No", "Yes")  
  
 Abnormal\_DYX <- factor(CP)  
 levels(Abnormal\_DYX) <- c("No", "Yes")  
  
 Abnormal\_Stress <- factor(PETdiff\_2)  
 levels(Abnormal\_Stress) <- c("No", "Yes")  
})  
  
# Remove NA  
tmp <- tmp[!is.na(tmp$CP),]  
  
# Comparison group of excluded patients  
cmpOut <- compareGroups(Abnormal\_Stress ~ Age + BMI + Smoking + Hypertension + Diabetes + PTSD + Depression + Antidepressant + Antilipidemic + Antihyperglycemic + Physical\_Activity + FRS + Abnormal\_DYX, data = as.data.frame(tmp))  
  
# Table  
cmpTableOutcome <- createTable(cmpOut, show.n = TRUE, show.ratio = FALSE, show.p.overall = TRUE, show.ci = TRUE)  
  
# Export to markdown  
export2md(cmpTableOutcome)

*Table 2. Between-group differences of patient characteristics by abnormal myocardial perfusion imaging*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Normal Stress  N=204 | Abnormal Stress  N=50 | p-value | N |
| Age (SD) | 55.2 (3.33) | 55.0 (2.85) | 0.667 | 254 |
| BMI (SD) | 29.3 (4.74) | 29.9 (5.01) | 0.435 | 253 |
| Smoking: (%) |  |  | 0.792 | 254 |
| No | 79 (38.7%) | 21 (42.0%) |  |  |
| Yes | 125 (61.3%) | 29 (58.0%) |  |  |
| Hypertension: (%) |  |  | 0.458 | 254 |
| No | 148 (72.5%) | 33 (66.0%) |  |  |
| Yes | 56 (27.5%) | 17 (34.0%) |  |  |
| Diabetes: (%) |  |  | 0.785 | 254 |
| No | 186 (91.2%) | 45 (90.0%) |  |  |
| Yes | 18 (8.82%) | 5 (10.0%) |  |  |
| PTSD: (%) |  |  | 0.436 | 254 |
| No | 182 (89.2%) | 42 (84.0%) |  |  |
| Yes | 22 (10.8%) | 8 (16.0%) |  |  |
| Depression (SD) | 4.90 (6.41) | 5.04 (6.06) | 0.887 | 253 |
| Antidepressant: (%) |  |  | 0.606 | 254 |
| No | 171 (83.8%) | 44 (88.0%) |  |  |
| Yes | 33 (16.2%) | 6 (12.0%) |  |  |
| Antilipidemic: (%) |  |  | 1.000 | 254 |
| No | 202 (99.0%) | 50 (100%) |  |  |
| Yes | 2 (0.98%) | 0 (0.00%) |  |  |
| Antihyperglycemic: (%) |  |  | 0.528 | 254 |
| No | 192 (94.1%) | 46 (92.0%) |  |  |
| Yes | 12 (5.88%) | 4 (8.00%) |  |  |
| Physical Activity (SD) | 7.48 (1.67) | 7.24 (1.80) | 0.404 | 251 |
| FRS (SD) | 6.14 (2.23) | 5.80 (2.63) | 0.399 | 254 |
| Abnormal DYX: (%) |  |  | 0.372 | 254 |
| No | 176 (86.3%) | 40 (80.0%) |  |  |
| Yes | 28 (13.7%) | 10 (20.0%) |  |  |

*Table 2.. We compared patients with normal and abnormal myocardial perfusion. We compared stress test result groups to demographic and clinical covariates, as well as the exposure variable of normal versus abnormal DYX. FRS = Framingham risk score, PTSD = post-traumatic stress disorder, BMI = body mass index. Continuous normal variables were tested using t-test, and chi-square test (Fisher test when frequency < 5) for categorical variables.*

# Table Questions

## Question 1. Titles

Each table should have a descriptive title. List the titles for your Table 1 and Table 2 here:

**1A. Table 1 title:**

*Description of patient characteristics by low and normal HRV*

**1B. Table 2 title:**

*Between-group differences of patient characteristics by abnormal myocardial perfusion imaging*

## Question 2. Exposure table

**2A. What was the primary exposure variable chosen for Table 1?**

Low versus normal heart rate variability (HRV), measured by the non-linear HRV measure Dyx at the cutpoint of 2.0

**2B. How many participants were in each of the two exposure categories?**

Low = 216 Normal = 38

**2C. How were participants who had missing values of the primary exposure handled?**

The exposure was generated through ECG analysis mulitple steps prior. They were excluded from the sample population based on missing-values at initial data preparation. For this subset of hte study, we used a single hour of 24-hours of data to calculate measures. Durng this 1 hour, out of the 262 patients, 8 patients had missing HRV data. This group was excluded in the tables, leading to a total of 254. Roughly 3% data loss. They were kept in the larger data pool because other hours they were able to contribute, and in more complex analyses we used imputation to help generate values (time series analysis prior to HRV calculation from ECG data).

## Question 3. Outcome table

**3A. What was the primary outcome variable chosen for Table 2?**

Abnormal versus normal stress test, determined by myocardial perfusion imaging using adenosine stress.

**3B. How many participants were in each of the two outcome categories?**

Normal stress = 204 Abnormal stress = 50

**3C. How were participants who had missing values of the primary outcome handled?**

The patients that were unable to tolerate stress tests were excluded from the study initially. The purpose of the study was obtaining stress tests, thus that data was excluded from the study onset. In terms of HRV values, as above, the same population of 8 individuals were excluded for missingness during 1 hour of the 24 hour sample.

## Question 4. Continuous variable statistical testing

**4A. State a participant characteristic that was included as a continuous variable in Table 1.**

Age was a simple continous variable used in Table 1.

**4B. What statistical test was used to compare this continuous variable across exposure categories? Justify your decision to use this statistical test.**

We used a two-sample t-test because, although based on size alone we can suppose we have a normal distribution, the SD of population is not known, thus the sample SD has to be used. Thus, t-test.

**4C. What was the null hypothesis of the statistical test used?**

H0 = Mean age of those with normal stress tests is the same as that of those with abnormal stress tests.

**4D. Provide a correct interpretation of the p-value in one sentence.**

P-value of 0.667, with an alpha of 0.05, suggests that there is insufficient evidence to reject the null hypothesis.

## Question 5. Categorical variable statistical testing

**5A. State a participant characteristic that was included as a categorical variable in Table 2.**

PTSD was selected as a categorical variable in this study (presence or absence by clinical chart review).

**5B. What statistical test was used to compare this categorical variable across outcome categories? Justify your decision to use this statistical test.**

Because this was a categorical variable, the chi-squared test of independence was used. This test was appropriate since it looks for whether or not there is a significant difference between teh expected and observed frequences in multiple categories. Essentially, its testing if there is independence among the groups.

**5C. What was the null hypothesis of the statistical test used?**

H0 = There is not a diffence in the frequency of PTSD in patients with normal stress and in the frequency of PTSD in those with abnormal stress. (Could also say proportion instead of frequency)

**5D. Provide a correct interpretation of the p-value in one sentence.**

The p-value = 0.436, with alpha = 0.05, suggesting that there is not sufficient evidence to reject the null hypothesis such that the proportions between populations are similar (not independent).

## Question 6. Interpretation of findings

**6A. Using one sentence, describe the association between the primary exposure and primary outcome variables chosen. Report a measure of association and 95% confidence interval for the association.**

When considering the exposure as a continuous variable, abnormal/low Dyx (HRV measure) is associated with abnormal stress, via two-sample t-test (t = 2, df = 74, p = 0.04).

DYX is higher in patients with normal stress (0.395) than in abnormal stress (0.0392), such that there is a difference between the means of the populations based on a two-sample t-test (95% CI 0.0191 - 0.6935, t = 2, DF = 74, p = 0.04).

If considering off of an artificial cutpoint, as per Table 1, then there is no association between the exposure of low DYX and abnormal stress test by chi-square test of independence (X^2 = 0.798, p = 0.372), and a confidence interval can be generated with 2-sample test of proportions with continuity correction (95% confidence interval -0.0868 to 0.2427).

**6B. Which, if any, of the participant characteristics are likely confounding the relation between the primary exposure and outcome variables? Using one or two sentences, justify the answer with information generated in Table 1 and Table 2.**

Not enough is known about autonomic dysfunction to allow us to assess for confounding. PTSD and depression effect teh ANS centrally, while HTN and age effect ANS systemically, and integration at the level of the heart leads to the HRV measure. Depression would likely be the strongest candidate to be a confounder, if anything. However, based on teh DAG model, there would not be an unblocked open path from HRV to abnormal stress test.