Homework 03

Spencer Pease

10/21/2019

Overview of the Data

The data in this report comes from a study of the risk factors associated with cardiovascular and cerebrovascular disease among generally healthier older adults. Specifically, this study enrolled a cohort of older (65+) and generally healthy adults by randomly sampling individuals enrolled in the United States Medicare system. Overall **735** adults agreed to participate in the study.

This report focuses on five variables from the study data. They are briefly described below (see the study documentation for more details):

Table 1: Variable Descriptions

Variable	Units	Description
packyrs	pack years (1 pack of cigarettes per day for 1 year)	Smoking history
crt	mg/dl	Measure of creatinine in the participant's blood
male	Binary (0 = female, 1 = male)	Binary Indicator of patient sex
obstime	days	Total time the participant was observed on the study
death	binary	Indicator that the participant died during the study

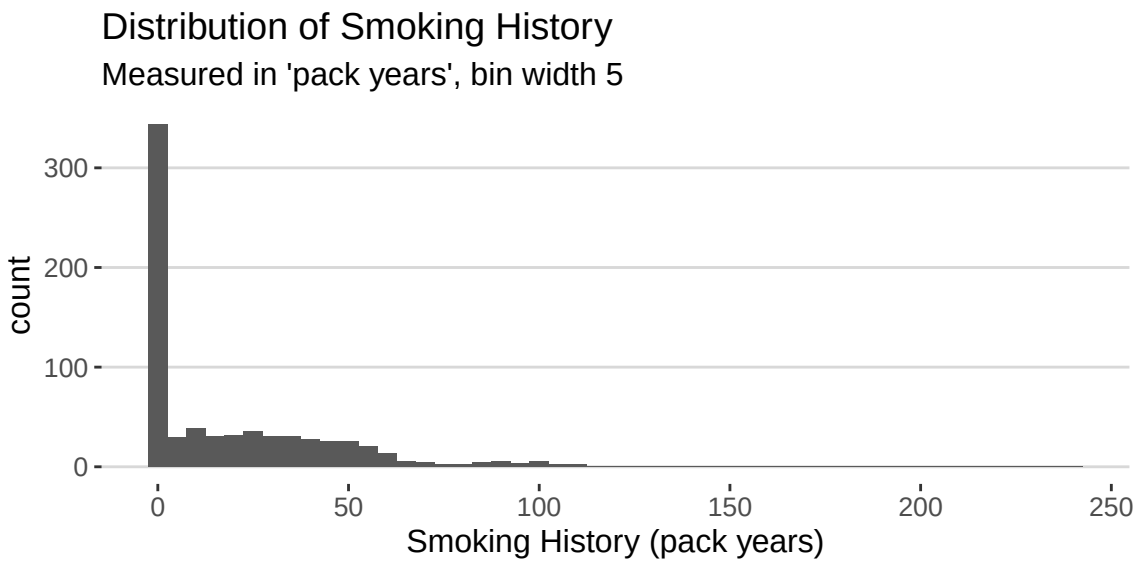
(*Q1*) Survival Status and Smoking History

In the first section of this report, we examine the relationship between smoking history (measured in pack years), and survival status at five years.

Descriptive Statistics

(*Q1.a*)

We start by examining the overall distribution of smoking history in our study participants. Notice that the most participants have between 0 and 5 pack years, while at least one participant has over 200 pack years of smoking history. This leads to the distribution having a strong **right skew**.



(***Q1.b,c***)

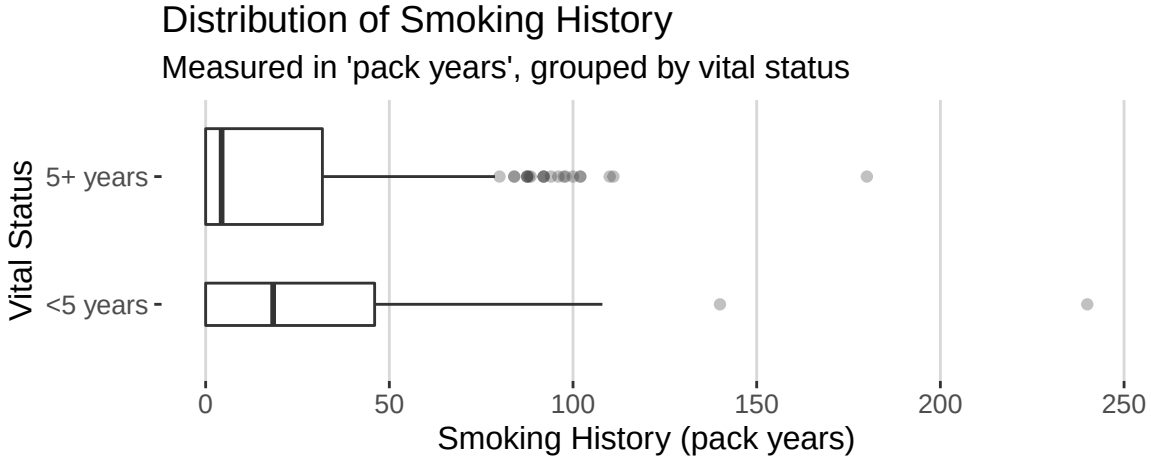
With an idea of the overall distribution of smoking history, we can now break our study participants into two groups - those who survived at least five years from the date of their MRI, and those who did not survive - and look at the descriptive summary statistics of smoking history within for each group.

Table 2: Summary of smoking history distribution (measured in pack years) by vital status group

Survival Status	valid	missing	mean	sd	min	q25	median	q75	max
<5 years	120	1	28.05	36.04	0	0	18.38	46.00	240
5+ years	614	0	17.95	24.69	0	0	4.35	31.79	180

(***Q1.d***)

We can also visualize these distributions using box-plots:



From the box-plots, we see that participants who survived at least five years generally had smoked fewer pack years than those who survived fewer than five years as indicated by the lower median and 75^{th} percentile. Despite these differences, both groups have a 25^{th} percentile of 0 pack years.

Inferential Statistics

(*Q1.e,f*)

Aside from examining the descriptive statistics of the participants, it is also useful to explore the population inferential statistics for populations fitting the two vital status groups. Below we examine the population point estimate and 95% confidence interval.

Table 3: Population inference of smoking history (measured in 'pack years') by vital status group

Survival Status	point est.	95% CI (lower)	95% CI (upper)
<5 years	28.05	21.6	34.49
5+ years	17.95	16.0	19.90

(*Q1.g*)

We can also test for significant differences between the two populations defined by vital status at five years. One way to do this is by testing the difference of two sample means, assuming that we know the population standard deviations are known and not necessarily equal. Below we calculate the point estimate and 95% confidence interval of the difference in population means under the assumption that there is no difference in means.

Table 4: Distribution of difference between sample means across vital status group populations for 'pack years'

Point est.	95% CI (lower)	95% CI (upper)
10.1	3.36	16.83

Since the confidence interval does not include 0, we can say with reasonable confidence that there is in fact a significant difference between our two populations.

(*Q1.h*)

Another way to explore any potential difference between the two populations is to not not assume that we know the population standard deviations while still allowing them to vary by population. This design can be captured by a *heteroscedastic t-test*, where we define our hypotheses as follows:

- H_0: $\mu_1 = \mu_2$ There **is not** a difference in the means between the two populations.
- H_a: $\mu_1 \neq \mu_2$ These **is** a difference in the means between the two populations.

Table 5: Two-sided heteroscedastic t-test results of smoking history by vital status group

95% CI (lower)	95% CI (upper)	p-value
3.3	16.892	0.004

The results of this test produce a *p-value* of **0.0039**. Since our test was at the $\alpha = .05$ level, this result indicates that there is a statistically significant difference in the means between the two populations, telling us that there is a relationship between smoking history and five-year vital status.

(*Q1.i*)

Both the *difference in mean pack years* and *heteroscedastic t-test* inferential tests indicate that there is an association with mean pack years of smoking and five-year vital status. This is seen in neither confidence interval including zero.

(*Q1.j*)

Using the same set of hypotheses, we can also perform a test for association under the assumption that we don't know the population standard deviations, but are treated as equal. This *homoscedastic t-test* produces the following results:

Table 6: Two-sided homoscedastic t-test results of smoking history by vital status group

95% CI (lower)	95% CI (upper)	p-value
4.83	15.36	0

The *p-value* for this test is 2×10^{-4}, which beats the threshold for significance set by our α level of .05. This further confirms that there is a significant relationship between smoking history and five-year vital status in our study population.

(*Q1.k*)

Both the *heteroscedastic* and *homoscedastic t-tests* assert that the null hypothesis should be rejected, and that there is a statistically significant association between smoking history and vital status at five years. *A priori*, We should prefer the heteroscedastic t-test, however, since we do not have any indication that the variance in the point estimate of the two populations would be the same. Assuming that the variances are the same when they in truth aren't hurts the validity of our tests.

(*Q2*) Patient Sex and Smoking History

In the this next section of the report, we assess if there is a significant difference in smoking history across groups defined by sex.

(Q2.a)

We can again look at the point estimate for the difference in mean pack years of smoking history between similar population groups defined by sex, as well as the confidence interval of this point estimate.

Table 7: test

Point est.	95% CI (lower)	95% CI (upper)
-11.02	-14.87	-7.17

Given that our confidence interval does not include zero, we can say there is an actual association between smoking history and patient sex that leads to a difference in mean pack years of smoking history between males and females.

(Q2.b)

Another way to test the relationship between smoking history and sex is the *two-sided heteroscedastic t-test*, which allows for the possibility of unequal variances across groups. For this test, we define our *null* and *alternative* hypotheses as:

- H_0: $\mu_1 = \mu_2$ There **is not** a difference in the mean pack years between the two populations.
- H_a: $\mu_1 \neq \mu_2$ These **is** a difference in the mean pack years between the two populations.

Table 8: Two-sided heteroscedastic t-test results of smoking history by sex group

95% CI (lower)	95% CI (upper)	p-value
-14.87	-7.16	0

The *p-value* for testing the null hypothesis is 3.02×10^{-8} (below the threshold set by $\alpha = .95$), which leads us to the conclusion that that we should reject the null hypothesis and that there is an association between smoking history and sex.

(Q2.c)

Both of these tests produce similar confidence intervals and assert that there is a statistically significant relationship between sex and smoking history.

(Q3) Survival Status and Creatinine Level

Lastly, we will examine the relationship between survival status at five years and creatinine level in our population.

(Q3.a)

Here we looks at the point estimate for the difference in mean pack years of creatinine level between similar population groups defined by survival status and the 95% confidence interval it produces.

Table 9: test

Point est.	95% CI (lower)	95% CI (upper)
0.18	0.09	0.27

From these results, we see that while the difference between populations means is small, it is still significant because 0 falls outside of our interval, so we can still say that there is an association between creatinine level and survival status, though we may want to collect more data to be sure.

(*Q3.b*)

We also perform the *two-sided t-test allowing for heteroscedasticity*, where we define our hypotheses as:

- H_0: $\mu_1 = \mu_2$ There **is not** a difference in the mean creatinine level between the two populations.
- H_a: $\mu_1 \neq \mu_2$ These **is** a difference in the mean creatinine level between the two populations.

Table 10: Two-sided heteroscedastic t-test results of creatinine level by vital status group

95% CI (lower)	95% CI (upper)	p-value
0.09	0.27	0

Given an α level of .95 and a *p-value* of 7×10^{-5} this test confirms there is a statistically significant association between creatinine level and five-year survival status.

(*Q3.c*)

Again, both tests have similar confidence intervals, supporting the claim that there is an association between creatinine level and survival status among our population. The fact that both test produce similar results may be a sign that the population variance of each group is the same as the the sample variance of each group.