Homework 02

Spencer Pease

10/14/2019

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 three variables from the study data. They are briefly described below (see the study documentation for more details):

Table 1: Variable Descriptions

Variable	Units	Description
crt	mg/dl	Measure of creatinine in the participant's blood
obstime	days	Total time the participant was observed on the study
death	binary	Indicator that the participant died during the study

Of particular note *obstime* is measured in days, but this report is interested in grouping by vital status (*death*) at five years. This report treats a year as exactly 365 days, and will still report results using units of days ($365\frac{days}{year} \times 5years = 1825days$).

Creatinine Levels Across Groups Defined by Vital Status at 5 Years

Summary Statistics

We begin this report by looking at the summary statistics of the creatinine level variable distribution for two subsets of the data: participants who survived at least five years since the date of MRI, and participants who did not survive at least five years.

Table 2: Summary of creatinine level distribution by vital status group

Vital Status at 5 Years	valid	missing	mean	sd	min	q25	median	q75	max
Died within 5 years	121	0	1.22	0.47	0.5	0.9	1.1	1.3	4.0
Survived 5+ years	612	2	1.03	0.25	0.5	0.9	1.0	1.2	1.9

Between these two groups, we observe that the mean, median, IQR, and minimum values are within .1 to .2 $\frac{mg}{dl}$ of each other. The bigger differences can be seen in the number of observations in each group, with the surviving group having approximately 5 times as many observation as the non-surviving group. The distribution of the non-surviving group also has a heavier right-skew, leading to a higher maximum value and standard deviation.

Population Inference

We can use the distribution of creatinine level in our sample to infer some parameters of the larger population. Below, we look at the point estimate (sample mean), standard error of the point estimate, and 95% confidence interval of the population for separate vital status groups at 5 years.

Table 3: Population inference of creatinine level by vital status group

Vital Status at 5 Years	point est.	std. error	95% CI (lower)	95% CI (upper)
Died within 5 years	1.22	0.04	1.13	1.30
Survived 5+ years	1.03	0.01	1.01	1.05

Looking at the 95% confidence intervals for both vital status groups, we see that there is no overlap in the intervals. Since the confidence interval is interpreted as having *a confidence of 95% that the true population mean of creatinine level for a given group lies in their respective intervals*, we can say with some confidence that the difference in point estimates between the two vital status groups is due to more than just variations in the sample mean.

It is important to note that we can compare these confidence intervals in this manner because we can approximate both the distribution of the sample means for both of these subsets to a normal distribution because of the Central Limit Theorem, which states *for a sufficiently large n, the distribution of sample means approximates a normal distribution.* Our sample sizes for each subset are sufficiently large (approximately 100 and 600), so we can say the distributions are normal. Because they are normal, we can then standardize them and use the a standard z-score to build a confidence interval, allowing the two distributions to be compared.

Exploring different levels of confidence

Narrowing our scope to look at only participants in the vital status group not surviving at least 5 years, we can further explore the confidence interval of creatinine level by looking at the intervals created from different confidence levels. Let's look at three common levels: 90%, 95%, and 99%.

Table 4: Confidence intervals for creatinine level among participants not surviving at least 5 years

Confidence Level	lower	upper
90%	1.14	1.29
95%	1.13	1.30
99%	1.10	1.33

Notice the pattern of widening intervals for higher levels of confidence. This pattern exists because, for the same set of data, we can only be more confident that the true mean of a population parameter falls within an interval if that interval encompasses a larger range of possible values. Conversely, we can set a smaller interval, but we can't be as confident that the true value falls within it.

Geometric Mean Analysis

Instead of using the the **arithmetic** mean, standard error, and confidence interval to build our population inference, we can also look at the **geometric** mean, standard error, and confidence interval. These geometric statistics are useful when the data has an underlying exponential nature. Below we provide a table of the geometric population inference as an alternative to the arithmetic equivalent.

Table 5: Geometric population inference of creatinine level by vital status group

Vital Status at 5 Years	point est.	std. error	95% CI (lower)	95% CI (upper)
Died within 5 years	1.15	1.03	1.08	1.22
Survived 5+ years	1.01	1.01	0.99	1.02

Examining "High" Creatinine Levels

With the population inferences we made above, we can begin to make some statements regarding what a reasonable population parameter would be, and if any of our target population groups include exceptional values.

This report defines a "high" creatinine level in an individual as anything over $1.2\frac{mg}{dl}$. Grouping our data by vital status at 5 years, we are able to examine if there are any differences in the inferred population mean of creatinine level for each group.

Looking at the subset of individuals who survived at least 5 years since the date of MRI, we see from the prior tables that the confidence interval of both the arithmetic and geometric point estimate fall below our defined threshold of $1.2\frac{mg}{dl}$. Since our confidence interval represents the range that we are 95% certain includes the true population mean of creatinine level, we can say that a similar population of individuals who survive at least 5 years would not be prone to high creatinine levels.

For the subset of individuals who die within 5 years of their MRI date, we again look at the prior tables to see that the confidence interval of the for both the arithmetic and geometric interpretations of the point estimate of mean creatinine level include the values above the defined threshold of $1.2\frac{mg}{dl}$ for high creatinine level. Because of this, we can't say definitively if a population of similar individuals who died within 5 years of their MRI have high mean creatinine level, since our 95% confidence interval includes values on either side of the threshold, so the true population mean could be anywhere in that range.