1. In the in-course exercises, we have been analyzing data with accident as the observation unit. This study uses a different observation unit. What is the unit of observation in the Brady and Li study? When you download the FARS data for a year, you get a zipped folder with several different datasets. Which of the FARS datasets provides information at this observation level (and so will be the one you want to use for this analysis)?
   1. This study looked at drug and alcohol use in deaths that occurred within one hour of a motor vehicle crash, so the unit of observation will be an individual person and their corresponding information. All of this information is available in the *person* dataset from the FARS data.
2. This study only analyzes a subset of the available FARS data. Enumerate all of the constraints that are used by the study to create the subset of data they use in their study (e.g., years, states, person type, injury type). Go through the FARS documentation and provide the variable names for the variables you will use to create filter statements in R to limit the data to this subset. Provide the values that you will want to **keep** from each variable.
   1. Years: database files for 1999 through 2010 were downloaded
   2. States:
      1. Variable name = state
      2. Values to keep = 6 (California), 15 (Hawaii), 17 (Illinois), 33 (New Hampshire), 44 (Rhode Island), 54 (West Virginia)
   3. Person Type:
      1. Variable name = per\_type
      2. Values to keep = 1 (driver)
   4. Injury Type:
      1. Variable name = inj\_sev
      2. Values to keep = 4 (fatal)
   5. Time from accident to death:
      1. Variable names = lag\_hrs, lag\_mins
      2. Values to keep: lag\_hrs = 1 and lag\_mins = 0
3. The study gives results stratified by age category, year, year category (e.g., 1999–2002), alcohol level, non-alcohol drug category, and sex. For each of these stratifications, give the variable name for the FARS variable you could use to make the split (i.e., the column name you would use in a group\_by statement to create summaries within each category or the column name you would mutate to generate a categorical variable). Describe how each of these variables are coded in the data. Are there any values for missing data that you’ll need to mutate to NA values in R? Are there any cases where coding has changed over the study period?
   1. Age category: This variable is coded as “age” in the FARS dataset, with discrete numeric values from 0-96 prior to 2009 and 0-120 after that. Prior to 2009, 97 is used for ages 97 and older, and 99 is used for unknown ages. For 2009 and later, 998 is used for unreported ages, and 999 is used for unknown ages.
   2. Year was taken from the name of the dataset, with each dataset containing information for a specific year. To code for a year, we will have to pull that number from the title of the dataset when combining datasets. We can then stratify by individual year, or combine years into year categories and stratify by those.
   3. Alcohol level is taken from the “alc\_res” variable. This is coded as a value from 0-93, and has some slight variations for years 2010 and prior to 2010. Values of 94 = 0.94 or greater blood alcohol content. Prior to 2010, 95 = test refused, while in 2010 it means not reported. In addition, 96 = not given, 97 = results unknown, 98 = positive reading with no actual value, and 99 = unknown.
   4. Sex is coded as 1 for male and 2 for female in the sex variable. 9 represents “unknown”, and in 2010, 8 represents “not reported.”
   5. Non-alcohol drugs were pulled from variables called drugres1, drugres2, or drugres3, which gives results from up to three drug tests that may have been performed. They are all coded similarly, with a range of numbers corresponding to a specific drug type. In 2010, 95 represents “not reported”, whereas prior to 2010 this is grouped into the 999 value which represents “unknown/no reported.”