

Homework 1

DATA 202 - Alexander - Fall 2023

Please submit **Homework 1** responses as a .pdf file on Canvas [here](#).

Exercise 1.1

The [USPHS Syphilis Study at Tuskegee](#) was one of many historical cases of scientific racism and unethical practices enacted by the U.S. government. Was the USPHS Syphilis Study at Tuskegee an *experimental* or *observational* study? Explain your reasoning.

Exercise 1.2

Conduct a web search and literature review to identify two (2) other cases where data and/or information was collected using unethical practices. Provide a brief explanation of each case and its significance to understanding ethics. Include a full citation of all sources.

Exercise 1.3

There are many definitions of statistics. What is the definition of statistics that has been used in our course lectures? Conduct a web search and find two (2) alternative definitions of statistics. Include a full citation of all sources.

Exercise 1.4

Based on your current understanding of statistics, what should it mean to be *critical* in the context of statistics? Explain your thinking. Include a full citation of all sources.

Exercise 1.5

You have been tasked with identifying data for a new study on [homelessness](#) and [housing insecurity](#) in your local area. Identify and describe two (2) of each variable type that could be collected for this study: **nominal**, **ordinal**, **discrete**, **continuous**. That is, identify and fully describe two nominal, two ordinal, two discrete, and two continuous variables that could be used to gather insights about homelessness and housing insecurity. Your descriptions should be detailed.

Exercise 1.6

Your local city council plans to conduct a 2024 census of the local homeless population. You have been hired as a data analyst to identify priorities for the project. This census is part of a broader effort to understand the issues experienced during homelessness and to find ways to mitigate issues given the recent increase in the homeless population. What are some potential *ethical issues* that could arise with the councils' plans to conduct a local homeless census? In the city's plan to collect data on those experiencing homelessness, should *informed consent* be obtained per the IRB? If so, why and how?

Exercise 1.7

Describe the contents of the data set below and what the values most likely represent.

```
# A tibble: 6 x 62
  country year_1960 year_1961 year_1962 year_1963 year_1964 year_1965 year_1966
  <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 Afghani~  8996967    9169406    9351442    9543200    9744772    9956318   10174840
2 Albania   1608800    1659800    1711319    1762621    1814135    1864791    1914573
3 Algeria  11057864   11336336   11619828   11912800   12221675   12550880   12902626
4 America~   20127      20605      21246      22029      22850      23675      24473
5 Andorra   13410      14378      15379      16407      17466      18542      19646
6 Angola   5454938    5531451    5608499    5679409    5734995    5770573    5781305
# i 54 more variables: year_1967 <dbl>, year_1968 <dbl>, year_1969 <dbl>,
#   year_1970 <dbl>, year_1971 <dbl>, year_1972 <dbl>, year_1973 <dbl>,
#   year_1974 <dbl>, year_1975 <dbl>, year_1976 <dbl>, year_1977 <dbl>,
#   year_1978 <dbl>, year_1979 <dbl>, year_1980 <dbl>, year_1981 <dbl>,
#   year_1982 <dbl>, year_1983 <dbl>, year_1984 <dbl>, year_1985 <dbl>,
#   year_1986 <dbl>, year_1987 <dbl>, year_1988 <dbl>, year_1989 <dbl>,
#   year_1990 <dbl>, year_1991 <dbl>, year_1992 <dbl>, year_1993 <dbl>, ...

# A tibble: 6 x 62
  country year_1960 year_1961 year_1962 year_1963 year_1964 year_1965 year_1966
  <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 Vietnam   32670048   33666111   34683410   35721213   36780001   37858947   38958046
```

```

2 Virgin ~      32500      34300      35000      39800      40800      43500      46200
3 West Ba~      NA        NA        NA        NA        NA        NA        NA
4 Yemen, ~    5315351    5393034    5473671    5556767    5641598    5727745    5816241
5 Zambia     3070780    3164330    3260645    3360099    3463211    3570466    3681953
6 Zimbabwe   3776679    3905038    4039209    4178726    4322854    4471178    4623340
# i 54 more variables: year_1967 <dbl>, year_1968 <dbl>, year_1969 <dbl>,
#   year_1970 <dbl>, year_1971 <dbl>, year_1972 <dbl>, year_1973 <dbl>,
#   year_1974 <dbl>, year_1975 <dbl>, year_1976 <dbl>, year_1977 <dbl>,
#   year_1978 <dbl>, year_1979 <dbl>, year_1980 <dbl>, year_1981 <dbl>,
#   year_1982 <dbl>, year_1983 <dbl>, year_1984 <dbl>, year_1985 <dbl>,
#   year_1986 <dbl>, year_1987 <dbl>, year_1988 <dbl>, year_1989 <dbl>,
#   year_1990 <dbl>, year_1991 <dbl>, year_1992 <dbl>, year_1993 <dbl>, ...

```

Exercise 1.8

```
sum(1:51)
```

- What is the meaning of the code chunk `sum(1:51)`?
- What is the numerical output?

Exercise 1.9

The population of five countries is listed in a data set using computational scientific notation.

Numerically expand the population for each country.

- Country 1: 2.06139E8
- Country 2: 8.9561E7
- Country 3: 2.77E7
- Country 4: 2.72815E5
- Country 5: 6.077E3

Exercise 1.10

Describe the error in the following attempt to construct a data frame.

```

vec1 <- c(1, 2, 3, 4)
vec2 <- c("a", "b", "c", "d")
vec3 <- data.frame(T, F, F, T)
df <- data.frame(vec1, vec2, vec3)

```