````
---
title: "Problem Set 1"
output:
  pdf_document: default
  word_document: default
  html_document:
    df_print: paged
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
load("anes2020subset.RData")
```

**Problem Set 1
Francisco Brady
`r Sys.Date()`**

### Question 1

The `anes2020` dataset has 8280 records.

```{r, eval=TRUE}
nrow(anes2020)
```

### Question 2

Below is the output of `summary(anes2020$SciImptCovid)`.

```{r, eval=TRUE, echo=TRUE}
summary(anes2020$SciImptCovid)
```

Below is the output of `table(anes2020$SciImptCovid)`

```{r}
table(anes2020$SciImptCovid)
```

#### (a)

The difference between the two outputs is that the `summary()` command identified that the
`SciImptCovid` variable has 897 missing values or NA's.

#### (b)

It's important to take account of the missing values. We should be concerned about missing
values because if there are a lot of missing values, we may introduce __nonresponse bias__
into our understanding of the variable. It might not be possible to say anything meaningful
about the data if there are a sufficient number of missing values.  In the prompt it was
mentioned that the `ScienceExperts` question was only asked in the post-election survey. The
addition of that question, along with the `SciImpCovid` question may have impacted responses.


#### (c)

The most common response to `SciImptCovid` is "Extremely Important".

#### (d)

The `SciImptCovid` is a categorical variable who's responses form an ordinal scale. The
````

"lowest" response is "Not at all important", and the "highest" value is "Extremely important".

Question 3

```{r }
table(anes2020$EconWorse)
table(anes2020$DiscussPol)
table(anes2020$HomeOwnership)
table(anes2020$VotedBiden)
table(anes2020$Empathy)
```

a) EconWorse -- Categorical, Ordinal, Discrete
b) DiscussPol -- Quantitative, Interval, Discrete
c) HomeOwnership -- Categorical, Nominal, Discrete
d) VotedBiden -- Categorical, Nominal, Discrete
e) Empathy -- Categorical, Ordinal, Discrete

Question 4

a) EconWorse -- Mode, Median
b) DiscussPol -- Mean, Median, Mode
c) HomeOwnership -- Mode
d) VotedBiden -- Mode
e) Empathy -- Median, Mode

Question 5
```{r}
summary(anes2020$SCOTUStherm)
table(anes2020$SCOTUStherm)
#max(table(anes2020$SCOTUStherm))

```

a) Mean: 60.67, Median: 60, Mode: 50

```{r}
# using sd function to print standard deviation
sd(anes2020$SCOTUStherm, na.rm = T)
```

b) The standard deviation of this variable is 21.83507

c) A Histogram

```{r }
hist(anes2020$SCOTUStherm)
```

Question 6

a) A frequency table for `SpendHighways`. The measurement level is categorical, ordinal, discrete.

````{r}
table(anes2020$SpendHighways)
```

b) A bar graph for `SpendHighways`

```{r}
barplot(table(anes2020$SpendHighways),
        las=1, cex.names=.7,
        main = 'The Level of Government\nSpending on Highways has:')
```

c)  Treating the categorical variable as numeric, removes the labels from the variable. The counts correspond to the categories (Decreased a lot = 1, Decreased a little = 2, Kept the same = 3, Increased a little = 4, Increased a lot = 5). The ordering of the number does have meaning in that the response choices imply an increase in government spending. It's not possible to say whether each step on the scale is the same amount of change, so it would be difficult to interpret as an interval level variable.

```{r}
table(as.numeric(anes2020$SpendHighways))
```

d)  Reporting the `SpendHighways` as a numeric variable:

```{r}
summary(as.numeric(anes2020$SpendHighways))
```

The calculation here is using the response as it is stored by R to calculate the summary statistics. Calculating the mean or standard deviation for these numbers would not be meaningful.

### Question 7

a)  For the first method, the democracy variable would be categorical, nominal, and discrete. For the second method, the democracy variable would be categorical, ordinal, and discrete.
b)  It depends heavily on the context which measurement strategy is more reliable. Reliability is about how consistent the measure is. Since the first method relies on events that happen (elections), it could be more reliable. The second method may be less reliable because the values are contingent on a set of experts (humans) whose value assignment might be affected by many other circumstances or contexts.
c)  Again this depends on the context and the aims of the measurement. Validity of a measurement is dependent on how well it represents the concept it is measuring. A rating from experts on civil liberties and political rights could be a good indicator of how democratic a country is. A binary measurement of whether a country is a democracy or not based solely on elections may not be as valid of a representation.

### Question 8

a)  For the first method, the measurement levels would be categorical ("favorable", "neutral", "unfavorable"), ordinal, and discrete. For the second method the measurement levels would be quantitative (count of "negative" key words), interval, and discrete.
b)  Reliability might be a concern when using the first method, but could be addressed by having the graduate students overlap the news articles they read and assign codes to, since re-testing and getting the same grade would increase confidence in the reliability of the students' assessments. In the second method, since the measure is simply a count of key words, it seems like it would be more reliable.
c)  In terms of validity, depending on the key words picked, I think the second method would be more valid, since it is relying on opinions a bit less. The fact that the measurement is being done by computer software also increases the ease of validation by other researchers.

<!-- ## Test -->

```
<!-- ```{r} -->
<!-- # writng our own standard deviation function  -->
<!-- std_dev <- function(x, na.rm = T){ -->
<!--   # find sample variance: summation (x - x_bar)^2 / n - 1 -->
<!--   if(na.rm){ -->
<!--     x <- na.omit(x) -->
<!--   } -->
<!--   n <- length(x) -->
<!--   sample_var <- sum((x - mean(x))^2) / (n - 1) -->
<!--   # take the square root of that sample variance to get SD -->
<!--   sdev <- sqrt(sample_var) -->
<!--   return(sdev) -->
<!-- } -->

<!-- # test  -->
<!-- regular_sd <- sd(anes2020$SCOTUStherm, na.rm = T) -->
<!-- my_sd <- std_dev(anes2020$SCOTUStherm, na.rm = T) -->
<!-- identical(regular_sd, my_sd) -->
<!-- ``` -->
```