LA-7: Examining Relationships (15 points)

Learning Outcomes

In this assignment, you will:

- Select the appropriate statistical test.
- Conduct a statistical test.
- Interpret the results of a statistical test.

> **Tip**
>
> Read all the instructions carefully before starting the assignment.

Instructions

1) Load the packages below and download the `covid.csv` data file from Canvas. Read the dataset into R. Use the codebook on Canvas to familiarize yourself with the data in this file.
   - `tidyverse`
   - `summarytools`
   - `rstatix`
2) Create a new variable called `region`. Have the variable equal `Americas` if the country is Brazil, Canada, Mexico, or the United States. Have the variable equal `Asia` if the country is China, India, Japan, Singapore, or South Korea. Have the variable equal `Europe` is the country is Denmark, France, Germany, Italy, or Spain. Run a frequency distribution of region. Which region has the highest frequency?
3) Let's say we want to conduct a statistical test to determine whether the total number of contacts people had on the previous day differs by region. First, answer the questions below. Be sure to include your answers as comments in your R file.
   a) Which variables are involved in this statistical test? Which is the independent variable? Which is the dependent variable?
   b) Run a frequency distribution of `total_contacts`. Is this variable categorical or continuous?
   c) Is the variable `region` continuous or categorical?
   d) Given your answers to the previous questions, what statistical test should you use? Why? Refer to the flowchart and instructions on page 3 to help select a statistical test.
4) Conduct the statistical test from your answer to 3d. Answer the following questions.
   a) What is the test statistic and the probability value (i.e., p-value) associated with that test statistic?
   b) Based on the test statistic and the p-value, do the number of total contacts differ by region?
5) Next, let's compare how mask-wearing differs between the first and second year after the COVID outbreak. Follow the steps below to select and conduct the appropriate statistical test.

a) State your hypothesis. Do you think more people wore masks when they went outside in the first or second year after the outbreak? Why?
b) Identify the dependent variable. Is it continuous or categorical?
c) What variable can you use to create the independent variable? Describe how you might recode an existing variable to create the independent variable. Then, create this new variable using the `mutate()` and `case_when()` functions. Is this new variable continuous or categorical?
d) Select and conduct the appropriate statistical test.
e) Report the test statistic and probability value as a comment in your R script.
f) Calculate the means and standard deviations of the dependent variable for the first and second years after the COVID outbreak. Include these descriptive statistics in your R script. Did people wear masks more often when they went outside in the first or second year after the outbreak?
g) Does the statistical test support your hypothesis?

Submission

Submit your R script (named `LA-#_FirstName-LastName.R`) to Canvas.

Your R script should:

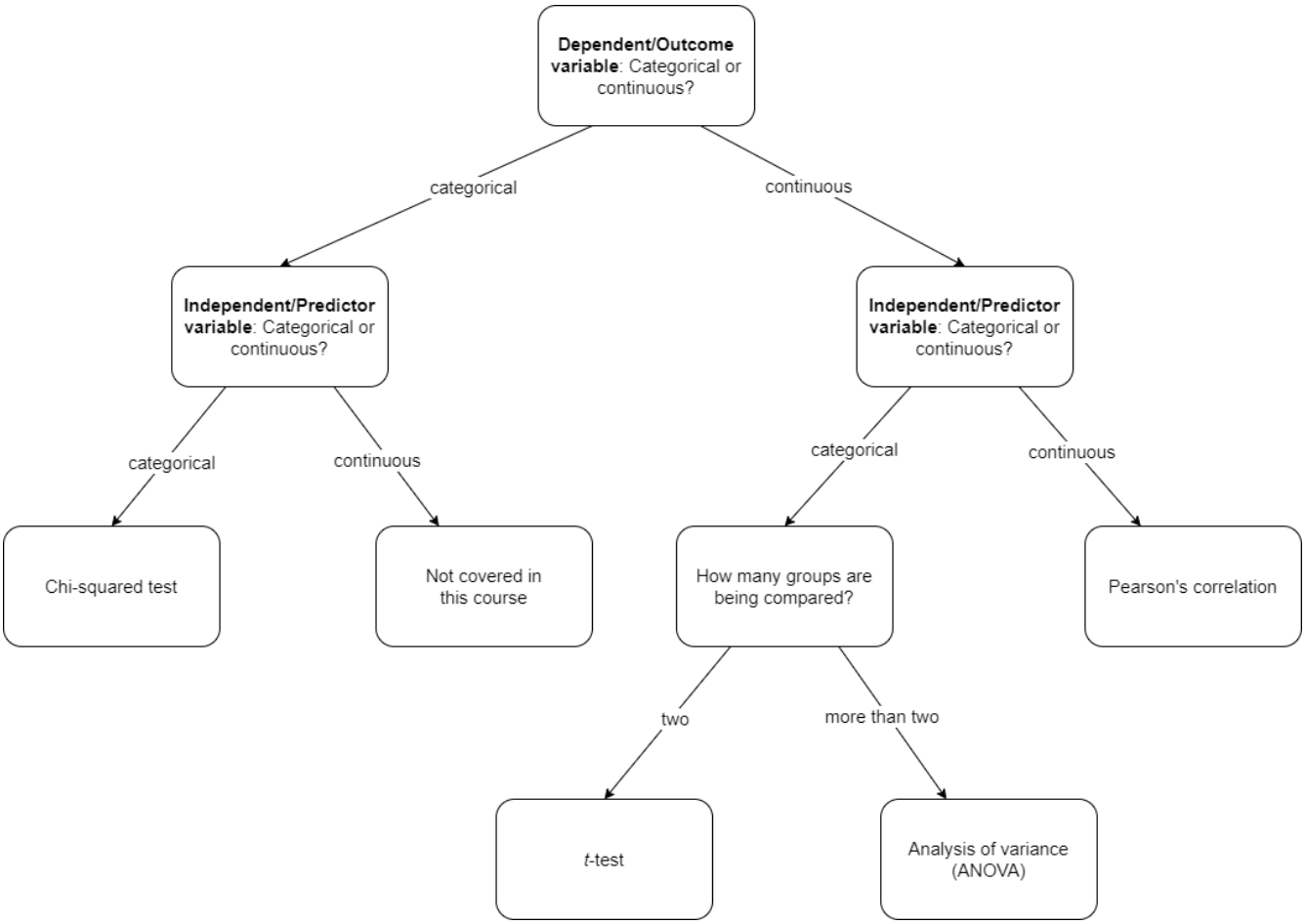
1) Include commands and functions that are necessary to address all the questions in the assignment.
2) Contain comments that answer the questions in the assignment.
3) Run in its entirety without errors.

To ensure that your R script runs without errors, you should:

- Save your script.
- Navigate back to Your Workspace on Posit Cloud.
- Reopen your project.
- Run the entire script line-by-line without editing it to ensure there are no errors.

> **!** Important
>
> These standards apply to all submissions in this course that require R scripts. You should follow these instructions for preparation, naming, and saving of your R script for all of your individual lab assignments.

Selecting a Statistical Test



Conducting a Statistical Test in R

Chi-squared test

A chi-squared test is used when we want to compare two categorical variables. To do so in R, you can use the following command (`df`, `dep_var`, and `ind_var` have to be replaced with the names of your dataframe, dependent variable, and independent variable, respectively). The output that results from running this command in R is shown below as comments.

```
chisq_test(x = df$ind_var, y = df$dep_var)

# A tibble: 1 × 6
#       n statistic         p    df method          p.signif
# * <int>     <dbl>     <dbl> <int> <chr>           <chr>
# 1   291      22.5 0.0000133     2 Chi-square test ****
```

To report the results of your chi-squared test in APA format, you would use the following format, replacing the text in bold font with the results from your Console:

$\chi^2(\mathbf{df}, N = \mathbf{n}) = \mathbf{statistic}, p = \mathbf{p}$

From the results of the above chi-squared test, I would report results like this: $\chi^2(df = 2, N = 291) = 22.5, p < .001$.

If the p-value is less than .001, you do not need to report the exact value. If it is greater than .001, then report the exact value rounded to two or three decimal places.

t-test

A t-test is used to compare the mean values of a continuous variable between two groups. To do so in R, you can use the following command (`df`, `dep_var`, and `ind_var` have to be replaced with the names of your dataframe, dependent variable, and independent variable, respectively). The output that results from running this command in R is shown below as comments.

```
df |>
        t_test(dep_var ~ ind_var)

# A tibble: 1 × 10
#   .y.        group1 group2    n1    n2 statistic    df        p    p.adj p.adj.signif
# * <chr>      <chr>  <chr>  <int> <int>     <dbl> <dbl>    <dbl>    <dbl> <chr>
# 1 dep_var ind_var1  ind_var2  54    70     -12.2  122. 6.49e-23 6.49e-23 ****
```

To report the results of your chi-squared test in APA format, you would use the following format, replacing the text in bold font with the results from your Console:

$t(\texttt{df}) = \texttt{statistic}$, $p = \texttt{p}$

Using the results above, I would report the results of a t-test like this: $t(df = 122) = -12.2$, $p < .001$.

Analysis of variance (ANOVA)

ANOVA is used when we want to compare the mean values of continuous variable between more than two groups. To do so in R, you can use the following command (`df`, `dep_var`, and `ind_var` have to be replaced with the names of your dataframe, dependent variable, and independent variable, respectively). The output that results from running this command in R is shown below as comments.

```
df |>
        anova_test(dep_var ~ ind_var)

# ANOVA Table (type II tests)
#
#   Effect  DFn DFd     F        p p<.05   ges
# 1 ind_var   2 157 7.505 0.000771     * 0.087
```

To report the results in APA format:

$F\,(\texttt{DFn}, \texttt{DFd}) = \texttt{F-value}$, $p = \texttt{p-value}$

Using the results above, I would report: $F(2, 157) = 7.51$, $p < .001$.

Correlation

We use Pearson's correlation to examine a linear relationship between two continuous variables. To run a correlation in R, you can use the following command (`df`, `dep_var`, and `ind_var` have to be replaced with the names of your dataframe, dependent variable, and independent variable, respectively). The output that results from running this command in R is shown below as comments.

```
df |>
        cor_test(dep_var, ind_var)

# # A tibble: 1 × 8
#   var1          var2                  cor statistic        p conf.low conf.high method
```

```
#   <chr>         <chr>               <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
# 1 masks_outside days_since_outbreak   0.2      3.48 0.000584   0.0875     0.308 Pearson
```

To report the results in APA format:

> Pearson's $r =$ `cor`, $p =$ `p-value` (or $p < .001$ if the p-value from your test is very small, i.e., less than .001)

Using the results above, I would report: Pearson's $r = 0.2$, $p < .001$.