**Type your answers in the appropriate fields; please make answer fields larger as needed. The text box font has been set to blue to make it easier to grade. Turn in your assignment on canvas by Friday, October 27st at 11:59 PM. This assignment will be graded by Simran.**

*Conceptual Questions*

Answer each of the questions completely in approximately one to three sentences. Please use bullet points for different ideas/components of your answer. Please delete any instance of “TYPE ANSWER HERE” and replace it with your own answer.

1. Explain why the following statement is FALSE, and include a comparison between Standard Deviation and Standard Error. Also explain how you would correct the sentence. [1 Point]

“The Standard Deviation is how much ***scores vary in a sample***, and the Standard Error of the Mean indicates how much ***scores vary in a population.***”

* The standard deviation definition is partially correct since it is a measure of variability of values, but it is not restricted to samples.
* The standard error does not measure score variation in the population. Instead, it is a measure of how much the sample mean is likely to vary from the population mean.
* The correct sentence should be: “The Standard Deviation is how much scores vary in relation to the mean, and the Standard Error of the Mean indicates how much sample means vary in relation to the population mean.”

1. Power is influenced by the Effect Size, Sample Size, Variance, and Alpha Level. For each factor, (A) state how power would change if the factor increased, and (B) use the *t*-test formula to help explain (in 2-3 bullet point sentences) why these factors logically influence power (when the other three factors are held constant).

Hint: For part B, many of the factors affect either the numerator or the denominator of the *t* calculation; think about how changes in each of these locations would affect the resulting *t* statistic. I.e., does increasing a certain factor lead to a larger or smaller numerator or denominator? How would that affect the resulting *t* statistic and the probability of rejecting the null hypothesis? [2 Points]

The t-test formula:

Diagram

Description automatically generated with medium confidence

←Denominator

←Numerator

←Numerator

**Effect Size:**

* **As effect size increases, power will:** Increase.
* **Why:**
  + As the difference between means (i.e., the numerator) increases, so does the effect size.
  + If the numerator increases and the denominator is held constant, it is expected that the *t* statistic will also increase.
  + Larger *t* values increase the likelihood of the test to obtain *t* > critical *t*, increasing power.

**Sample Size:**

* **As sample size increases, power will:** Increase.
* **Why:**
  + As the sample size increases and the variance is held constant, the standard error (i.e., the denominator) decreases.
  + Even with small effect sizes, decreasing the denominator will result in a larger *t* statistic, increasing power.

**Variance:**

* **As variance increases, power will:** Decrease.
* **Why:**
  + As the variance increases, the standard error will also increase, making the denominator larger if all the other factors are held constant.
  + The resulting t values will then be smaller, diminishing the likelihood of obtaining *t* > critical *t* and decreasing power.

**Alpha:**

* **As alpha increases, power will:** Increase.
* **Why:**
  + Increasing alpha results in smaller values of critical t to reject the null hypothesis.
  + If the other three factors are constant, the t value will not change. But making the critical *t* smaller implies increasing the likelihood of obtaining *t* > critical *t*, increasing power.

*Computational Problems*

**Use the lab data set “Age Religion Health \_ HW 4.csv”**. These data are a subset from the Age, Religion, and Health data set that we used in the lab (do NOT use the same data from the lab, you will not get the right answers). You will test hypotheses about health behaviors and religious behaviors. Please put each answer in line with each portion of the question, and put all code and syntax in the portion below in the space marked for it (the very last page of the homework).

1. Assess whether there are differences between smokers and non-smokers in frequency of Church attendance (HowOftenChurch\_04) and Relationship with God (RelationshipGod\_04). Please note that higher scores on these variables indicate greater attendance at Sunday School or a closer relationship with God.
   1. First, calculate the mean, *SD, N,* and *SE* for each group and report these values in the table below (please round to 2 decimal places). *Note*: Make sure that when calculating the summary statistics you filter out subjects that have missing (NA) data in the dependent variable (make sure to do this for each DV separately, as the same people missing data in one variable may not be the same people missing data for the other variable!). [1.5 Points]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Non-Smokers | | | | Smokers | | | |
| Variable | Mean | *SD* | *SE* | *N* | Mean | *SD* | *SE* | *N* |
| **Church Attendance** | 5.40 | 2.95 | 0.11 | 726 | 4.22 | 2.65 | 0.38 | 49 |
| **Relationship with God** | 3.54 | 0.65 | 0.02 | 721 | 3.54 | 0.58 | 0.08 | 50 |

* 1. Conduct an independent samples *t*-tests comparing smokers and non-smokers on Church attendance, and conduct a separate independent samples t-test comparing smokers and non-smokers on their Relationship with God. Assume unequal variances. Report the statistics in the table below. [2 Point]

Answer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | t | DF | *p* | *d* |
| **Church Attendance** | 2.99 | 56.3 | 0.004 | 0.42 |
| **Relationship with God** | -0.04 | 57.9 | 0.97 | -0.005 |

* 1. Make graphs depicting the means and *SE* bars for each group for each outcome variable (either a bar graph or points are OK). Make your graph as clean as possible (label the axes, give the graphs titles, remove missing data, etc.). [1.5 Points]

Answer:

A graph with lines and dots

Description automatically generated with medium confidence

* 1. Interpret the results of the two analyses (with one bullet-pointed sentence per analysis). Specifically: If there was a significant difference between groups on a variable, state the direction of differences. Otherwise state that there were no differences between the groups. In both cases, cite *p* values to support your claims (round to two or three decimal places). [1 Point]

**Church Attendance:** The results indicate that Non-smokers have a statistically significant higher mean frequency of Church attendance than people who smoke (*p* = 0.004)

**Relationship with God:** The mean frequency of Relationship with God was equivalent between smokers and non-smokers, as the test results were non-significant (*p* > 0.05)

* 1. Do these results surprise you? What do you think could be happening here to explain these results? There is no right or wrong answer per se, as long as you demonstrate critical thinking. [1 Point]
* With these results, we observed that although non-smokers have a higher mean of church attendance than smokers, this difference is not reflected in the way they perceive their relationship with God. We can hypothesize that smokers might cultivate spirituality similar to non-smokers but are less prone to go to church due to fear of judgment of their smoking habits.

**Extra Credit** [0.5 point]

Suppose you were conducting a one-sample t-test and had a hypothesized mean of 5, a sample standard deviation of 1.4, and a sample size of 32. Your test statistic was 2.83. What was your original sample mean?

5.70

**Code/Syntax for all Questions:**

# Load packages -----------------------------------------------------------

library(ggplot2)

library(dplyr)

# Load data ---------------------------------------------------------------

dat <- readr::read\_csv("Age Religion Health \_ HW 4.csv")

# Question 3 - A ----------------------------------------------------------

church <- dat |>

dplyr::filter(!is.na(HowOftenChurch\_04)) |>

dplyr::with\_groups(Smoke\_04,

summarise,

mean = mean(HowOftenChurch\_04, na.rm = T),

sd = sd(HowOftenChurch\_04, na.rm = T),

N = length(HowOftenChurch\_04)) |>

dplyr::mutate(se = sd/sqrt(N))

relationship <- dat |>

dplyr::filter(!is.na(RelationshipGod\_04)) |>

dplyr::with\_groups(Smoke\_04,

summarise,

mean = mean(RelationshipGod\_04, na.rm = T),

sd = sd(RelationshipGod\_04, na.rm = T),

N = length(RelationshipGod\_04)) |>

dplyr::mutate(se = sd/sqrt(N))

# Question 3 - B ----------------------------------------------------------

t.test(HowOftenChurch\_04 ~ Smoke\_04, data = dat, var.equal=FALSE)

effectsize::cohens\_d(HowOftenChurch\_04 ~ Smoke\_04, data = dat, var.equal=FALSE)

t.test(RelationshipGod\_04 ~ Smoke\_04, data = dat, var.equal=FALSE)

effectsize::cohens\_d(RelationshipGod\_04 ~ Smoke\_04, data = dat, var.equal=FALSE)

# Question 3 - C ----------------------------------------------------------

long\_data <- dat |>

dplyr::select(Smoke\_04, HowOftenChurch\_04, RelationshipGod\_04) |>

tidyr::pivot\_longer(cols = -Smoke\_04, values\_to = "score", names\_to = "variable") |>

dplyr::with\_groups(c(Smoke\_04, variable), summarise,

mean = mean(score, na.rm = T),

sd = sd(score, na.rm = T),

N = sum(!is.na(score))) |>

dplyr::mutate(se = sd/sqrt(N))

long\_data |>

ggplot(aes(x = variable, y = mean, color = Smoke\_04)) +

geom\_point(size = 3) +

geom\_errorbar(aes(ymin = mean - se, ymax = mean + se),

color = "black", size = 0.5, width = 0.2, alpha = 0.6) +

theme\_minimal() +

labs(

title = "Mean and Error Bars for outcome variables by smoking status",

x = "Outcome",

y = "Mean",

color = "Smoking status"

) +

scale\_x\_discrete(labels = c(

"HowOftenChurch\_04" = "Church attendance",

"RelationshipGod\_04" = "Relationship with God"

)) +

theme(

text = element\_text(size = 12),

legend.position = "top",

legend.title = element\_text(size = 10, face = "bold"),

plot.title = element\_text(size = 14, face = "bold"),

axis.title = element\_text(size = 12, face = "bold"),

legend.text = element\_text(size = 10)

)

# Extra question ----------------------------------------------------------

hyp\_mean <- 5

sample\_sd <- 1.4

sample\_size <- 32

t\_stat <- 2.83

sample\_mean = (t\_stat \* (sample\_sd / sqrt(sample\_size))) + hyp\_mean