**Assignment 1: Database Management and Sample Size Calculation**

Steven Jackson

Dalhousie University

Department of Psychiatry

Dr. Igor Yakovenko

March 7th, 2024

*Please note: I couldn’t figure out how to include APA formatting in Quarto (title page, page number, Times New Roman, double-spacing, etc.), so I manually formatted the report. The code I wrote will generate a docx file that does not include the APA formatting.*

The report herein summarizes key findings from the Avengers dataset, which is data collected about the avengers involved during the battle against Thanos.

First, the Tidyverse (Wickham et al., 2019) and Abind (Plate & Heiburger, 2016) R packages were loaded in.

#Library'ing in TidyVerse and aBind for access to useful operations.   
library(tidyverse)

Warning: package 'tidyverse' was built under R version 4.2.3

Warning: package 'ggplot2' was built under R version 4.2.3

Warning: package 'tibble' was built under R version 4.2.3

Warning: package 'readr' was built under R version 4.2.3

Warning: package 'purrr' was built under R version 4.2.3

Warning: package 'dplyr' was built under R version 4.2.3

Warning: package 'stringr' was built under R version 4.3.2

Warning: package 'forcats' was built under R version 4.3.2

Warning: package 'lubridate' was built under R version 4.2.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(abind)

Next, the dataset reposition was cloned from <https://github.com/iyakoven/PSYR6003-Assignment-1> .

#Importing the data.  
avengers<-read\_csv("avengers.csv")

Rows: 814 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (3): superpower, north\_south, died  
dbl (12): iq, agility, speed, strength, damage.resistance, flexibility, will...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

The data was then inspected for missing and unusual values. The data was subset to include only complete, clean cases.

#Question 1:  
  
#Inspecting the dataset for missing/unusual values.  
  
view(avengers)

A new custom variable was created called ‘CombatEffectiveness’, defined as a sum of agility, speed, strength and willpower.

#Question 2: Removing rows with missing data, create CombatEffectiveness.  
  
avengers\_clean <- na.omit(avengers)  
avengers\_clean1 <- mutate(avengers\_clean, CombatEffectiveness=agility+speed+strength+willpower)  
head(avengers\_clean1, 20)

# A tibble: 20 × 16  
 iq agility speed strength damage.resistance superpower flexibility  
 <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
 1 118 27 5.05 400 2.1 no -1.9  
 2 109 54 4.98 782 2.09 no -2.2  
 3 90 56 4.84 569 1.99 no -1   
 4 114 63 5.24 535 2.19 no -0.3  
 5 116 30 5.01 663 2.33 no -2.6  
 6 137 62 4.79 435 2.05 no -0.9  
 7 114 57 4.96 418 2 no -1.5  
 8 102 63 5.39 630 2.05 no -0.6  
 9 98 70 5.14 480 1.71 no -0.2  
10 99 52 5.05 663 2 no -1.2  
11 112 47 5.04 494 2.07 no -1.2  
12 109 -11 4.25 22 1.62 no -1.3  
13 108 33 4.85 591 2.11 no -2.6  
14 104 43 4.83 679 2.24 no -2.7  
15 118 33 5.15 599 2.15 no -1.4  
16 113 65 5.22 509 2.21 no 0.9  
17 107 74 5.21 466 1.73 no -0.9  
18 110 35 5.07 579 2.15 no -1.7  
19 116 36 4.75 282 1.81 no -1.3  
20 102 46 5.15 745 2.19 no -1.5  
# ℹ 9 more variables: willpower <dbl>, ptsd <dbl>, north\_south <chr>,  
# died <chr>, kills <dbl>, injuries <dbl>, minutes.fighting <dbl>,  
# shots.taken <dbl>, CombatEffectiveness <dbl>

#Double-checking filtration & new CombatEffectiveness variable.  
  
head(avengers\_clean1, 20)

# A tibble: 20 × 16  
 iq agility speed strength damage.resistance superpower flexibility  
 <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
 1 118 27 5.05 400 2.1 no -1.9  
 2 109 54 4.98 782 2.09 no -2.2  
 3 90 56 4.84 569 1.99 no -1   
 4 114 63 5.24 535 2.19 no -0.3  
 5 116 30 5.01 663 2.33 no -2.6  
 6 137 62 4.79 435 2.05 no -0.9  
 7 114 57 4.96 418 2 no -1.5  
 8 102 63 5.39 630 2.05 no -0.6  
 9 98 70 5.14 480 1.71 no -0.2  
10 99 52 5.05 663 2 no -1.2  
11 112 47 5.04 494 2.07 no -1.2  
12 109 -11 4.25 22 1.62 no -1.3  
13 108 33 4.85 591 2.11 no -2.6  
14 104 43 4.83 679 2.24 no -2.7  
15 118 33 5.15 599 2.15 no -1.4  
16 113 65 5.22 509 2.21 no 0.9  
17 107 74 5.21 466 1.73 no -0.9  
18 110 35 5.07 579 2.15 no -1.7  
19 116 36 4.75 282 1.81 no -1.3  
20 102 46 5.15 745 2.19 no -1.5  
# ℹ 9 more variables: willpower <dbl>, ptsd <dbl>, north\_south <chr>,  
# died <chr>, kills <dbl>, injuries <dbl>, minutes.fighting <dbl>,  
# shots.taken <dbl>, CombatEffectiveness <dbl>

Next, a new copy of the dataset in both SPSS and csv formats was created which only includes the avengers who did not have a superpower and have died. The haven package (Wickham et al., 2023) was loaded in to facilitate creating the SPSS-formatted file.

#Question 3: New dataset copies, summarizing properties of combat effectiveness, kills, & injuries.  
  
#CSV file  
avengers\_clean\_filtered <- filter(avengers\_clean1, superpower=="no" & died=="yes")   
write.csv(avengers\_clean\_filtered, "avengers\_clean\_filtered.csv", row.names=F, na="")  
  
#SPSS file  
library(haven)

Warning: package 'haven' was built under R version 4.2.3

write\_sav(avengers\_clean\_filtered, "avengers\_clean\_filtered.sav")  
  
#Double-checking that database includes avengers who both did not have superpowers and have died.  
head(avengers\_clean\_filtered, 20)

# A tibble: 20 × 16  
 iq agility speed strength damage.resistance superpower flexibility  
 <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
 1 114 57 4.96 418 2 no -1.5  
 2 109 -11 4.25 22 1.62 no -1.3  
 3 116 36 4.75 282 1.81 no -1.3  
 4 104 47 4.92 352 1.67 no -1.9  
 5 109 48 5.19 554 1.94 no -3.1  
 6 113 25 4.65 236 1.85 no -1.4  
 7 112 55 4.8 422 1.67 no -0.8  
 8 109 39 4.77 420 2.21 no -1.3  
 9 94 43 5.12 464 1.89 no -0.5  
10 99 -2 4.51 143 1.74 no -1.4  
11 109 63 5.2 704 1.91 no -1   
12 120 59 5.12 533 2.02 no -0.3  
13 114 33 4.93 300 2.08 no -0.6  
14 102 46 4.92 328 1.98 no 0.3  
15 103 58 4.81 311 1.69 no 0.3  
16 98 33 4.82 317 1.8 no -1.5  
17 107 12 4.73 230 1.85 no -1.1  
18 113 43 4.93 346 1.94 no 0.2  
19 102 34 5.14 603 2.14 no -1.2  
20 118 44 4.77 295 2.16 no -0.1  
# ℹ 9 more variables: willpower <dbl>, ptsd <dbl>, north\_south <chr>,  
# died <chr>, kills <dbl>, injuries <dbl>, minutes.fighting <dbl>,  
# shots.taken <dbl>, CombatEffectiveness <dbl>

Next, summary reports of the overall kills, injuries, and combat effectiveness were prepared. A second report grouped these results by battlefield.

#Summarize properties of combat effectiveness, kills, & injuries  
  
#Overall stats  
mean\_stats\_overall<- summarize(avengers\_clean\_filtered,  
   
 #Kills  
 mean\_kills = mean(kills),  
 sd\_kills = sd(kills),  
 min\_kills = min(kills),  
 max\_kills = max(kills),  
   
 #Injuries  
 mean\_injuries = mean(injuries),  
 sd\_injuries = sd(injuries),  
 min\_injuries = min(injuries),  
 max\_injuries = max(injuries),  
   
 #CombatEffectiveness  
 mean\_CE = mean(CombatEffectiveness),  
 sd\_CE = sd(CombatEffectiveness),  
 min\_ce = min(CombatEffectiveness),  
 max\_ce = max(CombatEffectiveness)  
 )  
  
#Stats by battlefield  
grouped\_avengers\_clean\_filtered <- group\_by(avengers\_clean\_filtered, north\_south)  
mean\_stats\_battlefield<- summarize(grouped\_avengers\_clean\_filtered,  
   
 #Kills  
 mean\_kills = mean(kills),  
 sd\_kills = sd(kills),  
 min\_kills = min(kills),  
 max\_kills = max(kills),  
   
 #Injuries  
 mean\_injuries = mean(injuries),  
 sd\_injuries = sd(injuries),  
 min\_injuries = min(injuries),  
 max\_injuries = max(injuries),  
   
 #CombatEffectiveness  
 mean\_CE = mean(CombatEffectiveness),  
 sd\_CE = sd(CombatEffectiveness),  
 min\_ce = min(CombatEffectiveness),  
 max\_ce = max(CombatEffectiveness)  
)

Overall, the mean kills in the dataset is 2.55 (range = 0, 79; SD = 8.81), mean injuries is 4.55 (range = 2, 5; SD = 0.74), and mean combat effectiveness is 497.53 (range = 67.25, 946.89; SD = 177.56).

Presenting the results by battlefield reveals some differences between the north and south battlefields. The mean kills in the north battlefield is 1.71 (range = 0, 34; SD = 4.57), and 4.75 (range = 0, 79; SD = 14.99) in the south battlefield. The mean injuries in the north battlefield is 4.60 (range = 2, 5; SD = 0.68), and 4.43 (range = 2, 5; SD = 0.88) in the south battlefield. The mean combat effectiveness in the north battlefield is 499.78 (range = 130.68, 897.06; SD = 174.07), and 491.68 (range = 67.25, 946.89; SD = 189.53) in the south battlefield.

The north battlefield had both the higher mean combat effectiveness and higher mean injuries. Kills is the most erroneous variable as it has the highest standard deviation relative to its mean value (i.e. a higher proportion of the measurement is attributable to error).

Next, we wanted to see if avengers with superpowers have a higher average IQ than those without. Below, we consider two options for estimating the sample size for an independent t-test of this analysis.

Two ways of estimating the required sample size are a) to include every existing avenger and b) run an a-priori power analysis using the expected effect size of the superpowers-IQ relationship. For the former, this is reasonable given that there was only one battle against Thanos, so the entire population of interest is accessible. For the latter option, running an a-priori power analysis will require estimating an effect size, which can be based on the smallest effect size of interest (SESOI). The SESOI is an estimate of the smallest effect size we would be interested in leveraging in the context of some hypothetical real-world application of the findings (i.e., an intervention, cost-benefit analyses, etc.).

As mentioned above, the effect size can be justified using the SESOI. In this case, we will say that an effect size of d = 0.3 would be of interest. In a real-life context, we might base this on other values already known to be associated with IQ in avengers, and pick an effect size that is equal to or higher than those existing values. If the effect we find is lower than those existing values, time and resources would be better spent investigating those variables.

Using the pwr package (Chameply, 2020) in RStudio, a power analysis was conducted to determine the sample size needed to detect a group difference with a significance level of α = 0.05, power = 0.80, and an effect size of d = 0.3. The alpha and power levels were chosen to be consistent with existing literature in the field, such that a statistically significant finding in one analysis would hypothetically still be considered significant by another researcher’s standard. The effect size is justified above on the basis of SESOI. The two-sided t-test is used, because we are interested in whether avengers with superpowers have either a higher or lower IQ.

library(pwr)

Warning: package 'pwr' was built under R version 4.2.3

pwr.t.test(n = NULL, d = .3, sig.level = 0.05, power = .8,   
 type = c("two.sample"),  
 alternative = c("two.sided"))

Two-sample t test power calculation   
  
 n = 175.3847  
 d = 0.3  
 sig.level = 0.05  
 power = 0.8  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

#Rounding up, this becomes 176 in each group, or 352 total.

To ensure there is sufficient power to detect a circumstance of zero effect, we conducted an equivalence test. Using the TOSTER package (Caldwell, 2022; Lakens, 2017), an equivalence test was conducted using a significance level of α = 0.05, a sample size of 176 per group, and lower and upper bounds of -0.3 and 0.3 respectively, we have .76 power for an equivalence analysis.

#Equivalence test to confirm if there is no effect.  
library(TOSTER)

Warning: package 'TOSTER' was built under R version 4.2.3

powerTOSTtwo(alpha=0.05, N=176, low\_eqbound\_d=-0.3, high\_eqbound\_d=0.3)

Warning: `powerTOSTtwo()` was deprecated in TOSTER 0.4.0.  
ℹ Please use `power\_t\_TOST()` instead.

The statistical power is 75.78 % for equivalence bounds of -0.3 and 0.3 .

[1] 0.757764

#Again rounding up, 191 per group or 382 in total.

Using a test statistic of t = 4.25, two samples of n = 176, and the effectsize package in RStudio (Ben-Shachar et al., 2020), the result is d = 0.45 [0.24, 0.67]. The effect size is closest to a medium effect size by Cohen’s standards, but can be anywhere from small to medium based on the CI. Whether the estimate is considered precise depends on the context of the analysis; for ours, we were only interested in whether it met the cutoff of a SESOI, that is, an effect size bigger than or equal to 0.3. Since the interval ranges below and above the SESOI, the estimate is not sufficiently precise to say it is of interest.

#Calculating effect size and 95% CI for t = 4.25  
  
library(effectsize)

Warning: package 'effectsize' was built under R version 4.2.3

t <- 4.25   
n1 <- 176   
n2 <- 176  
  
t\_to\_d(t, df\_error = n1+n2-2, paired = FALSE, ci = 0.95)

d | 95% CI  
-------------------  
0.45 | [0.24, 0.67]

**References**

Ben-Shachar M, Lüdecke D, Makowski D (2020). effectsize: Estimation of Effect Size Indices

and Standardized Parameters. Journal of Open Source Software, 5(56), 2815. doi: 10.21105/joss.02815

Plate T, Heiberger R (2016). \_abind: Combine Multidimensional Arrays\_. R package version

1.4-5, <<https://CRAN.R-project.org/package=abind>>

Caldwell, A.R. (2022). Exploring Equivalence Testing with the Updated TOSTER R Package.

PsyArXiv. https://doi.org/10.31234/osf.io/ty8de

Champely S (2020). \_pwr: Basic Functions for Power Analysis\_. R package version1.3-0,

<<https://CRAN.R-project.org/package=pwr>>

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-

analyses. Social Psychological and Personality Science, 8(4), 355-362. doi:10.1177/1948550617697177

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes

A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J,

Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H

(2019). “Welcome to the tidyverse.” \_Journal of Open Source Software\_, \*4\*(43), 1686.

<<https://doi.org/10.21105/joss.01686>>.

Wickham H, Miller E, Smith D (2023). \_haven: Import and Export 'SPSS', 'Stata' and 'SAS'

Files\_. R package version 2.5.4, <<https://CRAN.R-project.org/package=haven>>.