Not-so-standardized effect sizes

Range restriction

Ian Hussey

06 Mai, 2024

Contents

Dependencies & data

Real BDI-II data is taken from Cataldo et al. (2022) Abnormal Evidence Accumulation Underlies the Positive Memory Deficit in Depression, doi: 10.1037/xge0001268.

```
library(tidyverse)
library(janitor)
library(knitr)
library(kableExtra)
library(faux)

data_bdi <- read_csv("bdi_data.csv")
```

Cohen's d from different preselections

Note that in the below, only data at pre is real BDI-II data. Data at post is modified data (i.e., offset by known amounts).

Equivalent change in means, different change in Cohen's d

We know for a fact that the true difference in means is the same in both studies, because we create the data to be this way (i.e., scores at post are exactly pre - 5). The unstandardized effect sizes (pre-post difference in means) are the same, by definition.

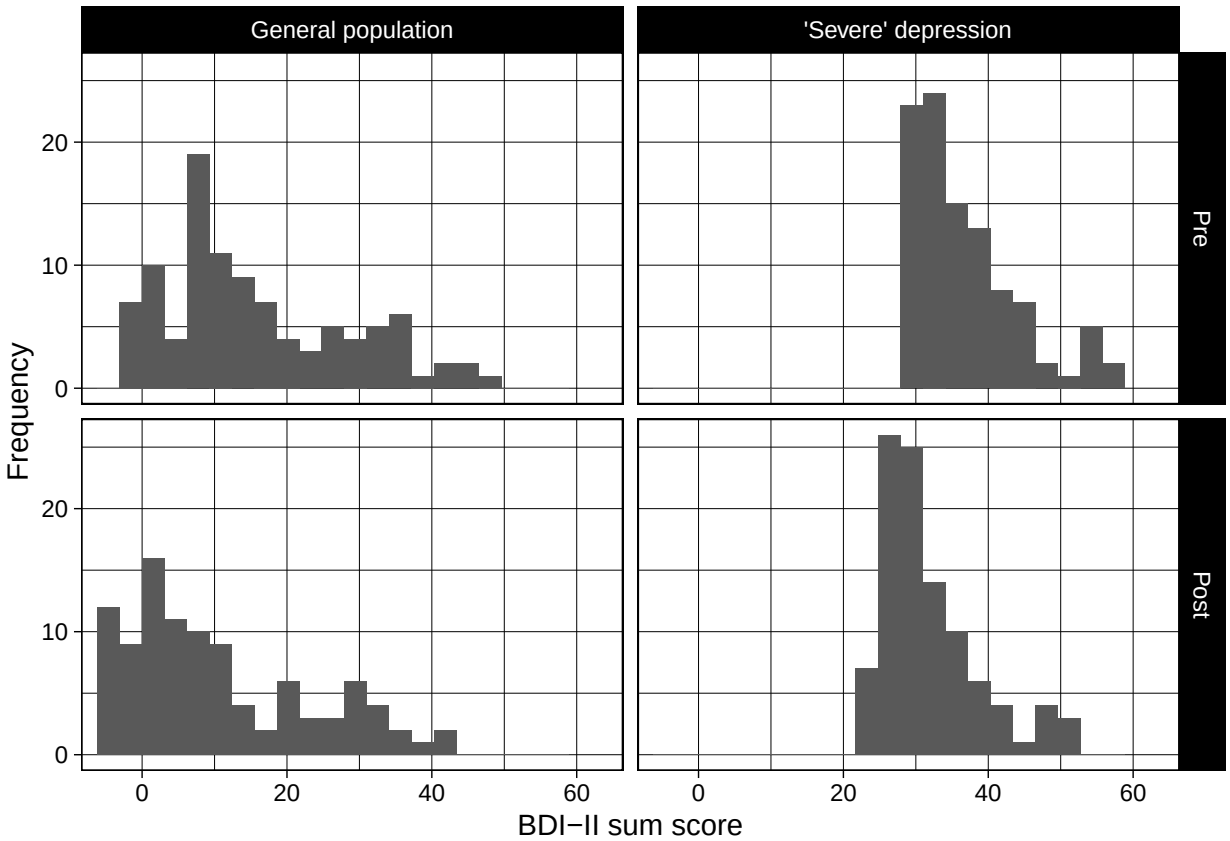
Despite this, the two studies produce the different Cohen's d values. The standardized effect sizes are the different, despite exactly the same pre-post differences between the studies. If the point of standardized effect

sizes is to be able to compare them between studies on a common scale, and they don't do this, what is their point?

```
set.seed(42)
#set.seed(49)

subset_no_preselection <- data_bdi |>
  rename(bdi_pre = bdi_score) |>
  # simulate a 'post' score that is 5 points lower than pre
  mutate(bdi_post = bdi_pre - 5) |>
  #mutate(n = n()) |>
  #mutate(bdi_post = bdi_pre + rnorm(n = n, mean = -5, sd = 1)) |>
  # sample 100 participants from the real data
  slice_sample(n = 100) |>
  mutate(recruitment = "General population")


subset_preselection_for_severe <- data_bdi |>
  rename(bdi_pre = bdi_score) |>
  # simulate a 'post' score that is 5 points lower than pre
  mutate(bdi_post = bdi_pre - 5) |>
  #mutate(n = n()) |>
  #mutate(bdi_post = bdi_pre + rnorm(n = n, mean = -5, sd = 1)) |>
  # simulate recruitment into the study requiring a score of 29 or more at pre ("severe" depression acc
  filter(bdi_pre >= 29) |>
  # sample 100 participants from the pre=selected real data
  slice_sample(n = 100) |>
  mutate(recruitment = "'Severe' depression")


subset_combined <-
  bind_rows(subset_no_preselection,
            subset_preselection_for_severe)

# table of results
subset_estimates <- subset_combined |>
  group_by(recruitment) |>
  summarize(n = n(),
            mean_pre = mean(bdi_pre),
            mean_post = mean(bdi_post),
            sd_pre = sd(bdi_pre),
            sd_post = sd(bdi_post)) |>
  mutate(mean_diff = mean_post - mean_pre,
         cohens_d = (mean_post - mean_pre) / ( (sd_post + sd_pre)/2 )) |>
  select(recruitment,
         n,
         mean_pre, sd_pre, mean_post, sd_post,
         mean_diff,
         cohens_d)

subset_estimates |>
  mutate_if(is.numeric, janitor::round_half_up, digits = 1) |>
  kable() |>
  kable_classic(full_width = FALSE)
```

recruitment	n	mean_pre	sd_pre	mean_post	sd_post	mean_diff	cohens_d
'Severe' depression	100	37.0	7.0	32.0	7.0	-5	-0.7
General population	100	15.7	12.6	10.7	12.6	-5	-0.4

```r
# plot
subset_combined |>
  rename(Pre = bdi_pre, Post = bdi_post) |>
  pivot_longer(cols = c(Pre, Post),
               names_to = "timepoint",
               values_to = "bdi_score") |>
  mutate(timepoint = fct_relevel(timepoint, "Pre", "Post"),
         recruitment = fct_relevel(recruitment, "General population", "'Severe' depression")) |>
  ggplot(aes(bdi_score)) +
  #geom_vline(xintercept = 29, linetype = "dashed") +
  #geom_histogram(boundary = 0, bins = 67) +
  geom_histogram(boundary = 0, bins = 21) +
  scale_fill_viridis_d(begin = 0.3, end = 0.7) +
  theme_linedraw() +
  coord_cartesian(xlim = c(-5, 63)) +
  facet_grid(timepoint ~ recruitment) +
  xlab("BDI-II sum score") +
  ylab("Frequency")
```



Equivalent Cohen's d, different change in means

We know for a fact that the true difference in means is different, because we create the data to be this way (i.e., pre-post difference is -5 in the no preselection study and -3 in the severe depression preselection study). The unstandardized effect sizes (pre-post difference in means) are different, by definition.

Despite this, the two studies produce the same Cohen's d value. The standardized effect sizes are the same, despite genuine differences in the pre-post changes between the two studies.

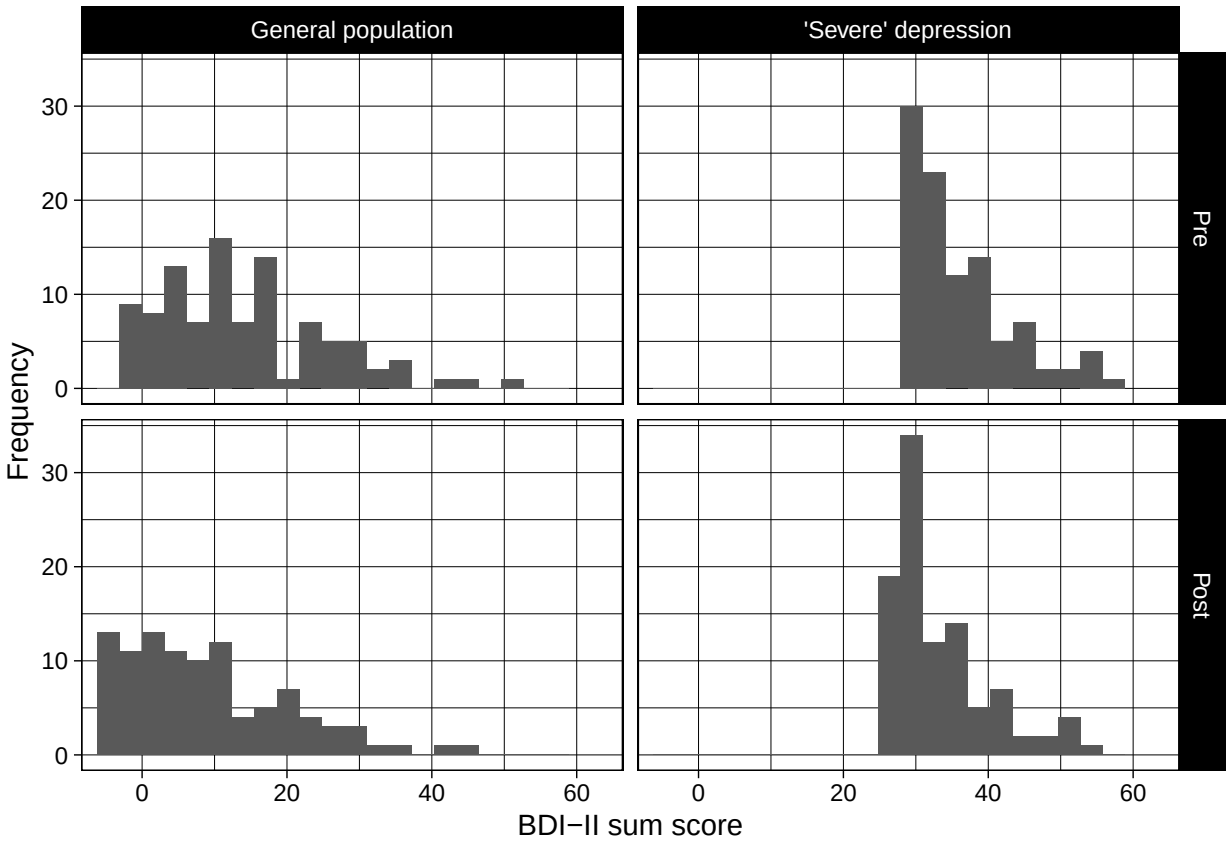
If the same standardized effect size estimate (Cohen's d) can represent different real changes in means, how can a Cohen's d of .2, for example, represent "small" effects? That is, if "small" effects on standardized effect sizes can represent unstandardized effect sizes of different sizes, how are standardized effect sizes 'standardized' at all?

```
set.seed(46)
#set.seed(45)

subset_no_preselection_2 <- data_bdi |>
  rename(bdi_pre = bdi_score) |>
  # simulate a 'post' score that is 5 points lower than pre
  mutate(bdi_post = bdi_pre - 5) |>
  #mutate(n = n()) |>
  #mutate(bdi_post = bdi_pre + rnorm(n = n, mean = -5, sd = 1)) |>
  # sample 100 participants from the real data
  slice_sample(n = 100) |>
  mutate(recruitment = "General population")


subset_preselection_for_severe_2 <- data_bdi |>
  rename(bdi_pre = bdi_score) |>
  # simulate a 'post' score that is 3 points lower than pre
  mutate(bdi_post = bdi_pre - 3) |>
  #mutate(n = n()) |>
  #mutate(bdi_post = bdi_pre + rnorm(n = n, mean = -3, sd = 1)) |>
  # simulate recruitment into the study requiring a score of 29 or more at pre ("severe" depression acc
  filter(bdi_pre >= 29) |>
  # sample 100 participants from the pre=selected real data
  slice_sample(n = 100) |>
  mutate(recruitment = "'Severe' depression")


subset_combined_2 <-
  bind_rows(subset_no_preselection_2,
            subset_preselection_for_severe_2)

# table of results
subset_estimates_2 <- subset_combined_2 |>
  group_by(recruitment) |>
  summarize(n = n(),
            mean_pre = mean(bdi_pre),
            mean_post = mean(bdi_post),
            sd_pre = sd(bdi_pre),
            sd_post = sd(bdi_post)) |>
  mutate(mean_diff = mean_post - mean_pre,
         cohens_d = (mean_post - mean_pre) / ( (sd_post + sd_pre)/2 )) |>
  select(recruitment,
```

```
        n,
        mean_pre, sd_pre, mean_post, sd_post,
        mean_diff,
        cohens_d)

subset_estimates_2 |>
  mutate_if(is.numeric, janitor::round_half_up, digits = 1) |>
  kable() |>
  kable_classic(full_width = FALSE)
```

recruitment	n	mean_pre	sd_pre	mean_post	sd_post	mean_diff	cohens_d
'Severe' depression	100	36.2	6.7	33.2	6.7	-3	-0.4
General population	100	14.2	11.2	9.2	11.2	-5	-0.4

```
# plot
subset_combined_2 |>
  rename(Pre = bdi_pre, Post = bdi_post) |>
  pivot_longer(cols = c(Pre, Post),
               names_to = "timepoint",
               values_to = "bdi_score") |>
  mutate(timepoint = fct_relevel(timepoint, "Pre", "Post"),
         recruitment = fct_relevel(recruitment, "General population", "'Severe' depression")) |>
  ggplot(aes(bdi_score)) +
  #geom_vline(xintercept = 29, linetype = "dashed") +
  #geom_histogram(boundary = 0, bins = 67) +
  geom_histogram(boundary = 0, bins = 21) +
  scale_fill_viridis_d(begin = 0.3, end = 0.7) +
  theme_linedraw() +
  coord_cartesian(xlim = c(-5, 63)) +
  facet_grid(timepoint ~ recruitment) +
  xlab("BDI-II sum score") +
  ylab("Frequency")
```



Explanation of the above results

The above results - where the same unstandarized effect sizes have different standardized effect sizes, or vice-versa - are due to the fact that standardized effect sizes involve dividing, in one way or another, unstandardized effect sizes by standard deviations.

E.g., for Cohen's d:

$d = \frac{M_{intervention} - M_{control}}{SD_{pooled}}$

Most researchers are far more interested in the numerator than the denominator.

- Researchers often care about how much the means differ between the intervention and control groups. Differences in the means determine whether the intervention 'worked' or not.
- They usually care very little about what the SD, except perhaps if they're assessing statistical assumptions (homogeneity of variances).

Despite this, the value of the SDs heavily influences the standardized effect size.

In the above examples, the *range restriction* in the 'severe' depression condition produces a narrower range of scores, and therefore smaller smaller SDs. Dividing the same difference in means by a smaller value of SD produces a different Cohen's d estimate.

Range restrictions like these are extremely common in psychology research, where studies can differ in their inclusion/exclusion strategies. This means makes it far harder to compare 'standardized' effect sizes between studies than you might think.

Solutions to this problem

There are solutions to this, to make "standardized" effect sizes actually standard between studies. But almost no one does them.

1. The when calculating standardized effect sizes, use a well established population norm estimate of the measure's SD rather than the sample SD. E.g., always set the BDI's SD to 12 (or whatever your best estimate is). Note that no implementations of Cohen's d in commonly used R packages recommend this, and only a few can directly handle it (e.g., {esci}).
2. Use math/R packages to correct your standardised effect size estimate for *range restriction* (see Wiernik & Dahlke, 2020, doi: 10.1177/2515245919885611).

Is this issue limited to Cohen's d?

No, it affects other forms of standardized effect sizes too, including correlations.

E.g., there is a perennial debate in the US about whether standardized university entrance tests like the SAT are useful or not, or indeed are biased or not (e.g., between gender and race/ethnicity), because straightforward analyses suggest that SAT scores (used to get a place at university) are poorly predictive of grades at university.

However, this poor predictive validity may be due in part to range restriction: because the SAT scores are used to determine who goes to university, data on university grades is only obtained from those individuals who already scored highly on the SAT. That is, there is a fairly narrow range of SAT scores among university students. Correlations, like Cohen's d, include SD in their denominator (i.e., $r = covariance_{xy}/(SD_x * SD_y)$), and therefore range restriction also distorts correlations.

It is therefore possible - indeed, likely - that SAT scores are usefully predictive of grades at university. The below short simulation demonstrates attentuation in correlations due to range constraint.

```
# Set seed for reproducibility
set.seed(42)

# Parameters
n <- 10000  # number of observations
rho <- 0.6  # correlation between x and y

# Generate correlated data using the faux package
simulated_data <- rnorm_multi(n = n,
                              mu = c(0, 0),
                              sd = c(1, 1),
                              r = matrix(c(1, rho,
                                           rho, 1), nrow = 2),
                              varnames = c("x", "y"))

# Calculate correlation in full data
full_correlation <- cor(simulated_data$x, simulated_data$y)
cat("Correlation in full data:", janitor::round_half_up(full_correlation, digits = 2), "\n")
```

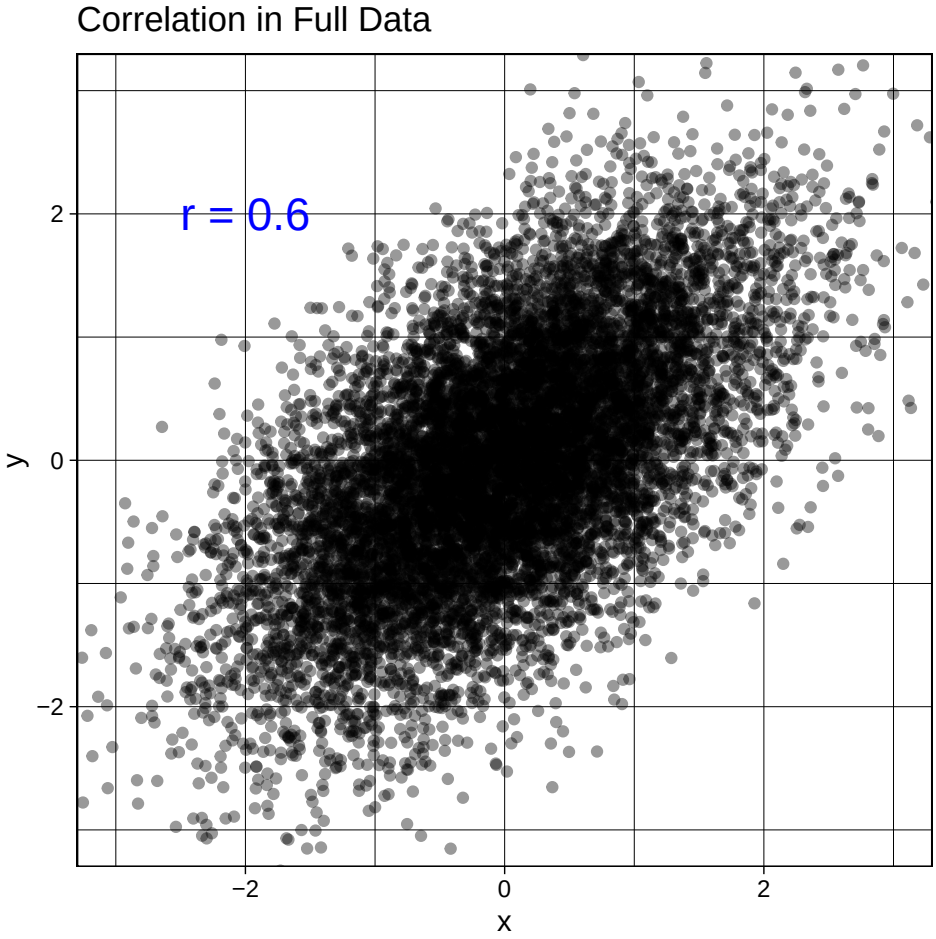
```
## Correlation in full data: 0.6
```

```
# Introduce range restriction (e.g., keep only x > -0.5 and x < 0.5)
simulated_data_range_restricted <- simulated_data |>
  filter(x > qnorm(0.75)) # top 25% of a normal population corresponds to SD > qnorm(0.75), ie 0.674489

# Calculate correlation in restricted data
```

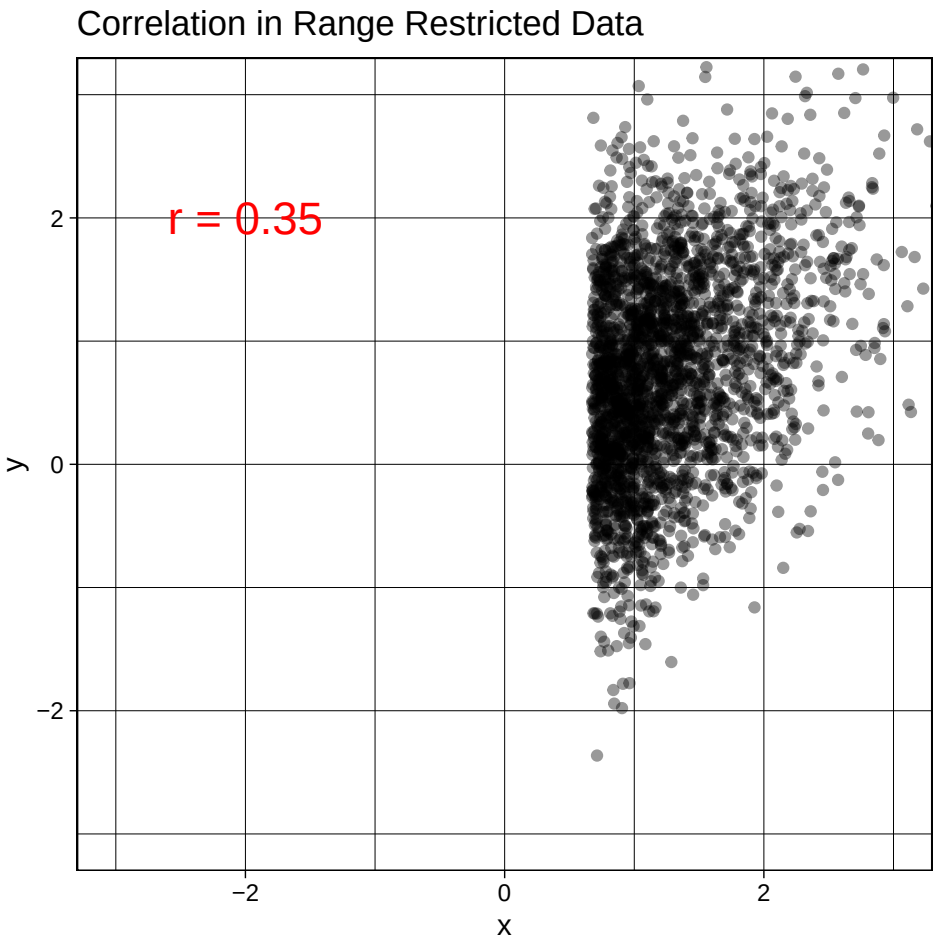
```r
restricted_correlation <- cor(simulated_data_range_restricted$x, simulated_data_range_restricted$y)
cat("Correlation in restricted data:", janitor::round_half_up(restricted_correlation, digits = 2), "\n")
```

```
## Correlation in restricted data: 0.35
```

```r
# Plot full data with correlation annotation
ggplot(simulated_data, aes(x = x, y = y)) +
  geom_point(alpha = 0.4) +
  #geom_smooth(method = "lm", se = FALSE, color = "blue") +
  ggtitle("Correlation in Full Data") +
  theme_linedraw() +
  annotate("text", x = -2, y = 2, label = paste("r =", round(full_correlation, 2)),
           hjust = 0.5, vjust = 0.5, size = 6, color = "blue") +
  coord_cartesian(xlim = c(-3, 3), ylim = c(-3, 3))
```




```r
# Plot restricted data with correlation annotation
ggplot(simulated_data_range_restricted, aes(x = x, y = y)) +
  geom_point(alpha = 0.4) +
  #geom_smooth(method = "lm", se = FALSE, color = "red") +
  ggtitle("Correlation in Range Restricted Data") +
  theme_linedraw() +
  annotate("text", x = -2, y = 2, label = paste("r =", round(restricted_correlation, 2)),
           hjust = 0.5, vjust = 0.5, size = 6, color = "red") +
  coord_cartesian(xlim = c(-3, 3), ylim = c(-3, 3))
```



Note that the observed correlations which have been distorted due to range restriction can be 'de-attentuated' or corrected if normative data is available to know what the unrestricted range looks like. However, this is very rarely done in studies and meta-analyses.

```
# Calculate the variance ratios as an estimate of the range restriction factor
variance_ratio <- var(simulated_data_range_restricted$x) / var(simulated_data$x)

# Deattenuate the observed correlation
corrected_correlation <- restricted_correlation / sqrt(variance_ratio)

# Output results
cat("Observed Correlation (Restricted):", janitor::round_half_up(restricted_correlation, 2), "\n")
```

```
## Observed Correlation (Restricted): 0.35
```

```
cat("Variance Ratio (Range Restriction Factor):", janitor::round_half_up(variance_ratio, 2), "\n")
```

```
## Variance Ratio (Range Restriction Factor): 0.25
```

```
cat("Corrected Correlation (Deattenuated):", janitor::round_half_up(corrected_correlation, 2), "\n")
```

```
## Corrected Correlation (Deattenuated): 0.69
```

Note that the corrected correlation is much closer to the original one.