Reproducible RMarkdown with GitHub HW:

V2 of IQ Report w/o `999` Outlier (actually `99`)

Lorenzo Pla Serrano

2025-11-30

Abstract

In this report, the relationship between residential proximity to a lead-emitting smelter and IQ scores in children was briefly explored using a subset of data from a 1975 CDC study. After changing the identified `999` outlier to the correct `99` IQ score, descriptive statistics and visualizations revealed that children who lived farther from the smelter had a slightly higher mean IQ than those who lived nearby, though formal inference was not conducted. Both groups displayed approximately normal IQ distributions with comparable variability.

Background

Lead is a highly toxic substance that affects nearly every organ system in the body when ingested directly. At lower, indirect exposure levels, the primary biological effect is damage to the nervous system, though the threshold for safe lead exposure remains a subject of scientific debate. To examine the relationship between low-level lead absorption and neurological function, researchers led by the CDC conducted a study of children aged 3 to 15 years in El Paso, Texas, who lived at varying distances from a large lead-emitting ore smelter (Landrigan et al., 1975). This brief exploratory analysis uses a subset of the original data, comprising 124 observations without any filtering and consisting of just two variables: residential proximity to the smelter (categorized as NEAR if within 1 mile or FAR otherwise) and IQ scores measured using the Wechsler Intelligence Scale for Children (WISC). The data come from the study *Neuropsychological dysfunction in children with chronic low-level lead absorption*, published in The Lancet.

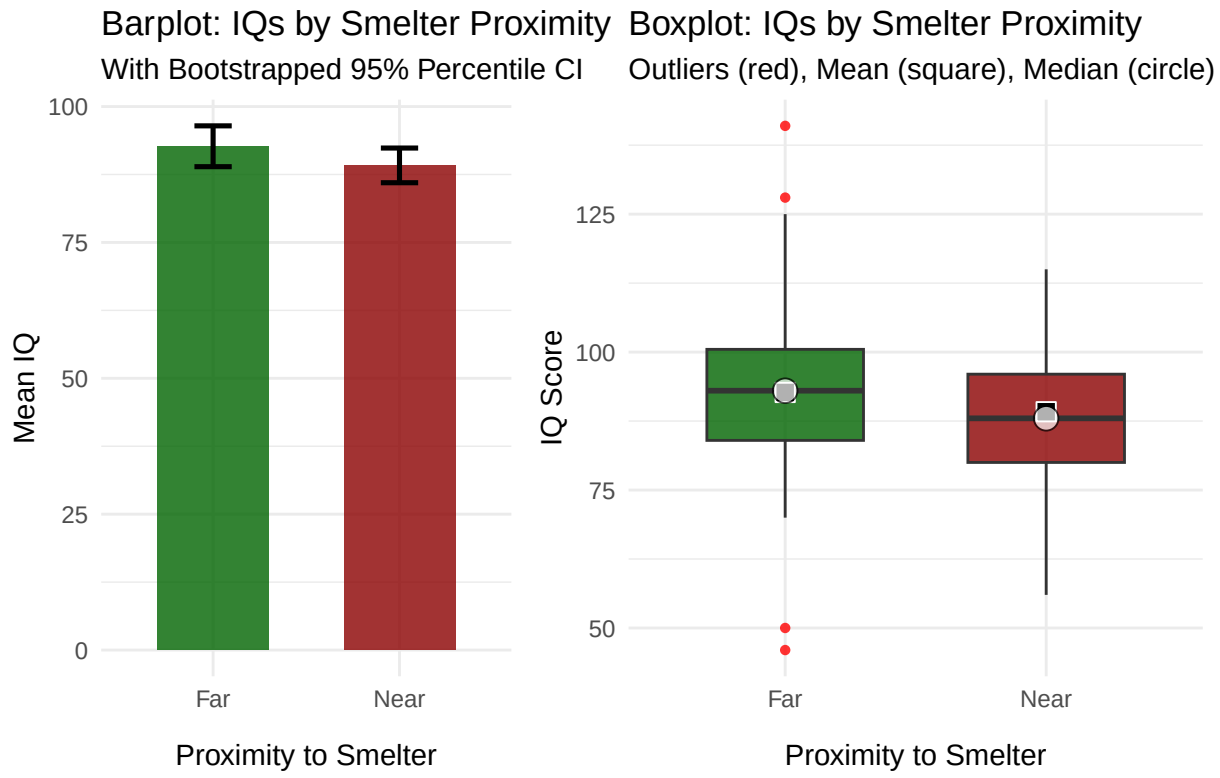
Comparing IQ Scores of Children Near or Far to Smelter

To compare the IQ levels of children who lived NEAR the smelter (within 1 mile) and were assumed to have greater lead exposure versus children who lived FAR and were presumably less exposed, a barplot and boxplot were created side by side in **Figure 1** to evaluate the distribution of IQ for each group. The barplot includes a bootstrapped percentile confidence interval to compare the variability between groups, whereas the boxplot displays outliers and indicates potential skewness. Common descriptive statistics were also included in **Table 1** (shown first).

Table 1: **IQ Scores of Children by Their Smelter Proximity**

Variable	N	Smelter Proximity	
		Far N = 67	**Near** N = 57
IQ Score	124		
Mean (SD)		92.7 (16.0)	89.2 (12.2)
Median (Q1, Q3)		93.0 (83.0, 101.0)	88.0 (80.0, 96.0)
Range (Min, Max)		46.0, 141.0	56.0, 115.0

Figure 1: **Distribution of IQs by Smelter Proximity**



Discussion & Conclusions

The descriptive statistics in **Table 1** and visualizations in **Figure 1** show that after correcting the `999` value to `99` (was confirmed with the PI as a data entry error), the `FAR` group now displays more realistic IQ statistics. The calculated mean IQ score for children `FAR` from a smelter was 92.7, whereas for children `NEAR` a smelter the mean IQ score was 89.2. In **Figure 1**, the barplot shows that both groups now have comparable confidence intervals, with the `FAR` group displaying slightly more variability than the `NEAR` group. The boxplot now allows for a meaningful comparison between groups, revealing that both distributions appear relatively symmetric with the mean and median closely overlapping in each group, suggesting approximately normal IQ distributions. The `FAR` group shows a few outliers at both the lower and upper ends of the distribution, while the `NEAR` group appears more compact. Overall, children who lived farther from the smelter had a slightly higher mean IQ score than those who lived nearby, though formal statistical testing would be required to determine if this difference is statistically significant.

Code Appendix

```r
# loading packages
library(knitr)
library(tidyverse)
library(ggplot2)
library(patchwork) # used to combine plots into one

# defining setup options
opts_chunk$set(tidy = F)

# NOTE: All r chunks are displayed in the Appendix section at the end

# NOTE: Before creating this RMarkdown file, the folder structure for this project
# was established using the `CIDAtools` package.
# See "/Code/creating_project_wCIDAtools.R",
# which was run once to set up the directory structure.
#
# Additionally, because this RMarkdown file was written on Posit Cloud
# (browser-based RStudio),
# the instructions in "/Code/connecting_PositCloud_wGitHub.R" were followed to ensure
# proper version control and reproducibility.

# NOTE, one must commit before pushing, which sends committed changes to GitHub
# loading IQ data
IQ_data <- read.csv("../DataRaw/lead-iq-01.csv")

# changing columns to appropriate data type
IQ_data <- IQ_data %>%
  mutate(Smelter = as.factor(Smelter),
         IQ = as.integer(IQ))

n <- nrow(IQ_data)
# changing `999` value to `99` after a brief conversation with the PI
IQ_data <- IQ_data %>%
  mutate(IQ = ifelse(IQ == 999, 99, IQ))
library(ggplot2)
library(dplyr)
library(patchwork)

# establishing colors for plots
smelter_colors <- c("Far" = "darkgreen",
                    "Near" = "darkred")

# creating box plot
p_box <- ggplot(IQ_data, aes(x = Smelter, y = IQ, fill = Smelter)) +
  geom_boxplot(outlier.colour = "red",
               outlier.shape = 16,
               width = 0.6,
               alpha = 0.80) +
```

```r
  # mean points
  stat_summary(
    fun = mean,
    geom = "point",
    shape = 22,
    size = 3.5,
    fill = "black",
    color = "white",
    alpha = 1
  ) +

  # median points
  stat_summary(
    fun = median,
    geom = "point",
    shape = 21,
    size = 4,
    fill = "white",
    color = "black",
    alpha = 0.7
  ) +

  scale_fill_manual(values = smelter_colors) +
  labs(
    x = "Proximity to Smelter",
    y = "IQ Score",
    title = "Boxplot: IQs by Smelter Proximity",
    subtitle = "Outliers (red), Mean (square), Median (circle)"
  ) +
  theme_minimal(base_size = 11) +
  theme(
    legend.position = "none",
    panel.grid.minor.x = element_blank(),
    axis.title.x = element_text(margin = margin(t = 12))    # <-- added spacing
  )

# creating bootstrapped bar plot
set.seed(123)

bootstrap_ci <- IQ_data %>%
  group_by(Smelter) %>%
  summarise(
    mean_IQ = mean(IQ),
    boot_mean = list(replicate(2000, mean(sample(IQ, replace = TRUE))))
  ) %>%
  mutate(
    CI_lower = sapply(boot_mean, function(x) quantile(x, 0.025)),
    CI_upper = sapply(boot_mean, function(x) quantile(x, 0.975))
  )
```

```r
p_bar <- ggplot(bootstrap_ci, aes(x = Smelter, y = mean_IQ, fill = Smelter)) +
  geom_col(width = 0.6, alpha = 0.80) +
  geom_errorbar(
    aes(ymin = CI_lower, ymax = CI_upper),
    width = 0.2,
    linewidth = 0.9
  ) +
  scale_fill_manual(values = smelter_colors) +
  labs(
    title = "Barplot: IQs by Smelter Proximity",
    subtitle = "With Bootstrapped 95% Percentile CI",
    x = "Proximity to Smelter",
    y = "Mean IQ"
  ) +
  theme_minimal(base_size = 11) +
  theme(
    legend.position = "none",
    panel.grid.minor.x = element_blank(),
    axis.title.x = element_text(margin = margin(t = 12))    # <-- added spacing
  )

# displaying plots side-by-side
library(ggtext) # for correct bolding in title

(p_bar + p_box + plot_layout(widths = c(5, 7))) +
  plot_annotation(
    title = "Figure 1: **Distribution of IQs by Smelter Proximity**",
    theme = theme(
      plot.title = element_markdown(
        hjust = 0.5,
        size = 12
      )
    )
  )

# loading gtsummary
library(gtsummary)

# gtsummary table with full descriptive statistics
IQ_data %>%
  tbl_summary(
    by = Smelter,
    include = IQ,
    statistic = all_continuous() ~ c(
      "{mean} ({sd})",
      "{median} ({p25}, {p75})",
      "{min}, {max}"
    ),
    type = all_continuous() ~ "continuous2",
```

```
    label = IQ ~ "IQ Score",
    digits = all_continuous() ~ 1    # <-- ensures at least 1 decimal place
  ) %>%
  add_stat_label(
    label = all_continuous() ~ c(
      "Mean (SD)",
      "Median (Q1, Q3)",
      "Range (Min, Max)"
    )
  ) %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(all_stat_cols() ~ "**Smelter Proximity**") %>%
  modify_caption("**IQ Scores of Children by Their Smelter Proximity**") %>%
  bold_labels()
# calculating means by Smelter
mean_far <- mean(IQ_data$IQ[IQ_data$Smelter == "Far"])
mean_near <- mean(IQ_data$IQ[IQ_data$Smelter == "Near"])
```

GitHub Project Repository (public): https://github.com/PostData-solutions/Pla_Project_01