

Example report: cloud seeding

The purpose of this document is to both serve as a template for reports created in this class, and to explain what should be found in each section of a statistical report. Note that this is a general example - exactly what a report includes will vary with the context. In this document, we break down a report about a data set that describes the impact of cloud seeding on rainfall.

Introduction

Overview

The introduction should describe how the data were collected and clearly articulate the research question.

- introduce data collection (briefly)
- define the research question in a way that implies the statistical test that will be used

Coding tips

- in code chunk options, you can set `include = F` to hide code and output, which is useful for librarying packages and loading in data

Does massive injection of silver iodide into cumulus clouds lead to increased rainfall? Between 1968 and 1972, researchers in southern Florida conducted a randomized experiment in which clouds on 52 suitable days were randomly assigned to be seeded with silver iodide or left unseeded. Then, using radar, the researchers recorded the total volume of rainfall from each cloud. We seek to determine whether the average total volume of rainfall from the seeded clouds exceeds that from the unseeded clouds.

Statistical procedures used

Overview

The statistical procedures section serves a couple of purposes. First, the data should be explained in greater detail, such that readers should understand the structure without having the Excel file. To achieve this goal, we often provide exploratory visualizations of the raw data, and explain how they relate to the research question. Second, all statistical procedures used should be explained in detail such that the work is completely reproducible.

- define the observational units, response, and explanatory variable(s)
- provide a visualization of the raw data and discuss what you see
- clearly articulate the hypotheses for your test and define all relevant symbols
- address the assumptions for the statistical method you are using
 - this step can be quite involved, particularly if a transformation is required
 - if a transformation is required, address the assumptions **before** and **after** the transformation
 - diagnostic plots (resid vs fitted, QQ, etc.) should be included in the appendix but **referenced in the body**

Coding tips

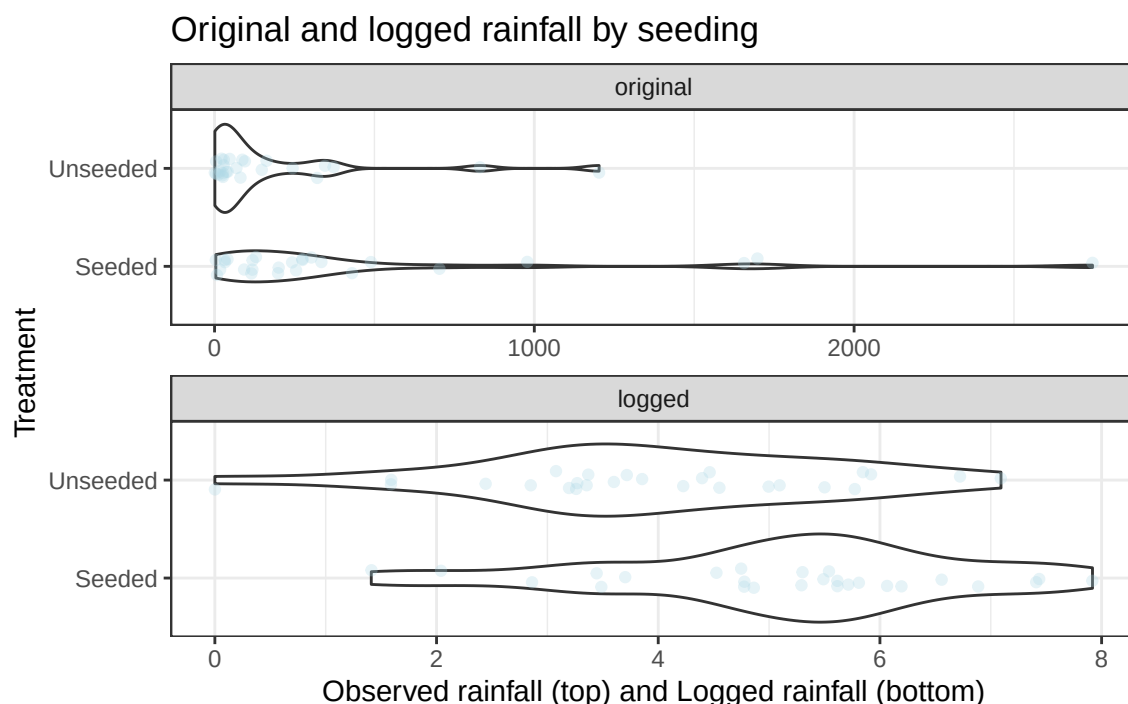
- in code chunk options, give the code chunk a name (such as `eda`); you can then easily repeat the code from this chunk in the appendix.
- set `echo = F`
- use `fig.dim = c()` to set the dimensions of any figures you'd like to include in this part of the report

In this study, the response variable is the total volume of rainfall and the explanatory variable is whether the cloud was seeded; each of these variables were measured on the 52 clouds deemed suitable for cloud seeding, which represent the observational unit in this study. To address the research question, we conducted a one-sided two-sample t-test, which tests the following hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

where μ_1 represents the true mean volume of rainfall in the seeded group and μ_2 represents the true mean volume of rainfall in the unseeded group. This procedure relies on three assumptions: (1) independence of observations, (2) normality of errors, and (3) equality of variance between groups. In regards to independence, we do not see any egregious violations; since clouds were (presumably) only measured once, and observations were not collected sequentially over time, there is no reason to believe that rainfall across days would be related to one another. To examine the remaining assumptions, we plotted rainfall distributions for seeded and unseeded clouds, both on the raw scale and after a log transformation. The raw data show a strong right skew in both groups, with noticeably greater variability in the seeded group.



These features suggest violations of the normality and equal-variance assumptions. Additionally, the quantile-quantile plot of the raw data (see Appendix) deviates sharply from the hypothesized normal line, consistent with a violation of normality. Similarly, the larger spread in the seeded group, visible in both violin plots and the residuals vs. fitted plot (see Appendix), indicates heterogeneity of variance. Under both of these likely violations, our t-based inference is suspect at best.

To address this issue, we applied a natural log transformation to the rainfall data. This reduced skewness and produced violin plots with more symmetric shapes and comparable spreads. The QQ plot of the transformed data aligns well with the hypothesized normal line, and the residuals vs. fitted plot shows roughly constant variance (see Appendix). Thus, the transformed data are not inconsistent with the normality and equal-variance assumptions, and we proceeded with t-based procedures on the log-transformed values.

Summary of Statistical Findings

Overview

The statistical findings section should generally be fairly succinct. It should address the question of interest based on the analysis, specifically referencing your evidence for a claim.

- in most cases, this is likely the result of a statistical test
- in some cases, this might be a specific figure

Generally, when describing results of statistical tests, you should:

- provide an evidence statement, being sure to include the type of test, p-value, test statistic, and reference distribution
- provide an estimate statement, being sure to include the level of confidence; if a transformation was performed, this should be interpreted on the original scale (where applicable)

If your conclusion is based on figures or other evidence, be sure to explicitly reference that evidence.

Coding tips

- same advice as in the previous section

We conducted a one-sided two sample t-test on 50 degrees of freedom, which yielded a test statistic of $t = 2.544$, with an associated p-value of 0.0141. This provides strong evidence to suggest that the mean log-rainfall for the seeded clouds is greater than that of the unseeded clouds. On the original scale, it is estimated that the median total volume of rainfall for the seeded group is 3.14 times the median volume of rainfall for the unseeded group, with a 95% confidence interval of 1.27 to 7.74.

Scope of Inference

Overview

The scope of inference should describe two things, in two separate statements:

- to whom you can generalize results and why
- whether you can infer any causal relationships and why

When applicable, you should also discuss opportunities for improvement in this section (usually with respect to data collection).

Since a random mechanism was used to determine whether clouds were seeded, we can conclude that the differences in rainfall that we observed were caused by the seeding. However, there is no reason to believe that the days on which clouds were seeded were randomly selected over the course of the four year experiment. Therefore, these results apply to only the 52 days deemed suitable to cloud seed, and only in the area in which the study was conducted.

Acknowledgements and author contributions

Overview

This section exists to acknowledge help you received, either from AI or other people, and describe who did what.

Meaghan and I both wrote the report together, and she made the exploratory graphics while I did the analysis with the help of ChatGPT 5.0, which was used to troubleshoot code to run the t-test.

Appendix

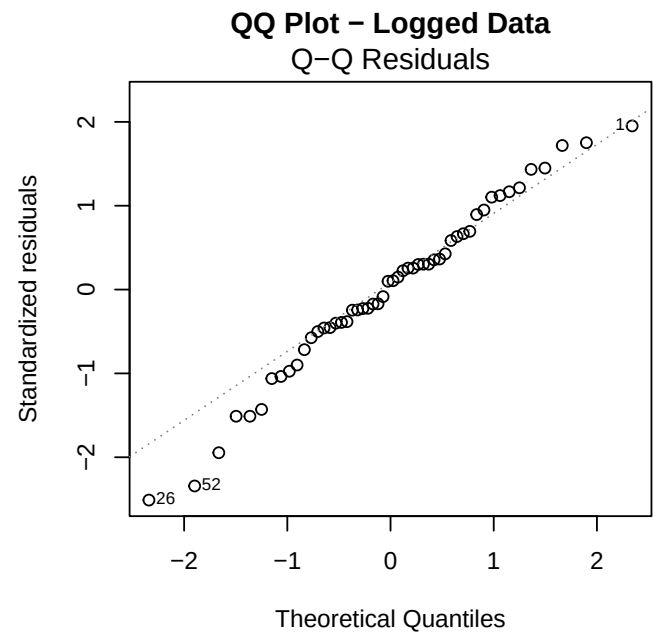
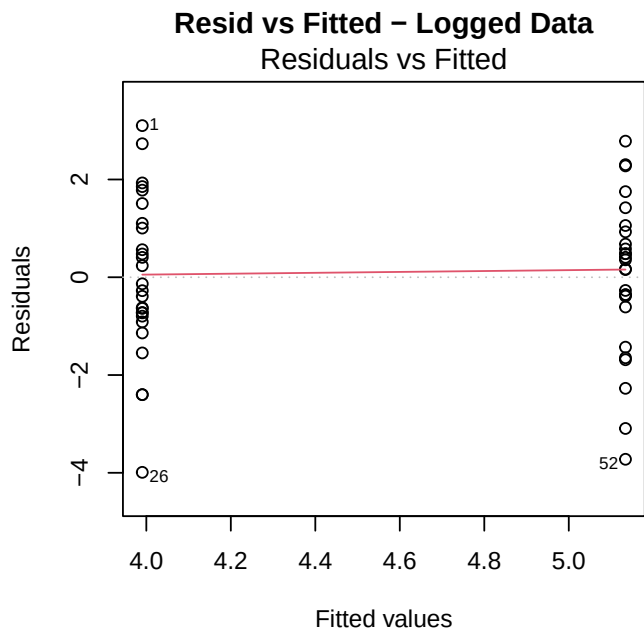
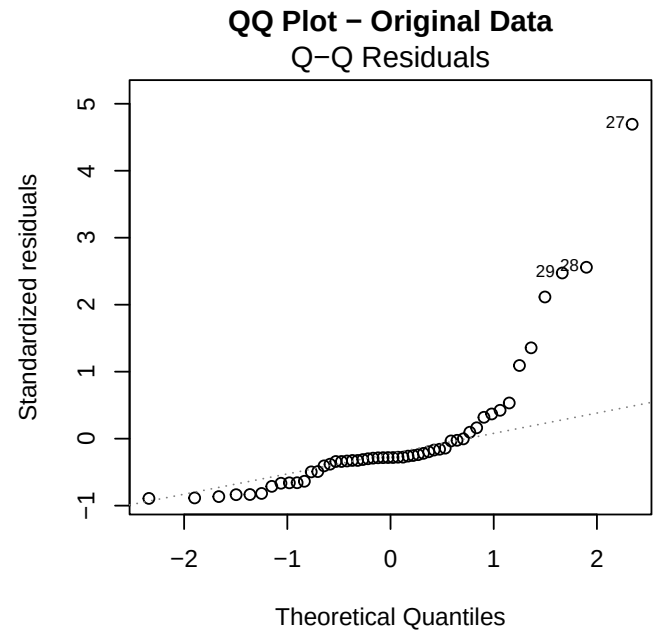
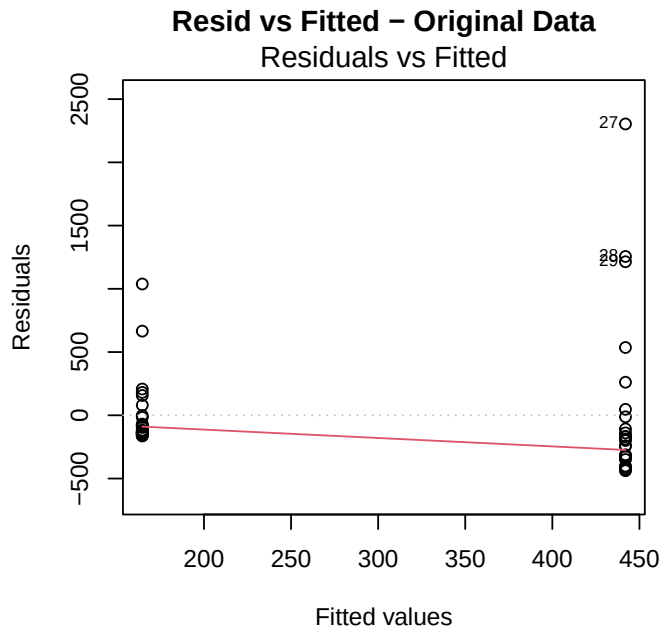
Overview

This section should provide all the code and plots used to create the analysis, if the plots were not included in the body of the document. Be sure to label your plots clearly.

Coding tips

- you can repeat the labeled code chunks from above, this time setting `eval = F` in the chunk options to easily repeat the code you used.
- code should be commented

Plots



Code

```
# create dataframe for ggplot
rainfall <- c(cloud$Rainfall, log(cloud$Rainfall))
```

```
treatment <- rep(cloud$Treatment, 2)
resp_type <- factor(c(rep('original', nrow(cloud)), rep('logged', nrow(cloud))))

plot_df <- tibble(rainfall, treatment, resp_type)
plot_df$resp_type <- relevel(plot_df$resp_type, ref = "original")

# plot
ggplot(plot_df, aes(x = treatment, y = rainfall)) +
  geom_violin() +
  geom_jitter(width = 0.1, alpha = 0.3, color = "lightblue") +
  facet_wrap(~ resp_type, scales = "free", nrow = 2) +
  coord_flip() +
  labs(title = 'Original and logged rainfall by seeding',
       x = 'Treatment',
       y = 'Observed rainfall (top) and Logged rainfall (bottom)') +
  theme_bw()
```

```
cloud$Treatment <- relevel(cloud$Treatment, ref = 'Unseeded') #relevel to match book
cloud.mod <- lm(log(Rainfall) ~ Treatment, cloud) #fit model
summary(cloud.mod) #summary
exp(1.1438) #estimate on original scale
exp(confint(cloud.mod)) #95% ci on original scale
```

```
#create diagnostic plots for original and transformed data
par(mfrow = c(2,2))
plot(
  lm(Rainfall ~ Treatment, data = cloud), which = 1,
  main = 'Resid vs Fitted - Original Data'
)
plot(
  lm(Rainfall ~ Treatment, data = cloud), which = 2,
  main = 'QQ Plot - Original Data'
)
plot(
  lm(log(Rainfall) ~ Treatment, data = cloud), which = 1,
  main = 'Resid vs Fitted - Logged Data'
)
plot(
  lm(log(Rainfall) ~ Treatment, data = cloud), which = 2,
  main = 'QQ Plot - Logged Data'
)
```