# Math 472: Homework 11

*Due Wednesday, May 6*

**Problem 1** (Exercise 10.117)**.** Lord Rayleigh was one of the earliest scientists to study the density of nitrogen.[1] He employed two types of methods to produce nitrogen gas: (1) a process to derive nitrogen from atmospheric gas via removal of oxygen, and (2) a process which produced nitrogen gas from chemical compounds not involving atmospheric gas.

In his studies, he noticed something peculiar. The nitrogen densities produced from chemical compounds tended to be smaller than the densities of nitrogen produced from atmospheric gas. Lord Rayleigh's measurements are given in the following table. These measurements correspond to the mass of nitrogen filling a flask of specified volume under specified temperature and pressure.

| Compound Chemical | Atmosphere |
|---|---|
| 2.30143 | 2.31017 |
| 2.29890 | 2.30986 |
| 2.29816 | 2.31010 |
| 2.30182 | 2.31001 |
| 2.29869 | 2.31024 |
| 2.29940 | 2.31010 |
| 2.29849 | 2.31028 |
| 2.29889 | 2.31163 |
| 2.30074 | 2.30956 |
| 2.30054 | |

(a) For the measurements from the chemical compound, the sample mean is $\overline{y} = 2.29971$ and the sample standard deviation $s = .001310$; for the measurements from the atmosphere, $\overline{y} = 2.310217$ and $s = .000574$. Is there sufficient evidence to indicate a difference in the mean mass of nitrogen per flask for chemical compounds and air? What can be said about the p-value associated with your test?

(b) Find a 95% confidence interval for the difference in mean mass of nitrogen per flask for chemical compounds and air.

(c) Based on your answer to part (b), at the $\alpha = .05$ level of significance, is there sufficient evidence to indicate a difference in mean mass of nitrogen per flask for measurements from chemical compounds and air?

(d) Is there any conflict between your conclusions in parts (a) and (b)? Although the difference in these mean nitrogen masses is small, Lord Rayleigh emphasized this difference rather than ignoring it, and this led to the discovery of inert gases like Argon in the atmosphere.

**Problem 2.** The octane number $Y$ of refined petroleum is related to the temperature $x$ of the refining process, but it is also related to the particle size of the catalyst. An experiment with a small-particle catalyst gave a fitted least-squares line of

$$\widehat{y} = 9.360 + .155x,$$

with $n = 31$, $\text{Var}(\widehat{\beta}_1) = (.0202)^2$, and SSE = 2.04. An independent experiment with a large-particle catalyst gave

$$\widehat{y} = 4.265 + .190x,$$

with $n = 11$, $\text{Var}(\widehat{\beta}_1) = (.0193)^2$ and SSE = 1.86.

Test the hypotheses that the slopes are significantly different from zero, with each test at the significance level of .05.

[1] Lord Rayleigh, On an Anomaly Encountered in Determinations of the Density of Nitrogen Gas, *Proceedings of the Royal Society of London*, (1894), `http://www.jstor.org/stable/115479`

**Problem 3.** A recent study of collected fossil data[2] investigated the relationship between species' traits and their risk of going extinct. In this problem we'll analyze a subset of their dataset, specifically a subset of 550 trilobita species, with fossil samples taken from various epochs. In particular, we'll look at data consisting of $n_1 = 40$ samples of trilobite fossils which date back to epochs of known mass extinctions, and compare them with $n_2 = 13070$ samples from non-mass-extinction epochs.

Download the trilobita fossil datafiles from

`https://github.com/max-hill/math-472/blob/main/datasets/trilobita-fossil-mass-extinctions.csv`

`https://github.com/max-hill/math-472/blob/main/datasets/trilobita-fossil-background.csv`

The first dataset consists of fossil samples dating to geologic epochs coinciding with mass extinction events. The second dataset consists of samples from "ordinary times" (non-mass-extinction epochs).

Each row of these datasets correspond to a fossil sample. The column labelled `extinct` is a boolean variable indicating whether the species represented by that fossil went extinction during that epoch. The columns `log_range` and `log_bodysize` correspond to the relative geographic range of the species and the body size of the fossil sample, respectively. (Here, `extinct` is the response variable $Y$; the covariates $x_1$ and $x_2$ are `log_range` and `log_bodysize`).

(a) Load the background dataset as a dataframe in `R` and run a logistic regression using

```
df_BG = read.csv("trilobita-fossil-background.csv")
fit_BG <- glm(extinct ~ log_range + log_bodysize, data=df_BG, family=binomial())
```

You can view the results of the logistic regression by running:

```
fit_BG
summary(fit_BG)
```

Interpret your results. What is the best fit logistic regression model that you got? What were the estimated coefficients. Which estimates were statistically significant and which were not? How do you interpret the $p$-values?

(b) Can you draw any inferences about the relationship between (1) body size and extinction risk, or (2) between geographic range and body size for this dataset? Consider the signs and magnitudes of your estimates $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ and their $p$-values.

(c) Perform a logistic regression analysis of the mass-extinction epoch dataset. Answer questions (a) and (b) for the fossil samples from the mass extinction epochs.

(d) Make a comparison between the two logistic regression analyses. How does the effect of body size and geographic range on extinction risk differ between (1) eras of mass extinction and (2) eras without mass extinction? Can you draw any biologically interesting conclusions? Explain your results in a way that a biologist would understand.

[2] P. Monarrez, et al. Reduced strength and increased variability of extinction selectivity during mass extinctions, *R Soc Open Sci* (2023) `https://royalsocietypublishing.org/rsos/article/10/9/230795/92261`