HOMEWORK 2 - SOLUTIONS

PSTAT 100 - DATA SCIENCE: CONCEPTS AND ANALYSIS

INSTRUCTOR: Ethan P. Marzban

Submission Instructions

This homework assignment consists of a mix of written and coding questions.

Written Portion

- Please show all of your work
- Answers may be handwritten or typeset (using LaTeX, Word, etc.)
- Please write legibly; if the grader cannot read your work, you will not receive full marks.

Coding Portion

- Please make sure to interpret **all** code outputs.
  - As a general rule-of-thumb: if there is a code chunk whose output is not being interpreted, you should move the code chunk to an Appendix.

Final Submission

- You should combine your written and coding answers into a **single** PDF, which you upload to Gradescope.
  - Here is a free online resource to help you merge PDFs.
  - Please note: Gradescope will only allow you to upload a single PDF.
- Ensure you match pages in your Gradescope submission; failure to do so may incur point penalties.

Due Date

You must upload your homework to Gradescope by no later than **11:59 pm on Sunday, July 20, 2025**.

Information on Grading

- A handful of parts will be selected from this homework to be graded on correctness; these parts will be graded collectively out of 12 points.
  - We will not reveal which parts are to be graded upon correctness until after the homework is graded, so please attempt all problems!
- You will be assigned 2 additional points for submitting the *entirety* of your homework, and 1 additional point for matching pages on your gradescope submission.
  - As such, if you fail to submit an attempt for all parts and fail to match pages, you will not receive anything above an 80%.

Written Portion

Problem 1: Simple Random Sampling Without Replacement

Motivation

In this problem, we develop a slightly more mathematical framework of simple random sampling without replacement. I encourage you to use this problem as:

- A review of some requisite PSTAT 120A (probability) knowledge
- Practice with manipulating sums and double sums

Let $S := \{\xi_1, \cdots, \xi_m\}$ be a set consisting of m distinct elements, and let $\vec{\boldsymbol{X}} := (X_1, \cdots, X_n)$ denote a random sample taken from S without replacement such that all subsets of size n, taken from S, are equally likely. Assume n, our **sample size,** is no larger than m, the number of elements in S.

If we view S as a **population**, this framework models $\vec{\boldsymbol{X}}$ as a **simple random sample, without replacement** taken from the population.

Define the **population mean** μ and **population variance** σ as

$$\mu := \frac{1}{m}\sum_{i=1}^{m} \xi_i; \qquad \sigma^2 := \frac{1}{m}\sum_{i=1}^{m}(\xi_i - \mu)^2$$

It may be useful to also define the **popualtion total** T to be the sum of all elements in S; that is, $T := \sum_{i=1}^{m} \xi_i$

(a) Show that $\sum_{i=1}^{n} \xi_i^2 = m(\sigma^2 + \mu^2)$. **Hint:** Consider starting with the definition of σ^2, expanding the square, and then simplifying terms.

Solution:

$$\begin{aligned}
\sigma^2 &:= \frac{1}{m}\sum_{i=1}^{m}(\xi_i - \mu)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}(\xi_i^2 - 2\xi_i\mu + \mu^2) \\
&= \frac{1}{m}\left[\sum_{i=1}^{m}\xi_i^2 - 2\mu\sum_{i=1}^{m}\xi_i + \mu^2\sum_{i=1}^{m}(1)\right] \\
&= \frac{1}{m}\left[\sum_{i=1}^{m}\xi_i^2 - 2\mu(m\mu) + m\mu^2\right] \\
&= \frac{1}{m}\left[\sum_{i=1}^{m}\xi_i^2 - m\mu^2\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\xi_i^2 - m\mu^2
\end{aligned}$$

So, in other words, we have shown that

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\xi_i^2 - m\mu^2$$

which, when rearranged, yields the desired result.

(b) Find $p_{X_i}(\xi_k)$, the marginal PMF (probability mass function) of X_i. Use this to show that X_i is an unbiased estimator for μ.

Solution: Since we are told to assume each subset of size n from S is equally likely, we know that we are using the equally-likely probability measure. This in turn implies that $X_i \sim \text{DiscUnif}(S)$, meaning

$$\boxed{p_{X_i}(\xi_k) = \frac{1}{m}, \quad \forall \xi_k \in S}$$

This allows us to conclude

$$\mathbb{E}[X_i] := \sum_{k=1}^{m} \xi_k p_{X_i}(\xi_k) = \sum_{k=1}^{m} \xi_i \cdot \frac{1}{m} = \frac{1}{m} \sum_{k=1}^{m} \xi_k = \boxed{\mu}$$

(c) Find $p_{X_i,X_j}(\xi_k, \xi_\ell)$, the joint PMF of (X_i, X_j) where $i \neq j$. Be sure to account for the cases $k = \ell$ and $k \neq \ell$ separately. **Hint:** Consider first calculating the conditional probability $\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k)$.

Solution: Let's follow the hint, and first consider calculating $\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k)$. First note that if $k = \ell$ this probability is zero: if $X_i = \xi_k$ and we are sampling without replacement, it is impossible for X_j to then equal ξ_ℓ. If $k \neq \ell$, then after assigning ξ_k to X_i there are a total of $(m-1)$ elements in S remaining, each of which are equally likely to be assigned to X_j. Hence, $\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k) = 1/(m-1)$ in this case, and so

$$\mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k) = \begin{cases} 1/(m-1) & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell \end{cases}$$

Finally, we use the multiplication rule:

$$\mathbb{P}(X_i = \xi_k,\ X_j = \xi_\ell) = \mathbb{P}(X_j = \xi_\ell \mid X_i = \xi_k) \cdot \mathbb{P}(X_i = \xi_k)$$

which, using our result of part (b) above, means the joint PMF of (X_i, X_j) is given by

$$\boxed{p_{X_i,X_j}(\xi_k, \xi_\ell) = \begin{cases} \frac{1}{m(m-1)} & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell \end{cases} = \left(\frac{1}{m(m-1)}\right) \mathbb{1}_{\{k \neq \ell\}}}$$

For parts (d) - (f): Since our sample was taken *without* replacement, the X_i's will *not* be independent. The question we will work toward answering over the next few parts is: what is the covariance between any two observations, X_i and X_j?

(d) Use the result of part (c) to show that

$$\mathbb{E}[X_i X_j] = \frac{1}{m(m-1)} \left[T^2 - \sum_{k=1}^{m} \xi_k^2 \right]$$

Hint: If you encounter a double sum over indices k and ℓ such that $k \neq \ell$, note that you can express this double sum as: a double sum over *all* (k, ℓ) minus a double sum over which $k = \ell$.

Solution: A formula from PSTAT 120A that we should remember is

$$\mathbb{E}[g(X_i, X_j)] = \sum_k \sum_\ell g(\xi_k, \xi_\ell) p_{X_i, X_j}(\xi_k, \xi_\ell)$$

Hence, taking $g(X_i, X_j) = X_i X_j$, we have

$$\begin{aligned}
\mathbb{E}[X_i X_j] &= \sum_{k=1}^{m} \sum_{\ell=1}^{m} \xi_k \xi_\ell p_{X_i, X_j}(\xi_k, \xi_\ell) \\
&= \sum_{k=1}^{m} \sum_{\ell=1}^{m} \xi_k \xi_\ell \left(\frac{1}{m(m-1)} \right) \mathbb{1}_{\{k \neq \ell\}} \\
&= \frac{1}{m(m-1)} \sum_{k \neq \ell} \xi_k \xi_\ell
\end{aligned}$$

where $\sum_{k \neq \ell}$ is a shorthand for a double sum over k and ℓ where $k \neq \ell$. At this point, let's follow the hint: we can express this double sum as the difference between a double sum over all k and ℓ minus a double sum where $k = \ell$:

$$\begin{aligned}
\mathbb{E}[X_i X_j] &= \frac{1}{m(m-1)} \sum_{k \neq \ell} \xi_k \xi_\ell \\
&= \frac{1}{m(m-1)} \left[\sum_{k=1}^{m} \sum_{\ell=1}^{m} \xi_k \xi_\ell - \sum_{k=\ell} \xi_k \xi_\ell \right] \\
&= \frac{1}{m(m-1)} \left[\left(\sum_{k=1}^{m} \xi_k \right) \left(\sum_{\ell=1}^{m} \xi_\ell \right) - \sum_{k=1}^{m} \xi_k^2 \right] \\
&= \frac{1}{m(m-1)} \left[T^2 - \sum_{k=1}^{m} \xi_k^2 \right]
\end{aligned}$$

(e) Show that

$$\frac{1}{m(m-1)} \left[T^2 - \sum_{k=1}^{m} \xi_k^2 \right] = -\frac{\sigma^2}{m-1} + \mu^2$$

Solution: We now apply the result of part (a):

$$\begin{aligned}
\mathbb{E}[X_i X_j] &= \frac{1}{m(m-1)} \left[T^2 - \sum_{k=1}^{m} \xi_k^2 \right] \\
&= \frac{1}{m(m-1)} \left[T^2 - m(\sigma^2 + \mu^2) \right]
\end{aligned}$$

Now, note that $T := \sum_{i=1}^m \xi_i = m[(1/m)\sum_{i=1}^m \xi_i] = m\mu$. Hence:

$$\begin{aligned}
\mathbb{E}[X_i X_j] &= \frac{1}{m(m-1)}\left[T^2 - m(\sigma^2+\mu^2)\right] \\
&= \frac{1}{m(m-1)}\left(m^2\mu^2 - m\sigma^2 - m\mu^2\right) \\
&= \frac{1}{m(m-1)}\left[(m^2-m)\mu^2 - m\sigma^2\right] \\
&= -\frac{\sigma^2}{m-1} + \mu^2
\end{aligned}$$

(f) Combine the results of previous parts to conclude that

$$\text{Cov}(X_i, X_j) = -\frac{\sigma^2}{m-1}$$

Is it true that as the population size grows, the correlation between X_i and X_j necessarily drops to zero? **Hint:** does σ^2 depend on m?

As an Aside: this question is a preview of the kinds of questions asked in the branch of Statistics known as **asymptotics**, which is primarily concerned with the long-term behavior of statistical quantities.

Solution:

$$\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_i] = -\frac{\sigma^2}{m-1} + \mu^2 - (\mu)(\mu) = -\frac{\sigma^2}{m-1}$$

Now, it is tempting to say: as $m \to \infty$, the above expression goes to zero. However, we need to be a bit careful: this would be true if σ^2 were independent of m, however σ^2 *does* implicitly depend on m.

Think of it this way: suppose $S = \{1, 2, 3\}$, and consider letting $m \to \infty$. Letting $m \to \infty$ amounts to adding additional elements to S. If we add the elements $\{10^2,\ 10^4,\ 10^6, \cdots\}$, we see that σ^2 will most certainly increase with m.

So, in order for $\text{Cov}(X_i, X_j)$ to tend to zero as $m \to \infty$, we need σ^2 to grow at a rate *slower* than m. (In words, this is saying we need to ensure that as we add more elements to S, we are doing so in such a way that the population variance doesn't blow up too quickly.) Mathematically, you might see this expressed as: $m \to \infty$ while $\sigma^2/m \to 0$.

Coding Portion

Problem 1: Industrial Wastewater Discharge

Goals

This problem will touch on the following topics:

- Exploratory Data Analysis (EDA)
- Conducting hypothesis tests (including ANOVA)
- Regression

I encourage you to think of this as a mini-project, with some additional guidance beyond that which would be provided on the final project for this course.

Important

All plots must be generated using `ggplot`; plots generated in Base `R` will not receive full marks.

In this problem, we'll continue our example of Industrial Wastewater Discharge (IWD) from Lecture 12. As a reminder: "Industrial wastewater discharge" refers to liquid waste produced as a result of industrial processes. Companies are often required to register for permits to produce industrial wastewater, and these permits are sometimes contingent on the average concentration of pollutants in the wastewater.

We'll consider the IWD from a fictitious company, called Company *X*, located in the fictitious country of *Gauchonia*. Suppose that the state of Gauchonia only offers permits to companies whose wastewater has an average pollutant concentration of 140 mg/L or less; to ensure regular compliance, government officials take annual audits of Company *X*'s IWD.

Audits of Company *X*'s IWD are conducted by taking water samples from a nearby river and recording pollutant levels (in mg/L). Because pollutant levels may differ across locations in this river (due to environmental factors), auditors take multiple water samples from 6 different locations in the river, marked *A* through *F*. Note that the number of samples taken from each location in a given audit are not the same; different locations may have different numbers of associated water samples. (In statistical terms, we call this an **unbalanced design**.)

The dataset `iwd.csv`, located in the `data` subfolder, contains the results of these audits dating back for the past 15 years. Specifically, it contains the following variables:

- `Year`: year of audit; ranges from 2010 to 2025.
- `Month`: month of audit.
- `Day`: day of audit
- `Location`: location marker, indicating the location of the sample; one of `A`, `B`, `C`, `D`, `E`, or `F`
- `Sample_No`: unique identifier for the sample
- `Pollutants`: concentration of pollutants in the given sample, measured in mg/L.

The first few rows of the `iwd` dataframe are displayed below:

Location	Year	Readings
A	2010	137.8

Location	Year	Readings
A	2010	136.9
A	2010	140
A	2010	137.3

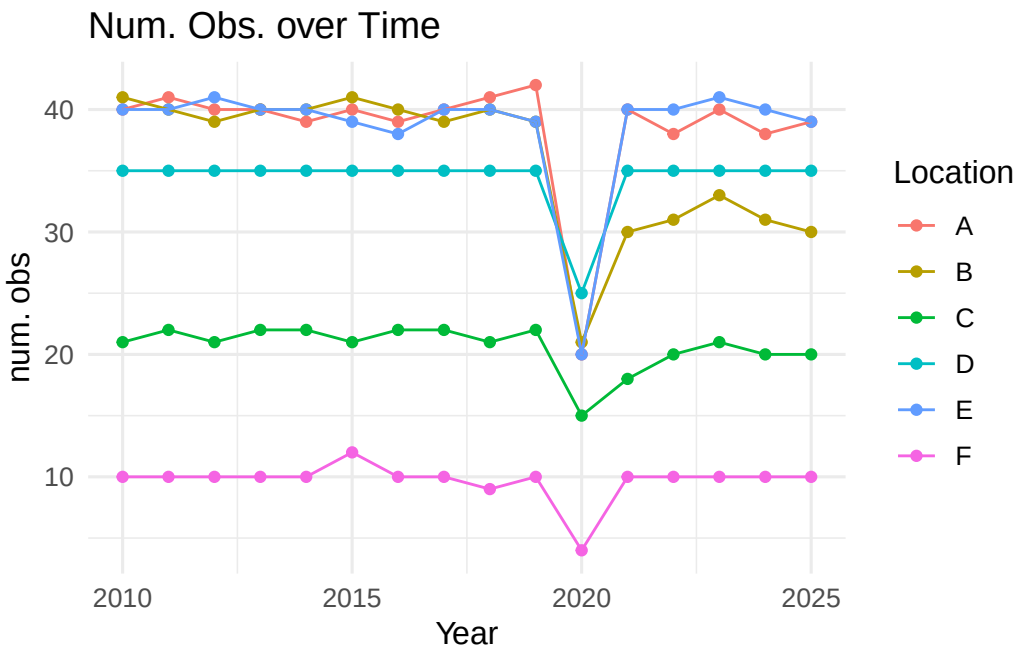
Part I: Exploratory Data Analysis

a) Generate a lineplot that tracks the number of observations taken at each location over time. Color your plot based on location, and select an appropriate color palette. Describe any patterns/trends you see, and provide a verbal explanation for why you think these trends may exist.

SOLUTIONS:

```
iwd %>% group_by(Location, Year) %>%
  summarise(n_obs = n()) %>%
  ggplot(aes(x = Year, y = n_obs)) +
  geom_point(aes(col = Location)) +
  geom_line(aes(col = Location,
                group = Location)) +
  ylab("num. obs") +
  theme_minimal(base_size = 12) +
  ggtitle("Num. Obs. over Time")
```

```
`summarise()` has grouped output by 'Location'. You can override using the
`.groups` argument.
```



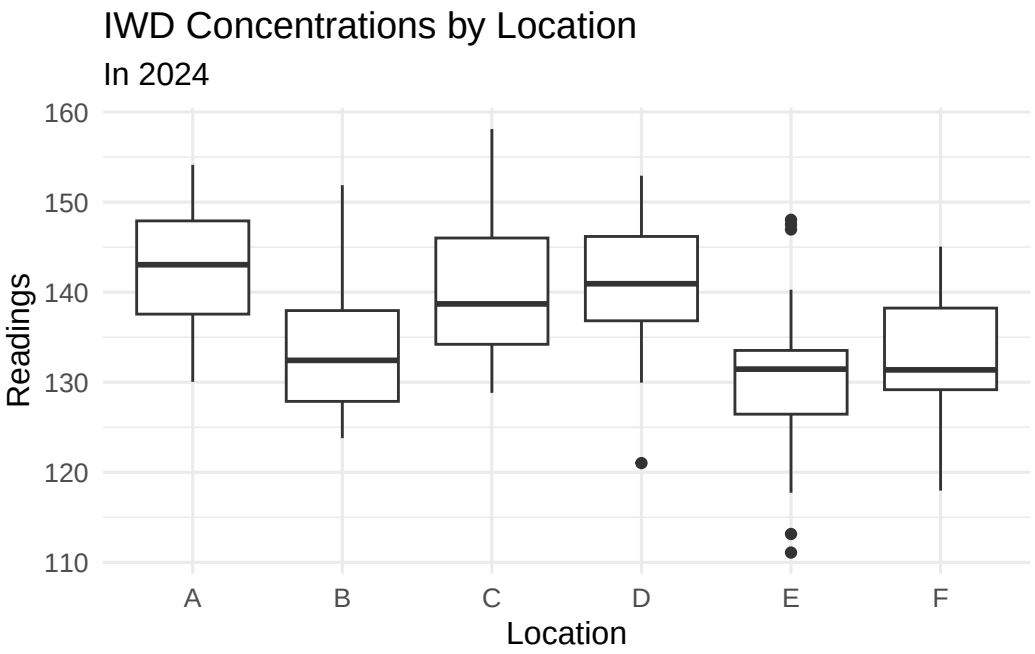
The number of observations appears to have stayed relatively constant over time, with a noticable dip in 2020. This dip is likely due to the COVID-19 pandemic.

b) Generate an appropriate plot (it's up to you to figure out which one!) that displays the distribution of pollutant measurements at each location; include only observations

from the year 2024. Based on this plot alone, do you think there are significant differences between the average IWD concentration at the different locations?

SOLUTIONS:

```
iwd %>% filter(Year == 2024) %>%
  ggplot(aes(x = Location, y = Readings)) +
  geom_boxplot() +
  theme_minimal(base_size = 12) +
  ggtitle("IWD Concentrations by Location",
          subtitle = "In 2024")
```



In 2024, there does appear to be some difference in the pollutant concentrations across locations. Specifically, locations E and F appear to have lower average IWD concentrations, and Location A appears to have the highest average IWD concentration.

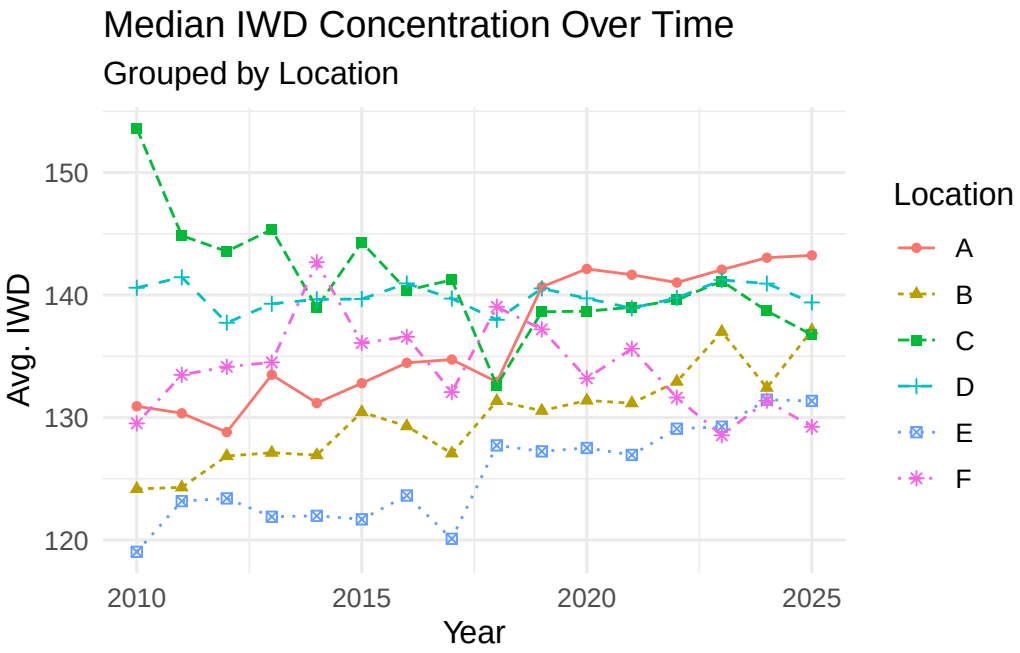
c) Generate a lineplot that tracks the average (based on the plot from part (b), it's up to you to decide whether the mean or the median is better) pollutant concentration at each location, over time. Color your plot based on location, and select an appropriate color panel. Describe any patterns/trends you see, and provide a verbal explanation for why you think these trends may exist.

SOLUTIONS:

```
iwd %>%
  group_by(Year, Location) %>%
  summarise(`Avg. IWD` = median(Readings)) %>%
  ggplot(aes(x = Year, y = `Avg. IWD`)) +
  geom_point(aes(colour = Location, shape = Location)) +
  geom_line(aes(colour = Location, linetype = Location)) +
  theme_minimal(base_size = 12) +
  ggtitle("Median IWD Concentration Over Time",
```

```
            subtitle = "Grouped by Location")
```

```
`summarise()` has grouped output by 'Year'. You can override using the
`.groups` argument.
```



Overall, there doesn't appear to be a trend in median pollutant readings over time. Locations A and E, howver, do appear to have experienced a steady increase in average IWD over time.

Part II: Statistical Tests

In this part, we imagine that the *Gauchonian* government is considering whether to renew Company *X*'s permits for 2026, based on their IWD in 2025.

d) For each location, test the null that the mean pollutant concentration falls below 140 mg/L against the upper-tailed alternative that the mean concentration exceeds 140 mg/L. Use an overall 5% level of significance, **but be sure to also implement appropriate safeguards to control for multiple hypothesis testing.** Report the conclusions of these six tests, and use this to determine whether you believe Company *X*'s IWD permits should be revoked in 2026 or not.

SOLUTIONS:

```
data_25 <- iwd %>% filter(Year == 2025)

filter_and_test <- Vectorize(function(loc) {
  loc_spec <- data_25 %>% filter(Location == loc) %>% pull(Readings)
  test_stat <- (mean(loc_spec) - 140) / (sd(loc_spec) / sqrt(length(loc_spec)))
  p_val <- 1 - pt(test_stat, df = length(loc_spec) - 1)
  return(p_val)
})
```

```
data.frame(`p val` = filter_and_test(LETTERS[1:6]) %>% round(6),
           check.names = FALSE
) %>% t() %>% pander()
```

	A	B	C	D	E	F
p val	0.000121	0.9991	0.9169	0.3763	1	1

To safeguard agasint multiple testing, we can simply divide our overall significance level by 6 (the number of tests we are conducting), and compare each of the above p-values to 0.0083. We see that at all locations except for Location A, we can safely fail to reject the null in favor of the alternative. As such, there was insufficient evidence to suggest that Company X's IWD falls below the threshhold, indicating that their license should not be revoked.

e) Let's suppose we want to statistically test whether we believe the mean pollutant concentrations in 2025 were the same across locations. Conduct an ANOVA to test this; use an overall 5% level of significance, but again **implement corrections for multiple testing.** Then, run a Kruskal-Wallis test and compare your results to that obtained by the ANOVA.

> **Information**
>
> The Kruskal-Wallis (KW) test is an example of what is known as a **nonparametric** test. Specifically, recall that one of the assumptions of ANOVA is that observations within each group are normally distributed. The KW test tests the same null and alternative hypotheses as the ANOVA, but does not make any distributional assumptions on the observations. Such "distribution-free" tests are called *nonparametric*.

SOLUTIONS:

```
aov(Readings ~ Location, data_25) %>% summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Location      5   4119   823.7   24.07 <2e-16 ***
Residuals   167   5716    34.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To safeguard agasint multiple testing, we should again divide our significance level by 6. Regardless, we see that the associated p-value in ANOVA is practically zero, meaning there is statistical evidence to suggest the average concentrations is not the same across locations.

```
kruskal.test(Readings ~ Location, data_25)
```

```
    Kruskal-Wallis rank sum test

data:  Readings by Location
Kruskal-Wallis chi-squared = 72.576, df = 5, p-value = 2.979e-14
```

Even the Kruskal-Wallis test agrees.