**Part I**

Exercise 1:

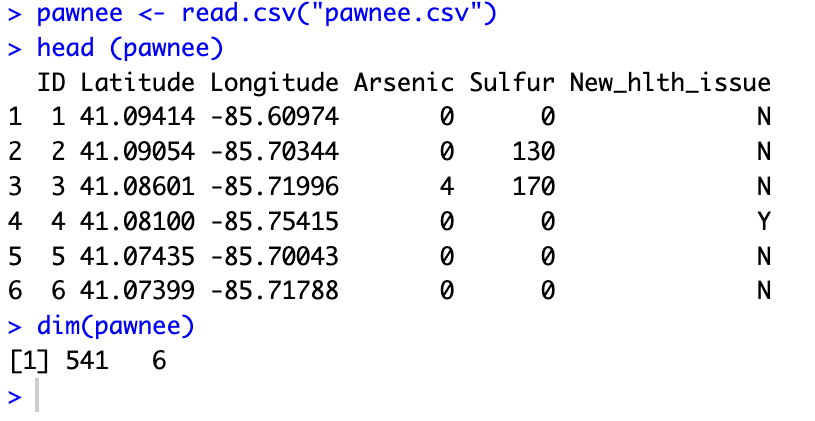
a. Using the head() and dim() functions to print out data.

**Command:**

**A close-up of a computer screen

Description automatically generated**

**Output:**

****

b. Setting the seed to 1337 and taking a simple random sample of size 30 from the data set and saving it as a separate R object.

**Command:**

A screenshot of a computer program

Description automatically generated

**Output:**

A screenshot of a computer code

Description automatically generated

c. Reporting the proportion of households experiencing a major health issue from sample and the population proportion of all households which experienced a new major health issue.

**Command:**

**A close up of words

Description automatically generated**

**Output:   
A screenshot of a phone

Description automatically generated**

d. Generating confidence intervals for sample proportion using the sample results and producing 90%, 95%, and 99% confidence intervals for the true population proportion.

**Command:**

A screenshot of a computer program

Description automatically generated

**Output:**

A screenshot of a computer program

Description automatically generated

Exercise 2:

a. Determining the null and alternative hypotheses and whether it’s a one-sided or two-sided test.

H0: The proportion of dangerous lead levels in Flint is less than or equal to 10%

Ha: The proportion of dangerous lead levels in Flint is greater than 10%

🡪 This is a one-sided test because we’re only interested in the case where the proportion is greater than 10%.

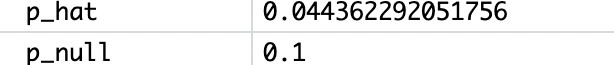
b. Calculating the sample proportion and sample standard deviation of the sample proportion of dangerous lead levels.

**Command:**

A computer code with black text

Description automatically generated with medium confidence

**Output:**





c. Calculating the SE of sample proportions, and the z-value for this test.

**Command:**

A black text on a white background

Description automatically generated

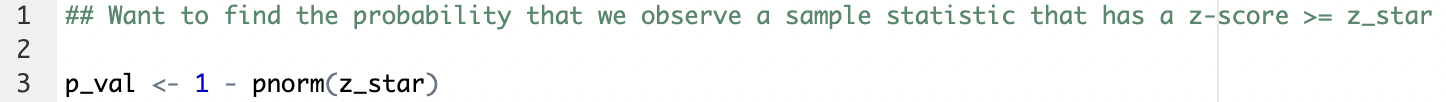
**Output:**

****

****

d. Calculating the p-value associated with this test using the z-statistic in (c).

**Command**



**Output:**

A screenshot of a cell phone

Description automatically generated

e. Using a significance level of 0.05, do you reject the null hypothesis?

**Command:**

A close up of a text

Description automatically generated

**Output:**

**A close-up of a sign

Description automatically generated**

🡪 Our p-value is larger than 0.05 so we fail to reject the null hypothesis.

f. Determining whether to be told to the EPA since remediation action is required to be taken by the EPA if greater than 10% of households in Flint contain dangerous lead levels.

🡪 Since we failed to reject the null hypothesis, we have no evidence to suggest that more than 10% of homes in Flint have a dangerous lead level. Therefore, we would tell the EPA that they are not required to take remediation action.

**Part II**

Exercise 1:

Research done in 2013 found that only 48% of all the site users reported getting their news about world events on this site. A 2018 poll of 3625 randomly selected users of a social media site found that 1830 get most of their news about world events on the site.

a. Does this sample give evidence that the proportion of site users who get their world news on this site has changed since 2013? Carry out a hypothesis test and use a 0.05 significance  
level.

Hypothesis Test

*1) Hypothesizing:*

🡪 Null Hypothesis **(H₀):** **p = p₀ =** **0.48,** the proportion of site users who get their world news on the site has not changed since 2013.

🡪 Alternative Hypothesis **(Hₐ):** **p ≠ p₀,** the proportion of site users who get their world news on the site has changed since 2013.

*2) Collecting and summarizing data:*

🡪 The conditions for using the z-test are **satisfied** since:

1. The sample is randomly selected

2. n x p₀ = 3625 x 0.48 = 1740 > 10 & n(1 - p₀) = 1885 > 10

3. Population N is large as N ≥ 10n = 10 x 3625 = 36250

🡪 We calculate the sample proportion (p̂) as 1830/3625 ≈ **0.5047**,

the standard error (SE) as √(p₀ (1-p₀)/n) ≈ **0.0083**, and

the z-test statistic as (p̂ - p₀) / SE ≈ **2.99**

*3) Assessing evidence:*

We calculated a z-score of approximately 2.99.

For the two-tailed test (Hₐ: p ≠ p₀), we assess the p-value = P(Z ≤ -|z|) + P(Z ≥ |z|)

= 2(1 – 0.9986) = 0.0028

The p-value was calculated to be approximately **0.0028.**

4*) Concluding and interpreting:*

We **reject the null hypothesis** (H₀) since the p-value ≈ 0.0028 is less than the significance level (α) of 0.05. There is statistical evidence to suggest that the proportion of site users who get their world news on the site has changed since 2013.

b. After conducting the hypothesis test, a further question one might ask is what proportion of  
all of the site users get most of their news about world events on the site in 2018. Use the  
sample data to construct a 95% confidence interval for the population proportion. How does  
your confidence interval support your hypothesis test conclusion

🡪 *Calculating 95% confidence interval:*

p̂ ± z\* x SE = 0.5047 ± 1.96 x 0.0083 = 0.5047 ± 0.016268 = **(0.489, 0.521)**

*🡪 Interpretation:*

The confidence interval provides a range of plausible values for the population proportion based on the sample data. Since this confidence interval does not include the 2013 proportion of 0.48, it suggests that there has been a significant change in the proportion of site users getting their news from the site since 2013 with the true population proportion being higher in 2018. Further, the interval suggests that if we were to take many samples of the same size from the population, about 95% of those samples would result in a sample proportion between 0.489 and 0.521, further indicating an increase from the 2013 proportion. In summary, the confidence interval does not contain the 2013 value of 0.48 and lies entirely above it, which is **consistent with the hypothesis test** **result indicating a significant increase from the 2013 proportion.**

Exercise 2:   
According to the Brookings Institution, 50% of eligible 18- to 29-year-old voters voted in the  
2016 election. Suppose we were interested in whether the proportion of voters in this age group who voted in the 2018 election was higher. Describe the two types of errors one might make in conducting this hypothesis test.

🡪 *Type I Error:*

This occurs when the null hypothesis (H₀) is true, but we incorrectly reject it. In the context of the hypothesis test concerning the proportion of 18- to 29-year-old voters in the 2018 election, a Type I error would happen if the true proportion of young voters who voted was actually 50% (or less), indicating no increase from the 2016 election, but our hypothesis test incorrectly concluded that there was an increase.

*🡪 Type II Error:*

This error occurs when the null hypothesis is false, but we fail to reject it. In the context of our hypothesis test, a Type II error would occur if the true proportion of young voters who voted in the 2018 election was actually higher than 50%, but our hypothesis test failed to detect this increase and incorrectly concluded that there was no increase from the 2016 election.

In hypothesis testing, we always try to balance these errors. However, it's important to note that reducing the probability of making one type of error typically increases the probability of making the other type. For instance, setting a lower significance level (like 0.01 instead of 0.05) will reduce the likelihood of a Type I error but increase the likelihood of a Type II error.