Exercise 3 - Hypothesis Testing

Karlo Angelo C. Lazaro

2023-06-08

Problem

Refer to the Christmas spending dataset (see christmas_spending.csv in the class drive), which is fictional data on the total amount a random sample of customers spent in a department store during the Christmas season. It has been claimed that the customers of the department store spend a total of 5000 pesos on the average during the Christmas season. Test the claim at 0.05 level of significance.

Q1. What is your parameter of interest? What is your point estimator?

A1. The parameter of interest is the average spending of customers during Christmas season in the department store. The point estimator in this instance would be the sample mean ($\bar{x}$) computed from the provided data.

Q2. Load the data into R and compute the point estimate.

```r
# Compute for the point estimate x-bar
s_mean <- mean(cspend$spending)
# Print the computed value of x-bar
print(s_mean)
```

A2.

```
## [1] 4658.872
```

The value of the sample mean ($\bar{x}$) is 4658.87 pesos.

Q3. State the problem as a two-sided hypothesis test. Give the null hypothesis and the alternative hypothesis.

A3. Here is the problem as a two-sided hypothesis:

H_0: The average Christmas spend of department store consumers is 5000 pesos.

H_a: The average Christmas spend of department store consumers is not 5000 pesos.

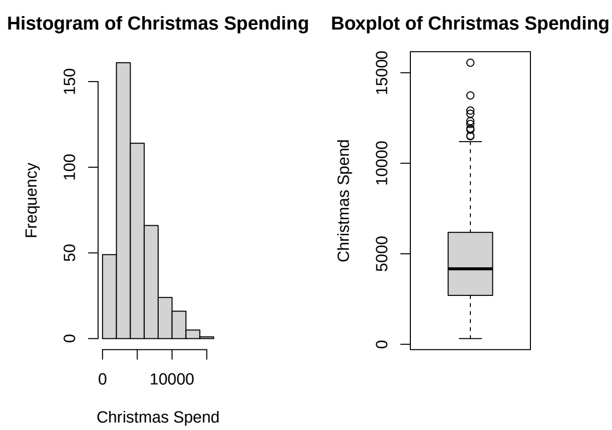
Q4. Why is the one-sample t-test appropriate for this problem?

A4. Looking at the requirements for a one-sample t-test, there are four main assumptions to be met.
1. The data must be continuous.
2. It should be independent.
3. The distribution is approximately normal.
4. Homogeneity of the variance.

Now, looking at our given data, for (1) we have spending in pesos which is continuous. (2), one persons spending is independent of another person's so that also meets the criteria. (3), based on the histogram, our data appears to be skewed to the right. And (4), our boxplot is showing no egregious outliers.

Although our data does not meet the normality assumption, we can still go ahead with using one sample t-test as our sample size is robust enough and the violation is not that extreme. Also, the t-test is concerned with approximate normality and not strict normality. Real world data usually shows some degree of deviation from normality but as long as these deviations are acceptable and not severe then we can go ahead and use the one sample t-test.

```
# Set up the grid layout and adjust the margin
par(mfrow = c(1, 2), mar = c(4, 5, 3, 5))
# Visual test of normality of the data
hist(cspend$spending, main = "Histogram of Christmas Spending", xlab = "Christmas Spend")
# Visual test of data variability
boxplot(cspend$spending, main = "Boxplot of Christmas Spending", ylab = "Christmas Spend")
```



Q5. Compute the observed value of your test statistic. The sample variance may be computed using the var() function in R.

A5. To compute for t we will use the following formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where,
$\bar{x}$ is the sample mean.
μ_0 is the population mean.
s is the standard deviation, and
n is the sample size.

And right now we have the following values computed/given:
1. $\mu_0 = 5000$
2. $\bar{x} = 4658.872$
3. $n = 436$

```r
# Given data from the problem
p_mean <- 5000
# Compute sample size
s_size <- NROW(cspend$spending)
```

To get the value of s, we use the following:

```r
# Compute the sample variance
s_var <- var(cspend$spending)
# Compute the standard deviation
s_dev <- sqrt(s_var)
# Print the computed standard deviation
print(s_dev)
```

```
## [1] 2581.068
```

Now that we have all the variables we need to compute the one sample t-test, we can just plug them in to compute for our t statistic.

```r
# Compute for the t-statistic
t <- (s_mean - p_mean) / (s_dev / sqrt(s_size))
# Print the computed t-statistic
print(t)
```

```
## [1] -2.759695
```

Q6. What is the null distribution? Don't forget to give the actual value of the parameter.

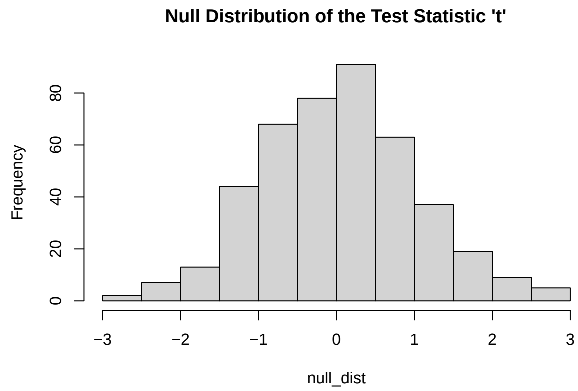
A6. The null distribution of our test statistic t follows a t-distribution with n-1 degrees of freedom.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

So, in this case, the null distribution follows a t-distribution with 435 degrees of freedom (t_{435}).

To visualize it further:

```r
# Set seed so that the computed null distribution will be fixed
set.seed(31)
# Set the degrees of freedom
deg_free <- 435
# Generate the null distribution
null_dist <- rt(s_size, deg_free)
# Plot the null distribution
hist(null_dist, main = "Null Distribution of the Test Statistic 't'")
```



Q7. Find the p-value corresponding to the observed value of your test statistic.

A7. For a two-sided t-test, the p-value is given by $p = P(|T| > t)$, which is equal to the R code 2*(1-pt(abs(t), n-1)).

Since we have all of these variables computed from previous questions, we can calculate the p-value by plugging in the different values we have.

```
# Plugging in the values
p_val <- 2*(1-pt(abs(t), deg_free))
# Print the computed p-value
print(p_val)
```

```
## [1] 0.006029875
```

Q8. Based on your answer in (7), what is your conclusion?

A8. As the p-value (0.0061) is less than our level of significance ($\alpha = 0.05$), then we reject the null hypothesis that Christmas spend of department store consumers is equal to 5000 pesos. There is sufficient evidence to suggest that the Christmas spend of department store consumers significantly differs from 5000 pesos.

Q9. Alternatively, what is the critical value corresponding to the significance level? You need to use the qt() function. Compare the critical value and the test statistic. What is your conclusion based on the critical value? Is your answer consistent with your answer in (8)?

A9. Using the critical value approach, we will get:

```
# Given alpha
alpha <- 0.05
# Compute for the critical value of a two-tailed test
crit_val <- qt(p=1-alpha/2, df=deg_free)
# Print computed critical value
print(crit_val)
```

```
## [1] 1.965432
```

Since the critical value |1.965432| < |-2.759695| t-value, we reject the null hypothesis that the average Christmas spend of department store consumers is equal to 5000 pesos. This indicates that there is sufficient evidence to suggest that the average Christmas spend of department store consumers significantly differs from 5000 pesos. As the t-value obtained is negative, it suggests that the average Christmas spend is significantly lower than 5000 pesos.

And yes, this is consistent with my answer in Q8.

Q10. Now perform the test again, but this time use the t.test() function in R. You should arrive at the same conclusion.

A10. Using the t.test() function, we get the following:

```r
# Run the t-test using the t.test() function
ttest <- t.test(cspend$spending, mu=p_mean, alternative = "two.sided")
# Print the computed t-test
print(ttest)
```

```
##
##  One Sample t-test
##
## data:  cspend$spending
## t = -2.7597, df = 435, p-value = 0.00603
## alternative hypothesis: true mean is not equal to 5000
## 95 percent confidence interval:
##  4415.924 4901.821
## sample estimates:
## mean of x
##  4658.872
```

The answers obtained in the t.test() function is consistent with what I got from the other questions.