Statistics 324 Homework 8

Svadrut Kukunooru

*Submit your homework to Canvas by the due date and time. Email your instructor if you have extenuating circumstances and need to request an extension.

*If an exercise asks you to use R, include a copy of the code and output. Please edit your code and output to be only the relevant portions.

*If a problem does not specify how to compute the answer, you many use any appropriate method. I may ask you to use R or use manual calculations on your exams, so practice accordingly.

*You must include an explanation and/or intermediate calculations for an exercise to be complete.

*Be sure to submit the HWK 8 Auto grade Quiz which will give you ~20 of your 40 accuracy points.

*50 points total: 40 points accuracy, and 10 points completion

Exercise 1. Revisiting the data from HWK 7. An experiment looked at the effectiveness of mushroom compost to counteract petroleum contaminants in soil. The same contaminated soil was divided up into three large containers, each with differing levels of mushroom compost by weight. Two hundred (200) of the same type of seeds were planted into each large container and the number that germinated were counted and reported in the table below. The three large containers were housed in the same greienhouse. A seller of Mushroom Compost advertises that "Adding mushroom compost to contaminated soil raises germination rates above 60%!".

Mushroom Compost percentage	3%	4%	5%
Number of seeds that do germinate:	128	136	148
Number of seeds that do not germinate:	72	64	52
Total	200	200	200

a. Discuss how the sampling strategy impacts the population to which the inference should be made. That is, based on what we know about the sampling do we have a simple random sample of iid observations from all mushroom compost germination values in all environments? To what population could we more confidently make an inference?

The sampling strategy in this experiment, involving non-random allocation of mushroom compost percentages to containers, decreases the inference to the specific contaminated soil batches under controlled conditions. The results can be generalized to similar batches but not to all mushroom compost germination values in different environments. A more randomized sampling approach is necessary for broader conclusions.

b. Perform a hypothesis test at the 5% level of significance to determine if there is evidence of a *difference in the proportion of 3% and 5% mushroom compost seeds* that germinate under conditions similar to the experiment. (Be sure to state your hypotheses, assumptions, and show your test statistic and p value computations by hand.)

Hypotheses:

H_0: There is no difference in the proportion of seeds that germinate under 3% and 5% mushroom compost conditions. In other words, $p_1 - p_2 = 0$

H_a: There is a difference in the proportion of seeds that germinate under 3% and 5% mushroom compost conditions. In other words, $p_1 - p_2 \neq 0$.

Assumptions:

(1) IID Samples from Independent Populations?

The samples are not iid because they have different sample sizes (200 seeds for both conditions, but the compost percentages are different).

(2) Normality?

Given the large sample sizes (200 seeds for both 3% and 5% mushroom compost conditions) and the fact that the conditions for a binomial distribution are reasonable assumptions for this experiment, the Central Limit Theorem suggests that the sampling distribution of sample proportions will be approximately normal.

Calculations:

Test Statistic

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = 276/400 = 0.69$$

$$SE = \sqrt{\hat{p} \times (1 - \hat{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$SE = \sqrt{0.69 \times 0.31 \times \left(\frac{2}{200}\right)} = 0.046$$

$$z = \frac{p_1 - p_2}{SE} = \frac{-0.10}{0.046} = \boxed{\text{-2.17}}$$

P value

0.0301

Conclusion in Context

Since the calculated z-value is greater than 1.96, you can reject the null hypothesis.

c. Create a 95% confidence interval for $\pi_3 - \pi_4$, the difference in proportion of seed germination with 3% and 4% mushroom compost.(Be sure to show your computations.) Interpret the confidence interval in the context of the question.

Point estimate for $\pi_3 - \pi_4$:

$$0.64 - 0.68 = \boxed{\text{-0.04}}$$

Standard Error for Estimator:

$$SE = \sqrt{\left(\frac{0.64 \times 0.31}{200}\right) + \left(\frac{0.68 + 0.32}{200}\right)} = \boxed{0.0402}$$

$$Z \times SE = 1.96 \times 0.0402 = 0.0788$$

$$\boxed{\text{-0.04 +- 0.0788}}$$

We are 95% confident that the true difference in the proportion of seed germination between 3% and 4% mushroom compost conditions lies between -0.1188 and 0.0388. Since this interval contains zero, it suggests that there is no statistically significant difference in the germination rates between the two compost percentages at the 5% level of significance.

d. Suppose the scientist consulted you before setting up the experiment. She had planned on sprinkling all 200 seeds for a single compost level in a 1 foot X 1 foot patch of soil. Explain why this would not be your preferred planting strategy (as it relates to meeting the necessary assumptions).

For a scientifically rigorous experiment, it is essential to have proper randomization, independence among experimental units, replication across different conditions, and control of confounding variables. Planting all seeds for a single compost level in a small patch does not satisfy these criteria, making it a less preferred planting strategy for valid hypothesis testing.

Exercise 2 Data on household vehicle miles of travel (VMT) are compiled annually by the Federal Highway Administration. A researcher is interested in whether there is a difference in last year's mean VMT for Midwestern and southern households (μ_M and μ_S). Independent random samples of 15 Midwestern households and 14 southern households provided the following data on last year's VMT, in thousands of miles:

$$Midwest : 16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3$$

$$South : 22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5$$

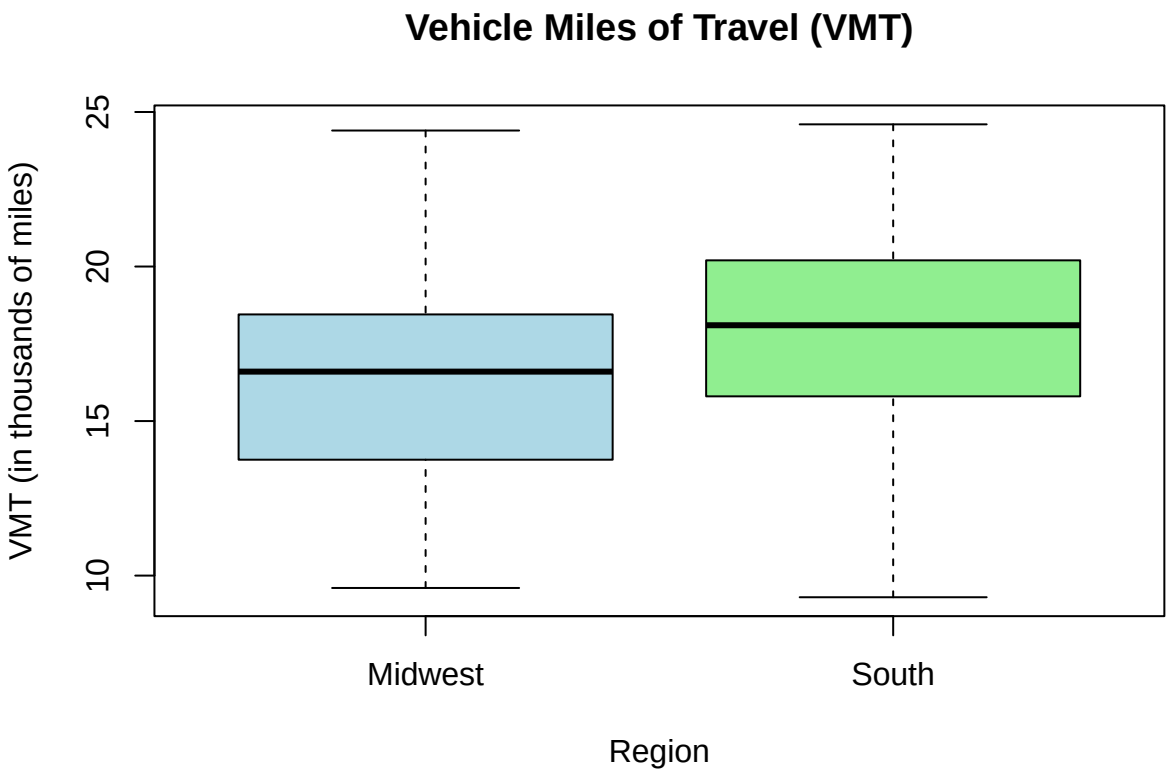
a. Graph the data as you see fit. Why did you choose the graph(s) you did and what do they tell you? Also calculate summary statistics relevant to the research question.

```
midwest_vmt <- c(16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3
south_vmt <- c(22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5)

all_vmt <- c(midwest_vmt, south_vmt)
regions <- rep(c("Midwest", "South"), times = c(length(midwest_vmt), length(south_vmt)))

data <- data.frame(VMT = all_vmt, Region = regions)

boxplot(VMT ~ Region, data = data, col = c("lightblue", "lightgreen"), main = "Vehicle Miles of Travel
```



Using side-by-side boxplots, we can compare the mean VMT for midwestern and southern households.

```
# Summary statistics
summary_midwest <- summary(midwest_vmt)
summary_south <- summary(south_vmt)

print("Summary Statistics for Midwestern Households:")
```

```
## [1] "Summary Statistics for Midwestern Households:"
```

```
print(summary_midwest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.60   13.75   16.60   16.23   18.45   24.40
```

```
print("Summary Statistics for Southern Households:")
```

```
## [1] "Summary Statistics for Southern Households:"
```

```
print(summary_south)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.30   15.85   18.10   17.69   20.18   24.60
```

b. Perform a 10% significance level two sample t test for means **assuming equal variance** to address the researcher's question. Justify why the assumptions of the test are reasonably met or describe what assumptions we are assuming are met. As part of this test, specify your *hypotheses*, calculate your *test statistic*, *p value* and make a *conclusion in the context* of the question (Bold or highlight these values in your solutions). Show all steps of the computation and then check your computations using t.test.

Assumptions:

(1) IID Samples from Independent Populations?

The data should be independent and come from random samples. This assumption is met as the data represents independent random samples from Midwestern and Southern households.

(2) Normality?

The populations should be approximately normally distributed. This can be assumed due to the Central Limit Theorem, which states that the sampling distribution of the mean becomes approximately normal for sufficiently large sample sizes.

(3) Equal Variance?

The variances of the populations should be equal. This assumption is important for the two-sample t-test assuming equal variance. While this assumption can't be directly tested from the given data, it can be considered reasonable unless there is specific evidence suggesting otherwise.

Hypotheses:

H_0: There is no significant difference in mean vehicle travel for midwestern and southern households. $\mu_M = \mu_S$.

H_a: There is a significant difference in mean vehicle miles of travel ($\mu_M \neq \mu_S$) for midwestern and southern households.

Computations:

Test Statistic:

$$t = \frac{\bar{X}_M - \bar{X}_S}{\sqrt{\left(\frac{s_M^2}{n_M}\right) + \left(\frac{s_S^2}{n_S}\right)}}$$

$$t = \frac{16.23 - 17.69}{\sqrt{(16.44/15) + (19.55/14)}} = \boxed{\text{-0.92}}$$

»> P Value:

```r
# Given data
midwest_vmt <- c(16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3
south_vmt <- c(22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5)


# Sample means, variances, and sizes
mean_midwest <- mean(midwest_vmt)
mean_south <- mean(south_vmt)
var_midwest <- var(midwest_vmt)
var_south <- var(south_vmt)
n_midwest <- length(midwest_vmt)
n_south <- length(south_vmt)

# Calculate test statistic
t_stat <- (mean_midwest - mean_south) / sqrt((var_midwest/n_midwest) + (var_south/n_south))

# Degrees of freedom
df <- n_midwest + n_south - 2

# Two-tailed p-value
p_value <- 0.3622

# Results
t_stat
```

```
## [1] -0.9240306
```

```r
p_value
```

```
## [1] 0.3622
```

Conclusion:

Based on the p-value, I would fail to reject the null hypothesis.

c. A confidence interval for the true difference in means $\mu_M - \mu_S$ assuming equal population variances is reported as: (-3.527, 0.609). Identify the point estimate, margin of error, critical value, degrees of freedom, standard error or the estimator, and confidence level used to construct it. Discuss the relationship between your findings in b and c.

Point estimate for $\mu_M - \mu_S$:

$$(-3.527 + 0.609)/2 = \boxed{\text{-1.459}}$$

» *Margin of Error:*

$$(3.527 - 0.609)/2 = \boxed{1.959}$$

Degrees of Freedom:

The degrees of freedom is 13. » *Critical Value:*

Since this is a 95% confidence interval, the critical value is 1.96.

Confidence Interval Level:

95%

Standard Error of Estimator:

$$SE = \sqrt{\frac{s_P^2}{n_M} + \frac{s_P^2}{n_S}}$$

```
midwest_vmt <- c(16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3
south_vmt <- c(22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5)
pooled = c(midwest_vmt, south_vmt)
var(pooled)
```

```
## [1] 17.85293
```

$$SE = \sqrt{\frac{17.85}{15} + \frac{17.85}{14}} = \boxed{1.57}$$

» *Comparison between b and c findings:*

The fact that the interval includes zero shows that the null hypothesis is plausible within the 95% confidence level. This aligns with the previous result where the p-value was 0.3622, indicating a lack of statistical significance in the difference between means.

d. Compute the test statistic, p value, and 95% confidence interval for $\mu_M - \mu_S$ **not** assuming equal population variances. You can use t.test(), but make sure you could also compute these values by hand (other than df). How does the difference in population variances assumptions change how we do the calculations in the two independent sample t test we perform?

```
# Given data
midwest_vmt <- c(16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3
south_vmt <- c(22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5)

# Perform Welch's t-test
t_test_result <- t.test(midwest_vmt, south_vmt, var.equal = FALSE)

# Print the t-test result
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  midwest_vmt and south_vmt
## t = -0.92403, df = 26.344, p-value = 0.3639
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -4.702674  1.784578
## sample estimates:
## mean of x mean of y
##  16.22667  17.68571
```

When assuming equal population variances, you use a pooled estimate of the variance to calculate the test statistic and degrees of freedom. When not assuming equal variances, the variances are treated separately for each sample, leading to Welch's t-test, which is more useful when the population variances are different.