GEOG 312 – DATA ANALYSIS IN GEOGRAPHY
ASSIGNMENT #3: COMPARING DISTRIBUTIONS

DATA

For this assignment, we will be using the measurements of surface temperature that were taken at different locations surrounding our lecture hall:
1) Grass of the Academic Plaza
2) Sidewalk around the Academic Plaza
3) Pavement of Lot 19

Using R, I have compiled all the properly-formatted tables that you submitted for Assignment #2 into three tables corresponding to the three surface types listed above. These CSV-formatted tables can be found on eCampus under Assignments → Assignment #3:
1) class_grass.csv
2) class_sidewalk.csv
3) class_lot19.csv

BACKGROUND

1. SAMPLE SIZE

As the size of the sample increases, the difference between the sample mean and the population mean decreases and the sampling error decreases. It is less likely that a sample mean will differ greatly from the population mean. The sampling error is also called the standard error of the mean (SE_x) and calculated as:

$$SE_x = \frac{s}{\sqrt{n}} \quad \text{(eq. 1)}$$

where s is the standard deviation of the sample and n is the sample size. As the sample size increases and the standard deviation of the population decreases the sampling error is reduced. The standard error calculation above is only appropriate where you have an infinitely large population. Where you have a finite population, you need to include a correction factor based on the size of the population (N), which we'll call the corrected standard error (same notation SE_x):

$$SE_x = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad \text{(eq. 2)}$$

When the population is relatively large compared to the sample size, the correction factor approaches 1. However, a small population or large sample will cause the correction factor to be less than 1 and the sampling error to decrease.

2. NULL HYPOTHESIS

In general, the null hypothesis (H_0) is a position that a researcher attempts to reject through a statistical test in a manner that proves significance. In this assignment, we are comparing the sample means between two groups. Our null hypothesis is thus:

- "The difference in the mean values between two groups is the same. Any differences between the two sample means are due to random and non-systematic errors.

If the difference between two means is significant, then we *reject the null hypothesis* and accept the alternative hypothesis (H_a). Significance levels commonly used are 90%, 95%, and 99%.

The two types of null hypothesis we will assume in this assignment:

1. The sample mean ($\bar{X}$) is the same as the population mean (μ).
2. The sample mean of group 1 ($\bar{X}_1$) is the same as the sample mean of group 2 ($\bar{X}_2$).

2. Z-STATISTIC: DIFFERENCE FROM POPULATION MEAN (#1 and #2)

We can look at our data in another way: does our sample mean differ significantly from the population mean. This can be calculated using the ***Z* test**:

$$Z = \frac{\bar{X} - \mu}{SE_x} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \qquad \text{(eq. 3)}$$

Keep in mind the difference between the different symbols (Greek letters typically represent population statistics and letters in the English alphabet represent sample statistics). Eq. 3 assumes an infinite population, but if we know our population size, the following equation is more appropriate:

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \qquad \text{(eq. 4)}$$

3. STUDENTS' *t*-TEST: DIFFERENCE FROM POPULATION MEAN (#3 and #4)

The *Z*-statistic is based on having a large sample size ($n \geq 30$). When working with a smaller sample size ($n < 30$), it is recommended you use the **Students' *t*-test**. You use the degrees of freedom ($df = n - 1$) as the sample size and the *t*-distribution for your calculations:

$$t = \frac{\bar{X} - \mu}{\left(s / \sqrt{n-1}\right)} \qquad \text{(eq. 5)}$$

4. TWO SAMPLE DIFFERENCE OF MEANS TEST

Above, the *Z* and Students' *t* tests were used to determine if a sample was statistically different from the mean. Similarly, we can determine if two samples are statistically different from each other.

_Z_-test:
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1 - \bar{X}_2}} \quad \text{(eq. 6)}$$

***t*-test:**
$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1 - \bar{X}_2}} \quad \text{(eq. 7)}$$

No, you are not mistaken. The right-hand-side of both equations is the exact same. What is different is the distribution that you use to find significance.

- For the *Z*-test, you use the normal distribution.
- For the *t*-test, you use the Students' *t*-distribution.

In general, the *t*-test becomes more appropriate as sample size decreases. Also, since the *Z*-test assumes normality, the *t*-test is more adaptable. In both tests, the calculation of the standard error ($s_{\bar{X}_1 - \bar{X}_2}$) depends on whether the variance from the populations in which the samples were drawn are similar or not. If assumed to be similar then the standard error is calculated as:

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{(eq. 8)}$$

If the variances are *not* assumed to be similar then the standard error is calculated as:

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{(eq. 9)}$$

DELIVERABLES

1. (20% of points) In the following table, use R to calculate the statistics for each surface. Assume the ~400 values measured for each surface by the class represents the population. Determine if your group's sample is *significantly different* from the mean at the different significant levels (Hint: If significant at 99%, then it will also be significant at 90% and 95%).

Surface	Statistics				Significant at this level?		
	$\bar{X}$	μ	*s*	*Z* (using eq. 4)	90%	95%	99%
Quad (Grass)							
Sidewalk							
Parking Lot							

2. (20% of points) Describe the results of the table in the above question in terms of the null hypothesis. Use R to create a three properly-labelled figures comparing your three samples to the three populations (e.g. histogram, boxplot, etc.). Include that figure here.

3. (15% of points) Using your group's first 15 observations from each surface, use the Students' *t*-test in R and determine if these samples are statistically similar to the population mean. Fill in the information of the following table. The values of *t* at the 90, 95 and 99% confidence level are 1.78, 2.15 and 2.95.

Surface	Statistics				Significant at this level?		
	$\bar{X}$	μ	*s*	*t* (using eq. 5)	90%	95%	99%
Quad (Grass)							
Sidewalk							
Parking Lot							

4. (5% of points) Describe the results of the table in the above question in terms of the null hypothesis.

5. (15% of points) Use the *Z*-test (not the Students' *t*-test) to determine if a statistically significant difference in surface temperature exists between the samples taken from the grass in the quad and the sidewalk surrounding the quad. Complete your work in R.

Surface	Statistics				Significant at this level?		
	$\bar{X}$	*s*	SE_x (using eq. 9)	*Z* (using eq. 6)	90%	95%	99%
Quad (Grass)							
Sidewalk							

6. (5% of points) Describe the results of the table in the above question in terms of the null hypothesis.

7. (15% of points) In R, use the Students' *t*-test (not the *Z*-test) to determine if a statistically significant difference in surface temperature exists between the samples taken from the grass in the quad and the parking lot.

Surface	Statistics				Significant at this level?		
	$\bar{X}$	s	SE_x (using eq. 9)	t (using eq. 7)	90%	95%	99%
Quad (Grass)							
Parking Lot							

8. (5% of points) Describe the results of the table in the above question in terms of the null hypothesis.

SUBMIT THE FOLLOWING ITEMS TO ECAMPUS (ONLY ONE PER GROUP)

1. Report:
   a. Format: Please submit as a PDF with the following file name: ######.pdf where "######" is your unique group 6-digit ID (same group as last assignment)
   b. Cover page with assignment title, unique 6-digit ID, and group members
   c. Deliverables above (please number questions)
   d. Include the R code you used in your analysis (please number questions)