Is there a difference in attendance between seminar classes and computer practical classes?

Introduction

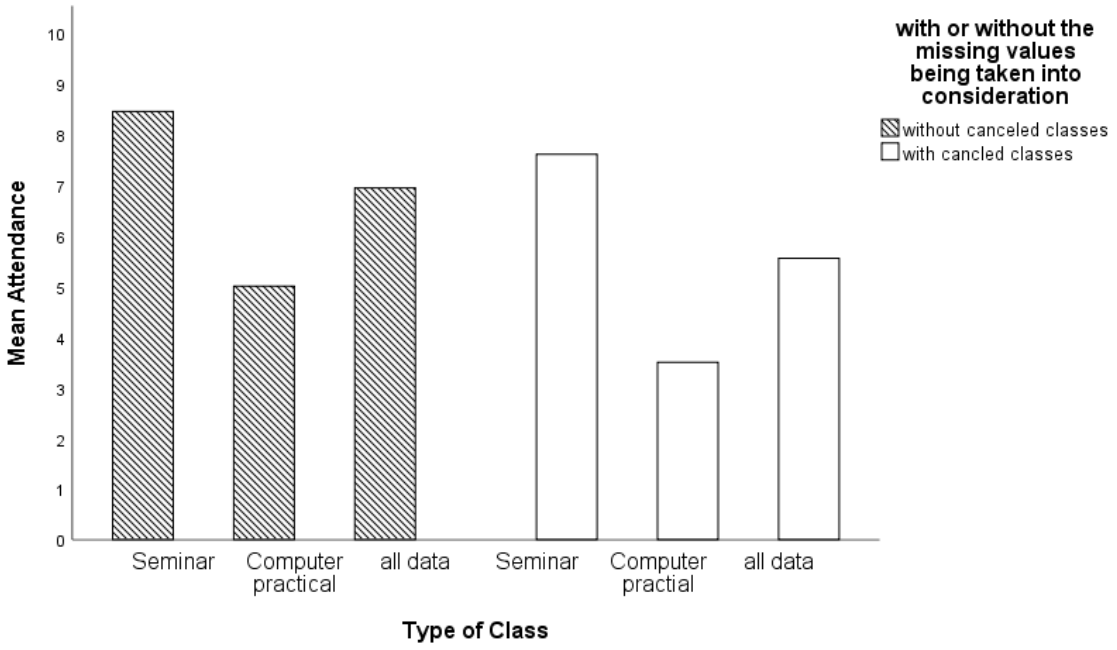
In this report I will be analysing the data provided and then by using the appropriate statistical test to evaluate if there is a difference between attendance based on class type. The report will start with some descriptive statistics and graphics about the data to try to understand it more by looking at the different aspects of it. Once an under standing has been gained a hypothesis shall be defined and then tested after deciding upon the appropriate test statistic.

Exploratory Data Analysis (EDA)

In the data that will be looked into for this report, the “Cancelled” values are when 2 or fewer people attended the class. For these missing values there is no data to tell as to why the attendance was so low. For example, the weather may have affected it, the lecturer on that day could have been really annoying. No one knows.

Under the style column there are 3 categories; Seminar, Computer practical and Combined. Since Seminar and Computer piratical classes are being analysed whenever analysis is taking place on something to do with the type of class the combined class data will be excluded.

When looking at the data it will be split 2 sets, one set that ignored the cancelled classes and another that sets them to 0. 0 was chosen because if the class was cancelled then there are 0 people sat in that room learning despite how many people may have tuned up for it. the data has been split up in this way because it’s a good balance between ignoring the data from the cancelled classes and analysing the data set for every possible combination of attendance for every class.



Firgure 1: Mean attendeace by class type and how the data is taken into consideration.

This bar chart shows the difference in mean attendance based upon if the cancelled classes are included as part of the calculation of the mean, or are they excluded assuming a cancelled classes attendance is 0. since no one was in the class despite people possibly arriving for it.

Figure 1 shows that on average the attendance for seminar classes is higher than computer practical classes. Then when including cancelled classes as 0 the divide between the means increases. Also considering the fact that more classes were cancelled for the computer piratical. You could conclude that computer practical classes are less popular than seminars. A statistical test should be done to formally analyse the data more accurately before making a conclusion.

Table 1. Desciptives looking at attendance for each type

		Style of Class			
		Seminar	Computer practical	Seminar with missing	Computer practial with missing
Attendance	Mean	8.444444	5.000000	7.60000	3.500000
	Lower Bound 5%	5.489540	3.150309	4.37957	1.414394
	Upper Bound 95%	11.39935	6.849691	10.8204	5.585606
	Std. Deviation	3.844188	2.000000	4.50185	2.9154759
	Variance	14.778	4.000	20.267	8.500

Table 1 excludes the combined class data and is similar to the bar chart except it gives more precise information about the data. Here you can see that there is a difference between the seminar and computer practical classes, for both with or without the cancelled classes being taken into consideration. Also, the variance is higher when the missing values are included.

Figure 2: Box plots showing the attendance distributed across different types of classes and different categorys based upon if the data falls into a sub group. With or Without are in relatoin to if the category included or excluded in the plot

Summary of EDA

In summary the data shows that there is a difference in the mean, and then further exaggerated when the 0s are included from the cancelled classes. The data also shows that the variance is higher when the 0s are included as part of the summary. Overall it does show grounds for there being a difference but that will have to be looked into in more detail.

Study design

Internal Validity

The study might have a problem with internal validity because the students may have been told that their attendance is being recorded to evaluate the classes popularity. This could cause the class to attended more out of pressure from the lecturer. This would impact the internal validity of the study.

The study doesn't mention the lecturer and the effect he may have on the classes, he may be really grumpy in the mornings etc. If this was true it could cause less students to show up for the class affecting the attendance and subsequently the internal validity.

The data doesn't show why the class was cancelled it could be because there was a snowstorm and class was cancelled for the student's safety. The lecturer could have been ill and couldn't find someone to teach the class on his behalf. These factors aren't reported and thus we don't know why the class was cancelled. It could even be that it did snow or there was a crash on major road grid-locking the commute and the class wasn't cancelled and thus only a few people where able to make it in. Furthermore because of the unlikely nature of these events if they had occurred the data should be excluded as to understand the effect it has on the attendance which would take far longer to collect the data for.

We don't know how many people are actually expected to show up for each class because it could be a subset of the 25 students that show up for a certain class and thus that why you have a lower attendance. It is not specifically told to us that all 25 students are required to show for every class. And also, we don't know if the attendance for certain classes counts towards a final grade as like an incentive to the students but only for a certain type of class.

Summary

we don't know about the external factors that would affect attendance, be that from lecturers telling the students they are being monitored and pressuring them to show up for more classes. Or an external factor that can't be controlled like weather, either way the internal validity of this study would not hold up. As there has been no discussion about how to prevent these things or record external factors.

External validity

If the study had good internal validity then that means that we can apply the result to a situation of the exact same type. But since we don't have information about the situation we don't know what else it could be applied to. For example, we don't know what the course is, or what time of year it was run. A far larger data set of this style across a far wider range of classes recorded over a longer period of time would be needed to get reliable external validity.

Similar to the point made about internal validity in how the test in run, it is not discussed as to what measures are taken place to ensure that the participants of the test are unaware that they are being studied this is referred to as 'psychological realism'. There would need to be particular steps made to ensure that no one knew that they where part of a study and the lecturer would have to behave in a certain way to prevent this.

There is also selection bias, since it is a new course there is more likely to be enthusiasm for it. To solve this, it would also need to be part of a larger study where multiple classes are looked into. By doing this it would remove the fact that it is new giving the study more external validity as it would be generalised and thus be able to be applied to more general situations.

Summary

The test has bad external validity because it is affected by the fact that the course is new, we don't know enough about external factors like the course type and the effect of psychological realism isn't considered/ prevented.

Test choice

A non-parametric test will be used because of the small sample size a normal distribution can't be confidently fit to data also it has been specifically instructed. Since two groups and subsequently the means of the groups are being compared, A non-parametric 2 sample test will be used.

Lots of statistical tests also require equal variance and thus an appropriate test should be used to evaluate if the data has equal enough variance as there was an observed difference in the EDA. Kendall's W test was chosen because its non-parametric and works for n samples thus making it a suitable choice.

Test of variance

Forming a hypothesis for this gives:

S_0: There is no difference between the variance across the 2 class types

S_1: There is a difference between the variance across the 2 class types

H_0: $\sigma^2_{\text{Seminar}} = \sigma^2_{\text{Comp prac}}$

H_1: $\sigma^2_{\text{Seminar}} \neq \sigma^2_{\text{Comp prac}}$

Table 2: Kendall's W test of variance

Kendall's W	7.200
Asymp. Sig.	.007

Table 2.1: Kendall's W test of variance with cancelled classes excluded

Kendall's W	1.000
Asymp. Sig.	.000

Table 2 shows us that for when the cancelled classes the data from that is omitted there is 0.7% chance that the variances are different by chance. This means that we have strong evidence to reject the null-hypothesis and state that the variance is not equal, thus we have to use a test that doesn't require equal variance. In table 2.1 there is 0% chance that the difference in variance is by chance when the cancelled classes attendance is taken into consideration and assumed to be 0.

Looking into what statistical test to use for the analysis of the difference between the groups. The Mann-Whitney test can't be used because it requires data that has no ties, subsequently data that is continuous. It also requires equal variance which was disproven by the test (Table 2 and 2.1). So, looking for a non-parametric test, that compares 2 independent groups, doesn't assume equal variance. the Kruskal-Wallis test will be used despite it requiring equal variance because it is more lenient to varying variances unlike the ANOVA test. As when it has data with varying variances it compares the mean ranks it also does well under small sample sizes in comparison to the ANOVA test.

Test of research question

Forming a hypothesis for this gives:

S_0: There is no difference between the attendance across the 2 class types

S_1: There is a difference between the attendance across the 2 class types

H_0: $\mu_{Seminar} = \mu_{Comp\ prac}$

H_1: $\mu_{Seminar} \neq \mu_{Comp\ prac}$

table 3: Kruskal-Wallis H ranked test of attendance by class type including canceled classes as 0

	Attendance
Kruskal-Wallis H	4.236
Asymp. Sig.	.040

table 3.1: Kruskal-Wallis H ranked test of attendance by class type excluding canceled classes

	Attendance
Kruskal-Wallis H	3.296
Asymp. Sig.	.069

For when the cancelled classes are taken into consideration the null hypothesis is rejected at the 5% confidence interval. Meaning that there is a real diference between the means of the groups when the cancelled classes have an attendance of 0 and aren't ommited.

For when the cancelled classes are excluded the null hypothesis is not rejected at the 5% significance level with a significance level of 6.9% there isn't enough evidence to show a real difference between the attendance between the class types.

Conclusion

There is a real difference in attendance between the class types this was observed from the EDA done to visualise and place numbers on the mean attendance of the groups. Then the requirements to choose the correct statistical test where chosen and the difference in variance had to be asses to see if there was a real difference in variance. After ascertaining that there was a real difference in attendance the requirements where redone and a proper statistical test was chosen. The chosen test was the Kruskal-Wallis H test this was chosen because it met the greatest number of requirements out of the tests we considered. Its main feature was its ability to handle groups with different variances. After performing the statistical test for when the cancelled classes were ignored there was insufficient evidence to reject the null hypothesis with only a 6.9% significance level. For when the attendance of the cancelled classes was set to 0 there was evidence to reject the null hypothesis at the 5% level. In conclusion based upon the tests performed I would say that there is a real difference between the attendance based upon the type of class.