# Descriptive data and inferential statistic

## Exercise worksheet.

This exercise is part of the resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20846>

You are reading a paper reporting about a randomised controlled trial, where researchers tested whether a digital mindfulness intervention was effective in preventing primary teachers’ burnout. The sample comprises of 257 teachers, and the descriptives table reports the following:

|  |  |  |
| --- | --- | --- |
| Participant characteristics | Intervention group (n = 130) | Control group (n= 127) |
|  |  |  |
| **Age. Mean (SD)** | 31.8 (3.7) | 26.9 (4.1) |
|  |  |  |
| **Gender. n (%)** |  |  |
| Female | 120 (51.9%) | 111 (48.1%) |
| Male | 10 (38.4%) | 16 (61.6%) |
|  |  |  |
| **Prior mindfulness experience** |  |  |
| None | 38 (29.3%) | 12 (9.4%) |
| Limited | 15 (11.5%) | 27 (31.3%) |
| Some | 50 (38.5%) | 67 (42.8%) |
| A lot | 27 (20.7%) | 21 (16.5%) |
|  |  |  |
| **Satisfaction with school leadership** |  |  |
| Dissatisfied | 42 (32.3%) | 51(40.1%) |
| Neither | 58 (44.6%) | 65 (51.2%) |
| Satisfied | 30 (23.1%) | 11 (8.7%) |

1. Age variable
	1. What do you notice at a first glance when considering mean age?
	2. What are your first thoughts about standard deviations?
	3. What would you do to check your assumptions?
2. Gender variable
	1. What do you notice at a first glance when looking at the raw counts of gender?
	2. What do you think about how percentages were calculated? Could have there been different ways to calculate the frequency distribution?
3. Prior mindfulness experience
	1. What do you notice at a first glance when looking at the frequency distribution of this variable between intervention and control group?
	2. Would you have any concerns? How would you check if your concerns were valid?
4. Satisfaction with school leadership
	1. What do you notice at a first glance when looking at the frequency distribution of this variable between intervention and control group?
	2. Would you have any concerns? How would you check if your concerns were valid?

Answers:

**1a.** The mean age of teachers in the intervention group is higher than that of the control group. I would be interested to check how this sample relates to the national average. The government publishes figures on the school workforce in England: <https://explore-education-statistics.service.gov.uk/find-statistics/school-workforce-in-england> but age data are reported in age categories, so we would not be able to see how our sample compares to the population.

**1b.** Upon a first glance, the standard deviations look similar. Remember, the standard deviation is the average amount by which all values deviate from the mean. This means that if you select any participant in the intervention group, their age is typically 3.7 years less or more than the mean, so their age is likely to be between 28.8 and 35.5 years. For the control group, any participant age will typically be between 22.8 and 31 years. If you look at the spread across groups, you notice that they are broadly similar, with the control group year ages deviating more from the mean than the intervention group ages. Approximately 95% of the sample was at an age that was within two standard deviations of the sample mean.

**1c.** Check, if available, the mean age of teachers in the country where the study is taking place. If the study sample differs substantially from the national average, then the results of the trial might not generalise to the teaching population, as they might be relevant only for younger teachers. You could also search the literature to check whether age plays a role in the effectiveness of mindfulness, for example, are younger people more likely to benefit from mindfulness than older people? Always remember: whatever your assumptions, if authors have conducted the sampling and randomisation process correctly, then you can be confident that then we should be reassured that any differences we see in the outcomes between groups are due to the intervention rather than to differences between group characteristics or other factors.

**2a**. Upon a first glance, the raw counts show that there are more females than males in the sample. We know that, broadly speaking, more women than men are teachers, but does this mirror the national distribution?

**2b**. The frequency distribution has not been displayed in the most informative manner. At the moment, the authors are reporting that 51.9% of females were in the intervention group, and 48.1% were in the control group. Instead, we are interested in the gender distribution between the intervention and control groups. Because the two groups display in separate columns, we will look for differences in the column percentages. Whether you focus on row percentages or column percentages in a table depends on the question you want to answer. If instead of having “intervention” and “control” as our columns we had “work setting” with “rural” and “urban” as columns, and our focus were on whether work setting distribution depends on gender, then it would have been sensible to focus on row percentages.

**3a.** At a first glance, it is apparent that more participants in the intervention group have no mindfulness experience than in the control group. However, this is the opposite when we consider the category “limited” experience. If we were to bundle up “none” and “limited” in one category, we would have 40.8% in the intervention group and 40.7% in the control group, which are very similar. This could reassure us that the distributions of participants with little mindfulness experience are similar between the intervention and control group.

**3b.** You could check whether prior mindfulness experience has been found to be a significant predictor of mindfulness effectiveness. Remember: whatever your assumptions, if authors have conducted the sampling and randomisation process correctly, then you can be confident that any differences we see in the outcomes between groups are due to the intervention rather than to differences between group characteristics or other factors.

**4a.** At a first glance, it appears that more participants in the intervention group are satisfied with their school leadership teams than in the control group. Do we have any national data that report teachers’ satisfaction with school leadership to compare this with?

4**b.** You could check whether satisfaction with leadership has been found to be a significant predictor of burnout. If it is, then you might expect researchers to adjust for this variable in their analysis. Adjustment for baseline variables can be considered when there is a known association between baseline characteristics (i.e. satisfaction with leadership) and the primary outcome (i.e. burnout). Remember: whatever your assumptions, if authors have conducted the sampling and randomisation process correctly, then you can be confident that any differences we see in the outcomes between groups are due to the intervention rather than to differences between group characteristics or other factors.

In the same paper, authors produce a table to complement the descriptive statistics. They worry that the differences in baseline characteristics are too wide, so they run a series of tests to check whether the variable distributions are significantly different between groups. Specifically, they are looking for statistical significance at p <.05.

They report the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Participant characteristics | Intervention group (n = 130) | Control group (n= 127) | t-value | p-value |
|  |  |  |  |  |
| **Age. Mean (SD)** | 31.8 (3.7) | 26.9 (4.1) | 10.063 | <.0001 |
|  |  |  | Χ2 | p-value |
| **Gender. n (%)** |  |  |  |  |
| Female | 120 (51.9%) | 111 (48.1%) | 1.700 | .192 |
| Male | 10 (38.4%) | 16 (61.6%) |  |  |
|  |  |  |  |  |
| **Prior mindfulness experience** |  |  |  |  |
| None | 38 (29.3%) | 12 (9.4%) | 20.136 | <.0001 |
| Limited | 15 (11.5%) | 27 (31.3%) |  |  |
| Some | 50 (38.5%) | 67 (42.8%) |  |  |
| A lot | 27 (20.7%) | 21 (16.5%) |  |  |
|  |  |  |  |  |
| **Satisfaction with school leadership** |  |  |  |  |
| Dissatisfied | 42 (32.3%) | 51(40.1%) | 10.040 | .006 |
| Neither | 58 (44.6%) | 65 (51.2%) |  |  |
| Satisfied | 30 (23.1%) | 11 (8.7%) |  |  |

**Questions:**

1. How do you interpret the t-value and p-value related to age?
2. How do you interpret the chi-square statistic value and p-value related to gender?
3. How do you interpret the chi-square statistic value and p-value related to prior mindfulness experience?
4. For variable “satisfaction with school leadership” what does the p-value .006 mean compared to the p-value of “prior mindfulness experience” <.0001?
5. What are the issues related to significance testing for different distribution of variables?

**Answers:**

1. The t-value per se does not convey any information. With the t-test, the sample mean is compared to the null hypothesis (i.e. means from the populations are equal). If t-value = 0 it indicates that the sample results equal the null hypothesis. There are tables available online that report t-distributions, but you might want to focus on the p-value instead. The p-value is <.0001, indicating that the null hypothesis can be rejected.
2. Similarly to the t-value, the chi-square statistic on its own does not convey any information. There are tables reporting critical values of the chi-square distribution that can be found online or in most statistics textbooks. This statistic is only meaningful when read together with the p-value. Looking at the p-value, we see that the null hypothesis cannot be rejected.
3. Similarly to the t-value, the chi-square statistic on its own does not convey any information. There are tables reporting critical values of the chi-square distribution that can be found online or in most statistics textbooks. This statistic is only meaningful when read together with the p-value. Looking at the p-value, we see that the null hypothesis can be rejected. However, it is important to note that if we collapsed “none” and “limited” in one category, we would have 40.8% of observations in the intervention group and 40.7% in the control group, which are very similar distributions.
4. While you technically can compare two p-values when data are from the same sample, it would be incorrect to say that the difference in the “prior experience of mindfulness” distribution between intervention and control group is more statistically significant than that of “satisfaction with poor leadership”. One should only infer that the frequency probability of the “prior mindfulness experience” between groups assuming a null hypothesis is 1 per 10,000 and that of “satisfaction with school leadership” is 6 per 1,000. It is also important to remember that p-values are not a measure of effect.
5. The biggest issue is the violation of the underlying assumption of these tests: they are based on detection of a significant difference in the population, or specifically, from two populations (intervention and control group), but we don’t have two populations, we have a sample that we recruited from a population (i.e. primary teacher) and the randomly divided the sample in two. Significance testing for different distributions in trials should be, therefore, avoided.

Further reading:

1. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med. 2018/08/01/ 2018;210:2-21. doi: https://doi.org/10.1016/j.socscimed.2017.12.005

2. Staggs VS. Pervasive errors in hypothesis testing: Toward better statistical practice in nursing research. Int J Nurs Stud. 2019/10/01/ 2019;98:87-93. doi: https://doi.org/10.1016/j.ijnurstu.2019.06.012

3. Griffiths P, Needleman J. Statistical significance testing and p-values: Defending the indefensible? A discussion paper and position statement. Int J Nurs Stud. 2019/11/01/ 2019;99:103384. doi: https://doi.org/10.1016/j.ijnurstu.2019.07.001

4. Cooksey RW, Cooksey RW. Descriptive statistics for summarising data. Illustrating statistical procedures: Finding meaning in quantitative data. 2020:61-139. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7221239/>

5. Meyer KE, Van Witteloostuijn A, Beugelsdijk S. What’s in a p? Reassessing best practices for conducting and reporting hypothesis-testing research. Research Methods in International Business. 2020:77-110. <https://link.springer.com/chapter/10.1007/978-3-030-22113-3_4>

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK