Inf2 – Foundations of Data Science 2022

Topic: A/B testing

22nd January 2023

Recommended reading:

- *Modern Mathematical Statistics with Applications*, Chapter 10

1 The principle of A/B Testing

A/B testing **A/B testing** is a method for assessing how changes to design of a system affect user behaviour. Figure 1 shows a hypothetical example, in which two versions of a website are presented to users selected at random. Group A gets the version with the blue button and group B gets the version with the green button with the inviting arrow. The numbers of users clicking-through is then measured. There are a number of commercial systems to implement A/B testing.

Statistical Questions in A/B testing

1. Is A *significantly* better than B?
2. How much better is A than B?

Generating confidence intervals for A/B learning using statistical simulations Let's imagine that we present the two versions of the page to group A and to group B the same number of times, n. We find that group A clicks through on 70% of occasions and group B on 72%. We'll call the underlying proportions of users that click through that we are trying to estimate p_A and p_B, and we will define the difference that we are trying to estimate:

$$d = p_A - p_B \tag{1}$$

The difference d is positive when A is better than B. We can address the question of how much better than A is than B by finding a point estimate of d – the larger d the better A is than B. We can address the question of if A is significantly better than B by finding a confidence interval.

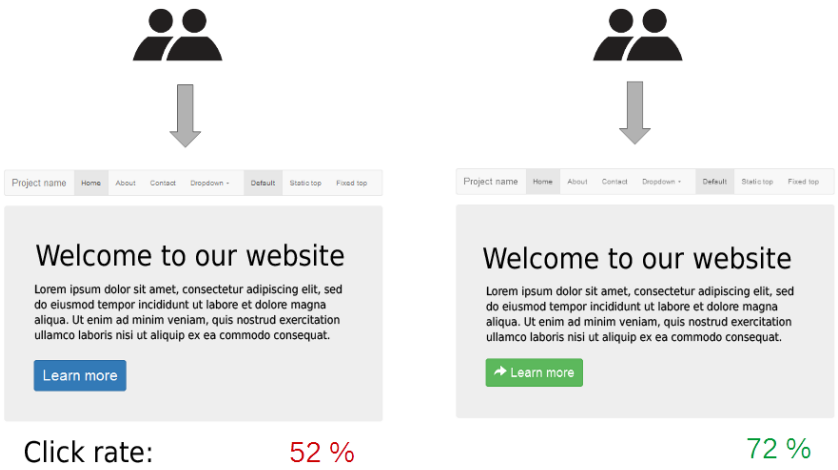

Figure 1: A/B Testing. Group A is shown the web page on the left; group B the one on the right. Image credit: Maxime Lorant, Wikimedia, CC SA 4.0.

The natural point estimators for p_A, p_B and d are:

$$\hat{p}_A = \frac{n_A}{n} \quad , \quad \hat{p}_B = \frac{n_B}{n} \quad \text{and} \quad \hat{d} = \hat{p}_A - \hat{p}_B \tag{2}$$

where n_A and n_B are the actual numbers clicking through from A and B.

To find the confidence interval, we can use a statistical distribution of d, assuming the underlying proportions in populations A and B are given by the point estimates $\hat{p}_A$ and $\hat{p}_B$. The routine to generate the sampling distribution of d looks like:

- For j in $1, \ldots, k$
  - Sample n_A^* from binomial distribution with parameters n and $\hat{p}_A$
  - Sample n_B^* from binomial distribution with parameters n and $\hat{p}_B$
  - Compute and store difference in proportions

$$d_j^* = n_A^*/n - n_B^*/n$$

- Plot the distribution of d^* and compute the desired quantities

The result is shown in Figure 2. The point estimate $\hat{d} = -0.02$, suggesting that B is better than A. However, the 95% confidence interval is $(-0.06, 0.02)$, which contains the value $d = 0$, suggesting that A and B could be equally effective.

Note that the area to the left of $\hat{d} = 0$ in the bootstrap distribution is about 85% and the area to the right is about 15%. We interpret this as meaning that there is an 85% chance that version B is better than version A – but there is still a 15% chance that it isn't.

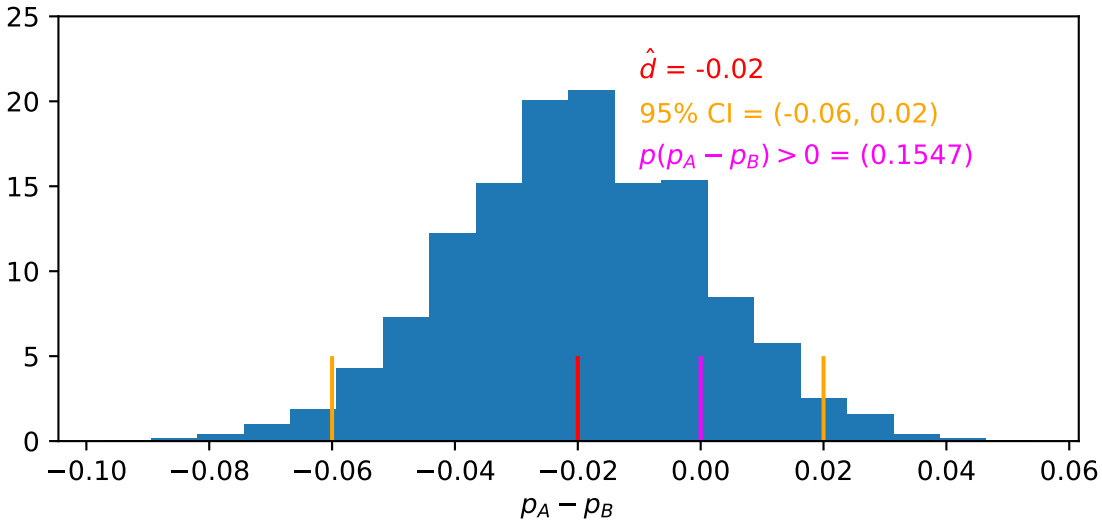

Figure 2: Bootstrap simulation of A/B test with $n = 1000$, $p_A = 0.70$ and $p_B = 0.72$.

Undertaking a hypothesis test for A/B learning using statistical simulations It's also possible to approach this A/B problem as a hypothesis test. We leave it as an exercise to formulate the problem in this way and write a statistical simulation.

2 Increasing certainty in A/B testing

Getting a more certain result To be more certain, we could keep the test running. But, assuming that the population proportions are $p_A = 0.70$ and $p_B = 0.72$ how many runs would we need to have a chance of (say) only 1% that A is better than B?

The brute force approach to find out how big n should be is to run the bootstrap again, with different values of n (Figure 3). As n increases, the distribution, and the 95% confidence interval gets narrower. By $n = 10000$, we can see that the upper end of the 99% confidence interval is now less than 0. The chance of the underlying proportion p_A being higher than p_B is around 0.0012. We can therefore say that a 99.999% confidence interval is $(-\infty, 0)$.

When to stop sampling Suppose we had collected our first $n = 1000$ A and B responses in 2 hours on a Monday afternoon. We're quite excited by the result, and reckon that we need to keep it running up to $n = 10000$ in order to be 99.999% certain. This will probably take us to Tuesday afternoon, we'll then write a report for the boss, and be done by Wednesday. What could possibly go wrong?

We've made a hidden assumption that every period of the week is like a Monday afternoon. What if people prefer blue to green in the evening? What if the Monday afternoon demographic is older, but the weekend demographic is younger? We may wish to collect at least a full week of data to check that our result really is robust – a week's worth of data should mean

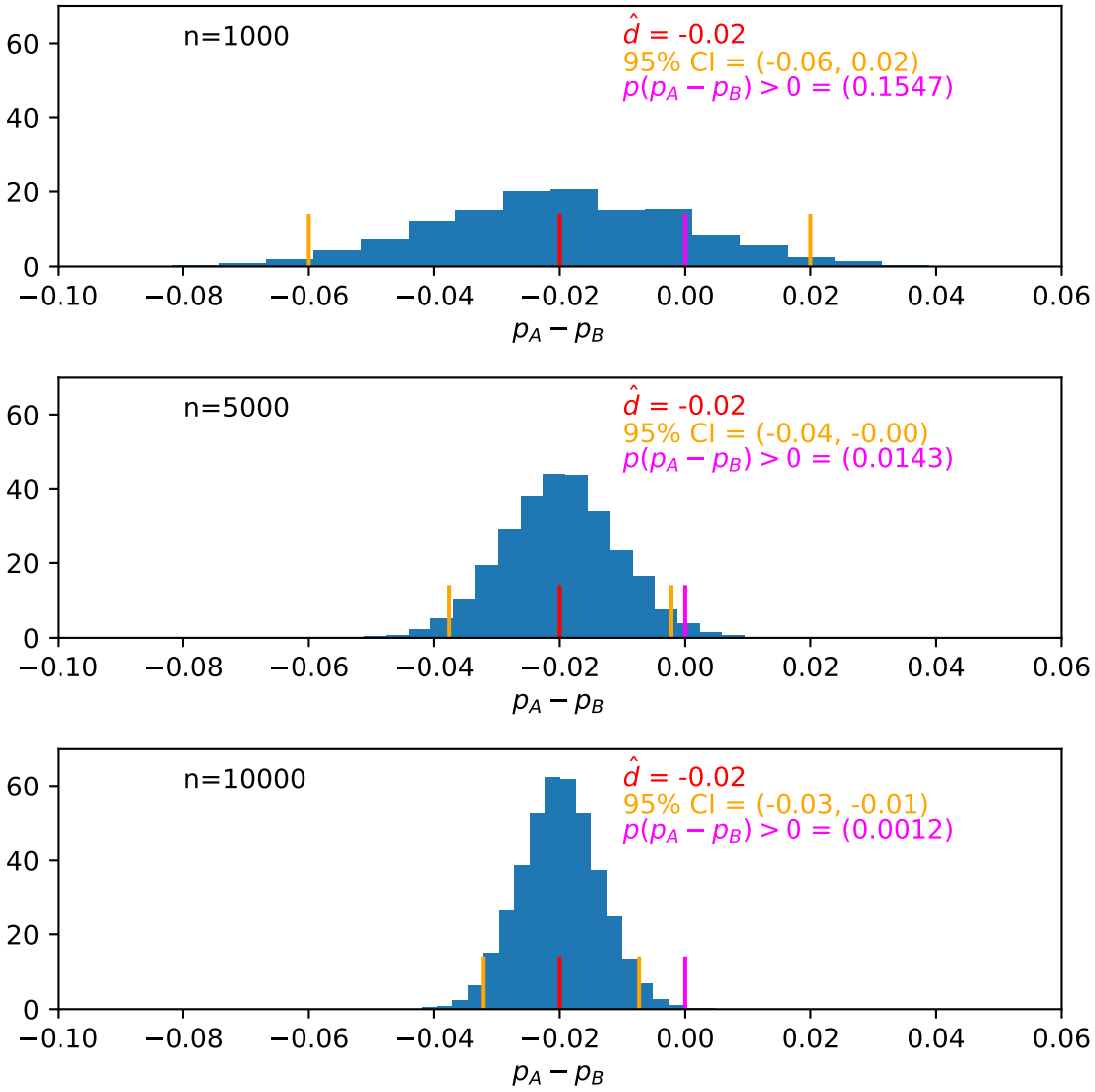

Figure 3: Bootstrap simulation of A/B test with $p_A = 0.70$ and $p_B = 0.72$, and varying numbers of n.

that any day- or time-specific effects are eliminated, or at least greatly reduced.

3 Large sample theory of A/B testing

As with the confidence intervals and hypothesis testing for the sample mean, we can use a theoretical approach to determine confidence intervals or undertake hypothesis testing when doing A/B testing. We have already determined that the estimators for the population proportions are:

$$\hat{p}_A = \frac{n_A}{n} \quad ; \quad \hat{p}_B = \frac{n_B}{n} \tag{3}$$

Now we are interested in the difference $d = p_A - p_B$ between our population proportions. An unbiased estimator of it is:

$$\hat{d} = \hat{p}_A - \hat{p}_B \tag{4}$$

Supposing the population proportions are p_A and p_B, we expect the number of successes in n trials to be binomially distributed, with the standard deviations of n_A and n_B being:

$$\sigma_{n_A} = \sqrt{np_A(1-p_A)} \quad ; \quad \sigma_{n_B} = \sqrt{np_B(1-p_B)} \tag{5}$$

Dividing through by n and replacing p_A and p_B by their estimates, we get the estimated standard errors of the estimators $\hat{p}_A$ and $\hat{p}_B$:

$$\hat{\sigma}_{\hat{p}_A} = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n}} \quad ; \quad \hat{\sigma}_{\hat{p}_B} = \sqrt{\frac{\hat{p}_B(1-\hat{p}_B)}{n}} \tag{6}$$

Since the samples from A and B are independent, the variance of the estimator of the difference in proportions $\hat{d}$ is equal to the sum of the variances of $\hat{p}_A$ and $\hat{p}_B$. We take the square root to get the standard error of the estimator $\hat{d}$:

$$\hat{\sigma}_{\hat{d}} = \sqrt{\hat{\sigma}^2_{\hat{p}_A} + \hat{\sigma}^2_{\hat{p}_B}} = \frac{\sqrt{\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)}}{\sqrt{n}} \tag{7}$$

We'll assume that n is large, in which case the Central Limit Theorem applies, and we can assume that there is little variance in the estimated standard error of $\hat{d}$. We can therefore assume that the statistic

$$Z = \frac{\hat{d} - 0}{\hat{\sigma}_{\hat{d}}} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{(\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B))/n}} \tag{8}$$

is normally distributed. We can then use the z-distribution to calculate confidence intervals.

Worked example We'll use figures we used for the bootstrap to find the 95% confidence interval theoretically:

$$\hat{d} = \hat{p}_A - \hat{p}_B = 0.70 - 0.72 = -0.02 \tag{9}$$

$$\hat{\sigma}_{\hat{d}} = \frac{\sqrt{\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)}}{\sqrt{n}} = \frac{\sqrt{0.70(1-0.70) + 0.72(1-0.72)}}{\sqrt{1000}} = 0.020 \tag{10}$$

For a 95% confidence interval (which makes sense here), we need to use the z critical value $z_{0.025} = 1.96$. The confidence interval for $\hat{d}$ is therefore

$$\begin{aligned}&(\hat{d} - z_{0.025}\hat{\sigma}_{\hat{d}}, \hat{d} + z_{0.025}\hat{\sigma}_{\hat{d}})\\ =&(-0.2 - 1.96 \times 0.020, -0.2 - 1.96 \times 0.020)\\ =&(-0.60, 0.20)\end{aligned} \tag{11}$$

This is almost exactly the same as the bootstrap estimates.

4 Issues in A/B testing

Statistical versus practical significance in A/B testing There is an important distinction between statistical significance and practical significance. We might test the run time of two versions (A and B) of a webserver program on random hits from users. In the example that we've just seen if we make n large enough, we can show that there B is better than A with a p-value of 0.001. However, is the difference of 2% actually that meaningful? In this case it is still probably worth it, since it requires little or no extra effort or energy to create a green button rather than a blue button. But if the processing required to serve version B used a lot more energy, maybe that 2% improvement wouldn't be worth it.

Quoting from the ASA statement on p-values again:

> A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs. (Wasserstein and Lazar, 2016)

Ethical questions in A/B testing There are a number of ethical issues we should consider in A/B testing:

- We are undertaking experiments on people
- In a commercial situation, users do not give informed consent.
- In an academic situation, informed consent is required – but how can we get this informed consent without affecting the experiment?
- What about data protection?

The experiment in which Facebook manipulated news feeds to explore the effect on users' moods (Kramer et al., 2014) was an example of A/B testing that was widely seen as problematic (Verma, 2014) because of a lack of informed consent, no opportunity to opt-out and no institutional review of the experiment, because it was carried out by a private company. Many A/B tests are arguably not having such a significant effect on users – but they may have some effect nonetheless.

It's therefore important to reflect on an A/B test before setting it up. Questions might include:

- Would I feel comfortable if this change was tested on me?
- What potential harms could be caused to users?

Comparing two samples more generally The methods we show here predate A/B testing, and come under the heading of "comparing two samples". Sometimes our groups may have numeric response variables, and need not be from a controlled situation – for example the scores of students on a calculus course in Semester 1 (Group A) and Semester 2 (Group B, Figure 4). There are a number of variations on the theoretical method presented here for generating confidence intervals and undertaking hypothesis tests. Which one to use depends on whether the data is approximately normal, how large the sample size is and if the data comes from a set of paired measurements, e.g. pairs of measurements of temperature at dawn and dusk from a number of different locations. *Modern Mathematical Statistics with Applications* chapter 10 has many more details, and we go through the maths example below.

5 Extras

There are no videos that go with these lecture notes. They are non-examinable, but you may find them of interest.

5.1 Samples with numeric responses

The problem of two samples A common problem is assessing whether the difference between distributions of a variable recorded in two populations is similar or different. The different populations may arise from a randomised controlled trial, in which participants are assigned randomly to a control group or a treatment group. However, we can also test differences between two observed groups.

For example, the maths scores of students taking a calculus course in Semester 2 seem to be lower than the grades in Semester 1 (Figure 4). We'll call the grades of the m Semester 1 students $x_1, \ldots, x_m$ and the grades in the n Semester 2 students $y_1, \ldots, y_n$. We can compute the two means of the grades of each group, and find that difference is $\overline{x} - \overline{y} = 2.37$. But, assuming that the maths scores are representative of performance in Semester 1 and Semester 2 in other years, we'd like to find a 95% confidence interval for the difference – in other words an interval that we would expect to contain the true, underlying difference 95% of the time.

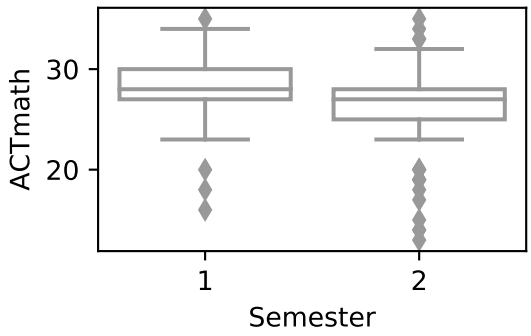

Figure 4: Maths scores in Semester 1 (Autumn) and Semester 2 (Spring). Data from Edge and Friedberg (1984), via Devore and Berk (2012).

We already know how to compute a confidence interval of one sample, using the bootstrap. Can we adapt it to give us a confidence interval around the difference between two means?

Applying the bootstrap To apply the bootstrap, on each bootstrap step we sample with replacement from both groups:

- For j in $1, \ldots, B$
  - Take sample x^* of size m from the sample *with replacement*
  - Take sample y^* of size n from the sample *with replacement*
  - Compute the sample mean of the new samples, $\overline{x_j^*}$ and $\overline{y_j^*}$
  - Compute and store the difference in the sample means $\overline{x_j^*} - \overline{y_j^*}$
- Plot the bootstrap distribution of $\overline{x_j^*} - \overline{y_j^*}$
- We can also compute the bootstrap estimator of the variance of the difference between the mean:

$$s^2_{\text{boot}} = \frac{\sum_{j=1}^{B} (\overline{x_j^*} - \overline{y_j^*})^2}{B-1}$$

The bootstrap distribution is shown in Figure 5. We can see that the 95% confidence interval is 1.11 to 3.66.

5.2 Relationship between hypothesis testing and confidence intervals

Suppose we had wanted to test the hypothesis that average performance in Semester 2 is different to average performance in Semester 1. Our null hypothesis would be

> H_0: The mean performance in semester 1 is the same as the mean performance in semester 2.

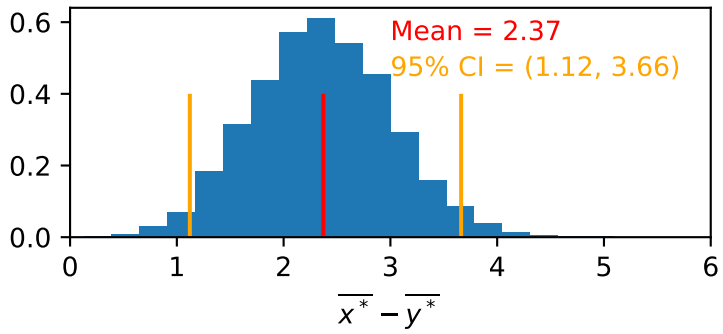

Figure 5: Bootstrap distribution of the maths grades example.

The alternative hypothesis would be:

> H_a: The mean performance in semester 1 is different from the mean performance in semester 2.

We've previously simulated the null hypothesis (here, no difference in means) to generate a distribution of what the test statistic (here, the difference in the sample means) would be under the null hypothesis, and then compared this distribution with the observed value of our test statistic.

It turns out that there is a duality between confidence intervals and hypothesis testing. Instead of the distribution of the sampling distribution generated under the null hypothesis, we have used the observed data to generate the distribution of the estimator for the parameter corresponding to the test statistic. Instead of the observed value of the test statistic, we have the value the parameter would take under the null hypothesis.

In this example, we used the bootstrap to estimate the distribution of the estimator of the difference of the means $\mu_x - \mu_y$ (Figure 5), and we could then ask if the null hypothesis value of the difference in the means (0) lies in either of the rejection regions in the tails of that distribution. If so, we can reject at the level corresponding to the size of the tails. Alternatively, we could compute a p-value, by finding at what quantile the null hypothesis value (here 0) lies on the distribution. In this case we would find $p = 0$, so the null hypothesis would be rejected.

For more on the duality between confidence intervals and hypothesis testing see this Quora article.

5.3 The theoretical method of testing for differences between groups

Samples with known variance We can also undertake hypothesis testing and compute confidence intervals for the difference between the means of two samples theoretically. The assumptions are:

- $X_1, \ldots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2

- $Y_1, \ldots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2
- The samples are independent of each other.

Estimator of the difference The difference between the sample means $\overline{X} - \overline{Y}$ is an unbiased estimator of the difference between the true means $\mu_1 - \mu_2$. This follows from $\overline{X}$ being an unbiased estimator of μ_1 and $\overline{Y}$ being an unbiased estimator of μ_2. The standard deviation of the estimator is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \tag{12}$$

This follows from the two samples being independent, so $V(\overline{X}) - V(\overline{Y}) = V(\overline{X}) + V(\overline{Y}) = \sigma_1^2/m + \sigma_2^2/n$.

Theoretical Distribution for large samples As we with the sample mean of one population, we define a standardised test statistic, which we expect to be zero in the case of a null hypothesis that the true difference between the population means is $\mu_1 - \mu_2$:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/m + S_2^2/n}} \tag{13}$$

The denominator is the sample standard deviation of the estimator, and is a random variable. As with the sample of one mean, for small m and n this will vary considerably between different samples, and so we do have to consider these random effects. However, for large m and n, it will approximate the true population means, and its variability is low enough to consider it as a fixed parameter. In the limit of large n and m, the central limit theorem suggests that the distribution of the statistic should be normal, so we can use a z-test.

Theoretical Distribution for small samples In the case of smaller samples, the variability of the sample standard deviation has to be taken into account, and it turns out that the sampling distribution of the standardised statistic is a t-distribution with a number of degrees of freedom ν that depends on the standard deviations of both distributions:

$$\nu = \frac{\left(s_1^2/m + s_2^2/n\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} \tag{14}$$

To determine a confidence interval of $1 - \alpha$, we can find the t critical value $t_{\alpha/2,\nu}$ which we will expect will containe the mean difference on $1 - \alpha$ percent of replications. We can then use the t critical value to transform back to the unstandardised variables using the standard error in the difference of the mean:

$$\overline{x} = \overline{y} \pm t_{\alpha/2,\nu}\sqrt{s_1^2/m + s_2^2/n} \tag{15}$$

References

Devore, J. and Berk, K. (2012). *Modern Mathematical Statistics with Applications*. Springer Texts in Statistics. Springer New York, New York, NY, second ed.

Edge, O. P. and Friedberg, S. H. (1984). 'Factors affecting achievement in the first course in calculus'. *Journal of Experimental Education* **52**:136–140

Kramer, A. D., Guillory, J. E. et al. (2014). 'Experimental evidence of massive-scale emotional contagion through social networks.' *Proc Natl Acad Sci U S A* **111**:8788–90. URL `https:dx.doi.org:/10.1073/pnas.1320040111`

Verma, I. M. (2014). 'Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks.' *Proc Natl Acad Sci U S A* **111**:10779. URL `https://dx.doi.org/10.1073/pnas.1412469111`

Wasserstein, R. L. and Lazar, N. A. (2016). 'The ASA statement on *p*-values: Context, process, and purpose'. *The American Statistician* **70**:129–133. URL `https://doi.org/10.1080/00031305.2016.1154108`