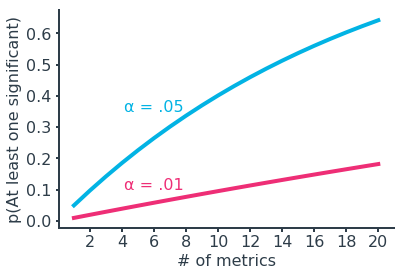
**Analyzing Multiple Metrics**

If you're tracking multiple evaluation metrics, make sure that you're aware of how the Type I error rates on individual metrics can affect the overall chance of making some kind of Type I error. The simplest case we can consider is if we have \_n\_ independent evaluation metrics, and that seeing one with a statistically significant result would be enough to call the manipulation a success. In this case, the probability of making at least one Type I error is given by \alpha\_{over} = 1 - (1-\alpha\_{ind})^n*αover*​=1−(1−*αind*​)*n*, illustrated in the below image for individual \alpha\_{ind} = .05*αind*​=.05 and \alpha\_{ind} = .01*αind*​=.01:

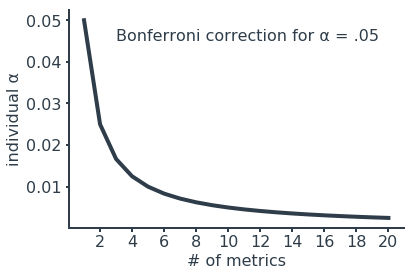
[[](https://classroom.udacity.com/nanodegrees/nd025-ent/parts/347da0f8-5670-4bbf-8587-b02b7c0fe111/modules/db86b648-24c4-4589-8755-458885e0503c/lessons/d64ebb3c-218c-4194-ba95-e50a19683be4/concepts/83ca614f-acbc-4123-8ec2-5739d12b97e9)](https://classroom.udacity.com/nanodegrees/nd025-ent/parts/347da0f8-5670-4bbf-8587-b02b7c0fe111/modules/db86b648-24c4-4589-8755-458885e0503c/lessons/d64ebb3c-218c-4194-ba95-e50a19683be4/concepts/83ca614f-acbc-4123-8ec2-5739d12b97e9)

To protect against this, we need to introduce a correction factor on the individual test error rate so that the overall error rate is at most the desired level. A conservative approach is to divide the overall error rate by the number of metrics tested:

\alpha\_{ind} = \alpha\_{over}/n*αind*​=*αover*​/*n*

This is known as the **Bonferroni correction**. If we assume independence between metrics, we can do a little bit better with the **Šidák correction**:

\alpha\_{ind} = 1-(1-\alpha\_{over})^{1/n}*αind*​=1−(1−*αover*​)1/*n*

[[](https://classroom.udacity.com/nanodegrees/nd025-ent/parts/347da0f8-5670-4bbf-8587-b02b7c0fe111/modules/db86b648-24c4-4589-8755-458885e0503c/lessons/d64ebb3c-218c-4194-ba95-e50a19683be4/concepts/83ca614f-acbc-4123-8ec2-5739d12b97e9)](https://classroom.udacity.com/nanodegrees/nd025-ent/parts/347da0f8-5670-4bbf-8587-b02b7c0fe111/modules/db86b648-24c4-4589-8755-458885e0503c/lessons/d64ebb3c-218c-4194-ba95-e50a19683be4/concepts/83ca614f-acbc-4123-8ec2-5739d12b97e9)

[The Šidák correction is only slightly higher than the line drawn by the Bonferroni correction.](https://classroom.udacity.com/nanodegrees/nd025-ent/parts/347da0f8-5670-4bbf-8587-b02b7c0fe111/modules/db86b648-24c4-4589-8755-458885e0503c/lessons/d64ebb3c-218c-4194-ba95-e50a19683be4/concepts/83ca614f-acbc-4123-8ec2-5739d12b97e9)

In real life, evaluation scenarios are rarely so straightforward. Metrics will likely be correlated in some way, rather than being independent. If a positive correlation exists, then knowing the outcome of one metric will make it more likely for a correlated metric to also point in the same way. In this case, the corrections above will be more conservative than necessary, resulting in an overall error rate smaller than the desired level. (In cases of negative correlation, the true error rate could go either way, depending on the types of tests performed.)

In addition, we might need multiple metrics to show statistical significance to call an experiment a success, or there may be different degrees of success depending on which metrics appear to be moved by the manipulation. One metric may not be enough to make it worth deploying a change tested in an experiment. Reducing the individual error rate will make it harder for a truly significant effect to show up as statistically significant. That is, reducing the Type I error rate will also increase the Type II error rate – another conservative shift.

Ultimately, there is a small balancing act when it comes to selecting an error-controlling scheme. Being fully conservative with one of the simple corrections above means that you increase the risk of failing to roll out changes that actually have an impact on metrics. Consider the level of dependence between metrics and what results are needed to declare a success to calibrate the right balance in error rates. If you need to see a significant change in all of your metrics to proceed with it, you might not need a correction factor at all. You can also use dummy test results, bootstrapping, and permutation approaches to plan significance thresholds. Finally, don't forget that practical significance can be an all-important quality that overrides other statistical significance findings.

While the main focus of this page has been on interpretation of evaluation metrics, it's worth noting that these cautions also apply to invariant metrics. The more invariant metrics you test, the more likely it will be that some test will show a statistically significant difference even if the groups tested are drawn from equivalent populations. However, it might not be a good idea to apply a correction factor to individual tests since we want to avoid larger issues with interpretation later on. As mentioned previously, a single invariant metric showing a statistically significant difference is not necessarily cause for alarm, but it is something that merits follow-up in case it does have an effect on our analysis.