Appendix A – The Mann-Whitney U test

In Section 2.1 we decided to use the Mann-Whitney U test to distinguish the performance of the optimization methods on a given test function f. While less prevalent in the literature than the standard two sample t test, the Mann-Whitney U test serves a similar purpose: determine whether two populations have the same central tendency (Fay and Proschan, 2010). For AutoML problems, these two populations could represent the performance of two optimization methods on the same model selection or hyperparameter tuning problem. The goal of the statistical analysis is to determine if one of them outperforms the other beyond a chosen statistical significance.

Both tests require i.i.d. samples of the relevant values drawn from each population; each sample requires conducting a full hyperparameter tuning and then using the f_{best} trace to compute the needed **Best Found** or **Area Under Curve** values. The t test studies the difference between i.i.d. sample means drawn from both populations. In contrast, the U test studies the relative rankings of the samples from both populations and is thus making a statement about the *medians* of the two populations instead of their means (but not a statement about sample medians). This distinction could be relevant when the populations have significant nonzero skew, as evidenced in Figure 2. A example drawing of sample means is presented below, where random search and PSO are both trying to minimize a function f.

rank	population	value
1	PSO	0.01
2	PSO	0.03
3	PSO	0.04
4	RANDOM	0.05
5	PSO	0.11
6	RANDOM	0.14
7	PSO	0.22
8	PSO	0.23
9	RANDOM	0.25
10	PSO	0.27
11	RANDOM	0.28
12	RANDOM	0.36
13	PSO	0.37
14	RANDOM	0.40
15	PSO	0.41
16	RANDOM	0.45
17	PSO	0.51
18	RANDOM	0.57
19	RANDOM	0.66
20	RANDOM	0.80

Information on these and other statistical tests can be found in, e.g., Wackerly et al. (2014). The t test demands the sample means for the two populations, $\bar{x}_R = 0.396$ for random search and $\bar{x}_P = 0.22$ for PSO, as well as an estimate of the standard error, $s_E = 0.092$ and estimate of the degrees of freedom $\nu = \lfloor 16.63 \rfloor = 16$. For a null hypothesis of equal means, $t = 1.92$, and the alternative hypothesis of unequal means, we observe a p value of 0.072.

To conduct a U test, we must consider the relative locations of PSO and RANDOM in the 1–20 rankings. In particular, we sum the ranking values for both samples, $r_R = 81$ and $r_P = 129$ and then uses the less of those two values to compute the U statistic,

$$U = \min\left(81 - \frac{10(10+1)}{2}, 129 - \frac{10(10+1)}{2}\right) = 26.$$

Looking this up in a table, or using software, provides a p value of 0.038. Notice that, in contrast to the t test, the U test considers only the ordinal relationship between samples; this prevents any single value from having a significant effect on the results.

Appendix B – Revisiting the skewness issue using the Berry-Esseen inequality

In Section 2.1.2, we alluded to the skewness of $Y_{(T)}$, the result of a random search after T function evaluations, where each observed function value Y_i, $1 \leq i \leq T$ is a random variable with distribution determined by the function of interest f and domain $\mathcal{X}$. Figure 2 demonstrated the impact of this empirically for the simple maximization problem involving $f(x) = |x|$ for $x \in [-1, 1]$. In particular, it showed that the n term sample mean of samples drawn from $Y_{(1)}$ failed a Kolmogorov-Smirnov test with greater likelihood for larger T and smaller n. The KS test tests whether the random variables are normally distributed, which they must be for a hypothesis test or confidence interval invoking the central limit theorem to be appropriate.

We can determine, at least for this simple problem, the exact nature of the Berry-Esseen inequality (Korolev and Shevtsova, 2010) governing the quality of the normal approximation. For a random variable $W_n = \frac{1}{n}(V_1 + \ldots + V_n)$ such that $E(V_i) = 0$, $E(V_i^2) = \sigma^2$ and $E(|V_i^3|) = \rho < \infty$, the Berry-Esseen inequality says

$$|F_{W_n}(w) - F_Z(w)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}, \quad \text{for all } w, n, \tag{1}$$

where $C > 0$ is a constant and $Z \sim N(0, 1)$.

Our goal is to study the i.i.d. summation $\bar{Y}_{(1)_n} = \frac{1}{n}\left(Y_{(1)_1} + \ldots + Y_{(1)_n}\right)$, but to apply the Berry-Esseen inequality we will have to subtract out the mean. In Section 2.1.2 we showed $F_{Y_{(1)}}(y) = 1 - [1 - F_Y(y)]^T$, and, for this problem, $F_Y(y) = y$ so $F_{Y_{(1)}} = 1 - (1 - y)^T$. Using this CDF, we can compute

$$E\left(Y_{(1)}\right) = \int_{\mathbb{R}} y \, \mathrm{d}F_{Y_{(1)}}(y) = \int_0^1 yT(1-y)^{T-1} \, \mathrm{d}y = \frac{1}{T+1}.$$

This gives us the random variable $V_i = Y_{(1)_i} - \frac{1}{T+1}$ which satisfies $E(V_i) = 0$. Using this, we can compute

$$\sigma^2 = E(V_i^2) = \int_0^1 \left(y - \frac{1}{T+1}\right)^2 T(1-y)^{T-1} \mathrm{d}y = \frac{T}{(T+1)^2(T+2)}$$

and

$$\rho = E(|V_i|^3) = \frac{-2T(T-1)}{(T+1)^3(T+2)(T+3)} + 12\left(\frac{T}{T+1}\right)^T \frac{T^3}{(T+1)^4(T+2)(T+3)}.$$

When we finally compute this ρ/σ^3 term that governs the quality of the normal approximation in the Berry-Esseen inequality, we see

$$\frac{\rho}{\sigma^3} = -2\frac{(T-1)(T+2)^{1/2}}{T^{1/2}(T+3)} + 12\left(\frac{T}{T+1}\right)^T \frac{T^{3/2}(T+2)^{1/2}}{(T+1)(T+3)}.$$

This quotient is monotonically *increasing* for $T > 1$, which can be determined with the derivative (albeit with a good deal of work and ignoring the fact that T is an integer) or graphically as in Figure 3.

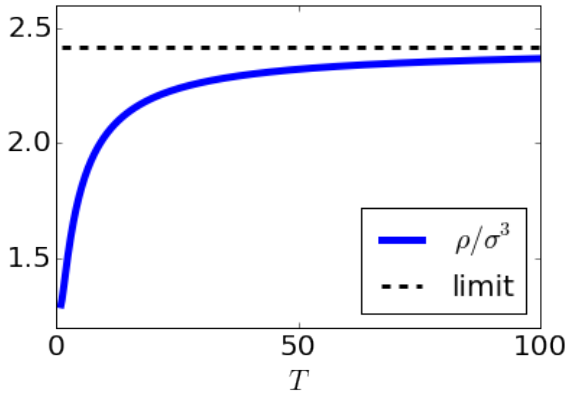

Figure 3: The ρ/σ^3 term in (1) is monotonically increasing with T, impugning the integrity of a t test for small n.

Of course, all is well as long as n is large because $\rho/\sigma^3 \to 12\mathrm{e}^{-1} - 2$ as $T \to \infty$, but for small n, Figure 3 demonstrates that as the quality of the solution improves (T increases) the validity of t tests and central limit theorem based confidence intervals actually diminishes. This is likely the cause of the similar behavior for the failed Kolmogorov-Smirnov test probability in Figure 2.