Confidence Sequences, p-values, and Confidence - Technical Details

Data Science, Adobe Journey Optimizer

Confidence Sequences for individual treatment arms

Let $\hat{\mu}$ be the sample mean, and σ the sample standard deviation after n sample haves been recorded for the treatment arm. Then for any pre-specified constant ρ, Waudby-Smith et al. [1] define:

$$CS_{1-\alpha} := \left\{ \hat{\mu} \pm \hat{\sigma} \sqrt{\frac{2(n\rho^2+1)}{n^2\rho^2} \log\left(\frac{\sqrt{n\rho^2+1}}{\alpha}\right)} \right\},$$

forms a $(1-\alpha)$ Confidence Sequence for the true mean, μ. The parameter ρ is a free parameter that must be tuned. Adobe uses a common constant $\rho^2 = 10^{-2.8} = 0.001585$ as this has been found to provide sufficiently tight bounds across most customers.

In terms of raw statistics, we have

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

as well as the running standard deviation

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2} \tag{1}$$

$$= \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 \right) - \frac{n}{n-1} \hat{\mu}_n^2} \tag{2}$$

For a definition of these quantities in terms of our raw statistics, and expressed in plain english:

$$n = \text{sum of users for particular treatment} \tag{3}$$

$$\sum_{i=1}^{n} x_i = \text{sum of metric, for particular treatment} \tag{4}$$

$$\sum_{i=1}^{n} x_i^2 = \text{sum of squared metric, for particular treatment} \tag{5}$$

Confidence Sequences for Difference Between Treatment and Control Arm, and derivation of p-values

Confidence Sequence for the difference in means

Consider two treatments, with ‘treatmentIds‘ 0 and 1, with $N = N_0 + N_1$ total visitors across the two treatments. In terms of the sample means, $\hat{\mu}_0$ and $\hat{\mu}_1$, as well as the sample standard deviations $\hat{\sigma}_0$ and $\hat{\sigma}_1$ (same definitions as earlier), the confidence sequence for the difference is given by:

$$CS_{1-\alpha} := (\hat{\mu}_1 - \hat{\mu}_0) \pm \Gamma_N$$

where

$$\Gamma_N = \sqrt{\frac{N}{N-1}\left[\frac{N}{N_0}\left(\hat{\sigma}_0^2 + \hat{\mu}_0^2\right) + \frac{N}{N_1}\left(\hat{\sigma}_1^2 + \hat{\mu}_1^2\right) - \left(\hat{\mu}_1 - \hat{\mu}_0\right)^2\right]} \cdot \sqrt{\frac{2(N\rho^2+1)}{N^2\rho^2}\log\left(\frac{\sqrt{N\rho^2+1}}{\alpha}\right)} + \ldots$$

This quantity is derived by considering an "inverse propensity weighted" estimator for the "Average Treatment Effect" in a randomized experiment.

Inverting the Confidence Sequence to find an Always Valid p-value

To find a p-value, we *interpret* the confidence bounds given by our confidence sequence as the normalizing factor for the test statistic. The derivation is as follows:

Recall that for a regular Hypothesis test for the difference in means, we have a test statistic defined as

$$z = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_p},$$

where $\hat{\sigma}_p$ is the pooled sample standard deviation.

For large enough sample sizes, the t-test and z-test are equivalent, and the p-value is then defined in terms of the cumulative distribution of the normal, Φ, as:

$$p\text{-value} = 2(1 - \Phi(|z|))$$

Then, a $1-\alpha$ ***confidence interval*** for the difference in means is given by:

$$CI_{1-\alpha} := \left\{(\hat{\mu}_1 - \hat{\mu}_0) \pm z^*_{1-\frac{\alpha}{2}} \hat{\sigma}_p\right\},$$

where $z^*_{1-\frac{\alpha}{2}}$ is a critical value for the standard normal. For $\alpha = 0.05$, we have $z^* \approx 1.96$.

To derive the equivalent *"always valid" p-value*, we take our expression for the $1-\alpha$ **Confidence Sequence**:

$$CS_{1-\alpha} := \{(\hat{\mu}_1 - \hat{\mu}_0) \pm \Gamma_N\},$$

where Γ_N is the complex expression shown above. We then create an analagous relationship between these confidence bounds and the test statistic:

$$\tilde{z} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{(\Gamma_N / z^*_{1-\frac{\alpha}{2}})},$$

And finally define the p-value to be:

$$p\text{-value} = 2(1 - \Phi(|\tilde{z}|))$$

Confidence

Finally, the confidence C is defined as:

$$C = 1 - p\text{-value},$$

REFERENCES

[1] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. Doubly robust confidence sequences for sequential causal inference. *arXiv preprint arXiv:2103.06476*, 2021.