Power Analysis of Strike Event Proportions

Revised: 2024.06.18

Assoc. Prof. Jeffrey A. Tuhtan, Dept. of Computer Systems, Tallinn University of Technology

Why is a power analysis needed?

A pre-study (*a priori*) power analysis is needed in order to estimate the minimum sample size of passive sensors subjected to one or more treatments, relative to a control group where a treatment has not been applied. A strike event in this study is defined as a sensor having direct contact with the pump impeller during passage through the pump, where the sensor enters the pumping station inlet region under suction and exits after passing completely through the pump impeller region via the discharge nozzle.

In the context of this study, a **successful sensor passage test** must meet the following criteria:

- a sensor which is turned on in the test environment before submersion into the water, where the local atmospheric pressure of each sensor is then recorded for a minimum duration of 15 s,
- a sensor which after activation is manually injected into the test environment,
- and then enters the test environment which is operating at a known fixed operational condition, or is subject to a control environment without influence of the pump impeller,
- which after injection passes completely through the test environment in a continuous manner without abnormal interruption or delay,
- and which is recovered without reasonable expectation of damage by the injection and / or recovery apparati, such that the data collected can be considered to be representative of the test environment as the sensor is subjected to treatment and / or control conditions,
- a **group** is defined in this study as a collection of sensor datasets collected under identical experimental physical conditions (e.g. same injection location, operating point and recovery system). The minimum group sample size for any test should always be n = 30 or greater for sufficient post-hoc testing. A **batch** is a set of deployments, usually 5, 10 or 20 sensors at a time, which are conducted in series, and are combined to form a group.

How is the power analysis performed?

- The null hypothesis (H0) in this study is that there is no difference in the proportion of strike events between two groups (e.g. between treatment and control group).
- The alpha value is the level of significance with 5% chance (Type I error, false positive) that a difference between groups is detected, when there was actually no difference ($\alpha = 0.05$)
- The statistical power of the study is taken as 0.85, which assumes a 15% chance (Type II error, false negative) that no difference between groups is detected, when there was actually a difference ($\beta = 0.15$).
- The effect size (h) is the difference in outcomes between two test groups where we assume: $h = 0.2$ for small effects, 0.5 for moderate effects and 0.8 for large effects.
- Power analysis is performed both before and after the tests are conducted. This is needed to estimate sample sizes prior to conducting the study, and afterwards to update the assumptions and improve sample size estimations in future studies.

How are the results of the power analysis reported?

A power analysis to estimate the suitable sample sizes in this study was carried out using G*Power (V3.1.9.7), which compared an assumed baseline strike probabilities p_{s0} = 0.1 (10%) to 1.0 (100%) against a range of increasing strike probabilities ranging from 0.10 to 0.5 (10% to 50%) in increments of 0.1. For each combination, the effect size and sample sizes required to achieve a significance criterion of α = 0.05 and power of 0.85 is reported in Table 1. Sample sizes were estimated to be sufficient using an exact z-test of proportions with two-sided distribution. A two-sided distribution is needed as it cannot be assumed beforehand whether different treatment conditions (e.g. two different operating points of a pump) are likely to have lower or higher proportions of impeller strikes on sensors relative to the baseline strike probability. The effect sizes of the baseline strike proportions and test strike proportions are provided in Table 2.

Table 1: Sample sizes (n) based on the difference in the proportion of impeller strikes for two groups. To use this table, first assume a baseline strike proportion (leftmost column), then select a test strike proportion (10% to 50%). The value in the table corresponding to this combination is reported as the required sample size, n for a two-sided z-test assuming equal sample sizes (e.g. the sample size, n_{base} for the assumed baseline strike proportion is equal to n_{test}. The coloring corresponds to the level of effort for conducting the tests where n ≤ 30 "easy", 30 < n ≤ 60 "possible", 60 < n ≤ 120 "difficult" and n > 120 "very difficult" are highlighted to assist in adaptive sampling strategies during testing. For all groups, the minimum sample size should be taken as n = 30, regardless of the results in order to perform sufficient post-hoc testing.

Assumed Baseline Strike Proportion	Test Strike Proportion				
	10%	20%	30%	40%	50%
10%		228	72	36	22
20%	228		335	93	44
30%	71	335		407	106
40%	36	93	407		443
50%	22	44	106	443	
60%	15	25	48	111	443
70%	11	16	27	48	106
80%	8	11	16	26	44
90%	5	8	11	15	22
100%	4	5	7	9	12

Table 2: Effect sizes calculated as the difference in proportions between the assumed baseline strike proportion and the test size strike proportions. The coloring corresponds to the effect sizes (h) where h < 0.2 is below detection, 0.2 ≤ h < 0.5 are small, 0.5 ≤ h <0.8 moderate and h ≥ 0.8 are large effects.

Assumed Baseline Strike Proportion	Test Strike Proportion Effect Sizes				
	10%	20%	30%	40%	50%
10%		0.1	0.2	0.3	0.4
20%	0.1		0.1	0.3	0.3
30%	0.2	0.1		0.1	0.2
40%	0.3	0.2	0.1		0.1
50%	0.4	0.3	0.2	0.1	
60%	0.5	0.4	0.3	0.2	0.1
70%	0.6	0.5	0.4	0.3	0.2
80%	0.7	0.6	0.5	0.4	0.3
90%	0.8	0.7	0.6	0.5	0.4
100%	0.9	0.8	0.7	0.6	0.5

Example 1: *We begin by assuming a baseline strike proportion of 100%. This corresponds to a strike rate expected on adult nase with a body length of 40 cm in a conventional axial flow pump.*

For this test, we create null hypotheses H_0 and corresponding alternative hypotheses H_a:

H_0: $p_{test} = p_{base}$ (the strike proportion from the test is the same as the baseline strike proportion)

H_a: $p_{test} \neq p_{base}$ (the proportions are not equal)

A batch of 10 sensors are deployed in the test environment, and 5 of them have strike events (50%). We deploy a second and third batch of 10 sensors to meet the minimum n = 30 for the test group. After recovering the second and third batches, we see that the second batch had a strike event rate of 40% (4 out of 10 had strikes) and the third 60%. The proportion of strikes in the test group is (5 + 4 + 6)/(10 + 10 + 10) = 0.5, or 50%. Looking at Table 1, we see that for a 50% test strike proportion we need n = 12. Based on this, we end the test based on strike proportions, as we have a sufficient number of strike events assuming a baseline strike proportion of 100%.

The input code in R for the post-hoc test:

```
prop.test(x = c(30, 15), n = c(30, 30), alternative = "two.sided", cor = T, conf.level = 0.95)
```

The output information from this test in R:

```
data:  c(30, 15) out of c(30, 30)
X-squared = 17.422, df = 1, p-value = 2.993e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2877473 0.7122527
sample estimates:
prop 1 prop 2
   1.0    0.5
```

Rule-based evaluation of the two-sided test results:

Because the p-value (0.00002993) is far less than our choice of α (0.05), we choose to reject the null hypothesis that the strike proportion of the test environment is equal to the assumed baseline strike proportion. The p-value is the chance, roughly 3/100 000 in this case, of a false positive if we repeated this experiment many times. The 95% confidence interval ranges from 28.8% to 71.2%, meaning that if we duplicated our tests with the same sample sizes, we can reasonably expect test environment strike proportions to occur within this range. If we want to reduce the range of the confidence intervals for this test, additional batches are required. The post-hoc power of this test is 0.99, which is even higher than our assumed power of 0.85 to start with.

Rule-based decision for this test:

There is sufficient evidence with 95% confidence to conclude that the test environment strike proportion of 50% is not the same as the assumed baseline strike proportion of 100%.

Now we check if the baseline strike proportion is greater than the test strike proportion. To do this, we need to run a single-sided test in R using the "greater" option:

H_0: $p_{base} = p_{test}$ (the strike proportion of the test is the same as than the baseline proportion)

H_a: $p_{base} > p_{test}$ (the strike proportion of the baseline is greater than the test proportion)

The input code in R for the post-hoc test:

```
prop.test(x = c(30, 15), n = c(30, 30), alternative = "greater", cor = T, conf.level = 0.99)
```

The output of this test in R:

```
data:  c(30, 15) out of c(30, 30)
X-squared = 17.422, df = 1, p-value = 1.497e-05
alternative hypothesis: greater
99 percent confidence interval:
 0.2543011 1.0000000
sample estimates:
prop 1 prop 2
   1.0    0.5
```

Rule-based evaluation of the single-sided test results:

Because the p-value (0.00001497) is much less than our choice of α (0.01), we choose to reject the null hypothesis that the strike proportion of the test environment is the same as the assumed baseline strike proportion. We are 99% confident that the baseline strike proportion is greater than the test environment strike proportion. The post-hoc power of this test is 0.99, which is even higher than our assumed power of 0.85 to start with.

Rule-based decision for this test:

There is sufficient evidence to conclude with 99% confidence that the test environment strike proportion of 50% is lower than the assumed baseline strike proportion of 100%.

Example 2: *We begin by assuming a baseline strike proportion of 50%. This corresponds to a strike rate expected on adult roach with a body length of 20 cm in a fish-friendly pump.*

For this test, we create null hypotheses H_0 and corresponding alternative hypotheses H_a:

H_0: $p_{base} = p_{test}$ (the strike proportion of the test is the same as than the baseline proportion)

H_a: $p_{base} > p_{test}$ (the strike proportion of the baseline is greater than the test proportion)

Our first batch of 10 sensors are deployed, and only 1 of them have strike events (10%). To satisfy our protocol, we carry out two more batches of 10 deployments get the minimum group sample size of n = 30. After three batches, the proportion of strikes in the test group is (1 + 3 + 2)/(10 + 10 + 10) = 0.2, or 20%. Looking at Table 1, we see that for a 20% test strike proportion we need n = 44. We continue to carry out two more batches, and end up with strike proportions from five batches in total as (1 + 3 + 2 + 3 + 4)/(10 + 10 + 10 + 10 + 10) = 0.26 or 26%. Looking to Table 1, we see that if we round up the test strike proportion from 26% to 30%, we would need 106 deployments. Considering the level of effort required, we decide to conclude the test with a total of 50 datasets (five batches of 10 deployments each).

The input code in R for the post-hoc test:

```
prop.test(x = c(50, 13), n = c(50, 50), alternative = "greater", cor = T, conf.level = 0.95)
```

The output information from this test in R:

```
data:  c(25, 13) out of c(50, 50)
X-squared = 5.1358, df = 1, p-value = 0.01172
alternative hypothesis: greater
95 percent confidence interval:
 0.0652788 1.0000000
sample estimates:
prop 1 prop 2
  0.50   0.26
```

Rule-based evaluation of the two-sided test results:

Because the p-value (0.01172) is less than our choice of α (0.05), we choose to reject the null hypothesis that the strike proportion of the test environment is equal to the assumed baseline strike proportion. The post-hoc power of this test is 0.80, which is lower than our target power of 0.85, but is still reasonably high enough to consider as a suitable test result.

Rule-based decision for this test:

There is sufficient evidence with 95% confidence to conclude that the baseline strike proportion of 50% is larger than the test group strike proportion of 26%.

Free and open sources resources

G*Power software for statistical power analyses (Mac and Windows)

https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

R project for statistical computing

https://www.r-project.org/