Experiment Design for Computer Sciences (0AL0400)

Topic 07 - Sample Size Calculations

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

Version 2021.1 (Updated May 27, 2021)

Part I – Introductio and Motivation

Lecture Outline

In this course we talked several times about the necessity of collecting **a sample**: A set of (multiple) observations that are used to calculate the value of interest.

However, up until now we have avoided talking about **how big** this sample should be. This is the topic of today's lecture.

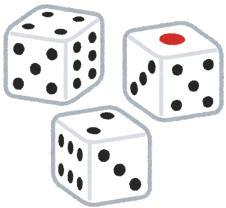
- What is sample size?
- Why do we need to worry about it?
- What factors influence the choice of sample size?
- How to calculate the desired sample size?

Why take samples?

Review: Noise Factors

As we discussed before, the result of an experiment is affected by several factors, some of which are unknown, or difficult to control.

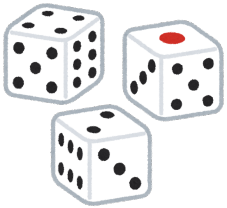
Because of these **Noise Factors**, sometimes there will be a small variance in the result of repeated experiments. To measure this variance, and take it into account for our analysis, we repeat the experiment several times, and gather those repetitions into a **sample**.



For example: The true mean of throwing three dice and adding their values is **10.5**. However, every time we throw the dice, the result will be a bit different.

Why take samples?

Sample size and estimation error



If we want to estimate the mean value of a noisy process (for example, the mean of three dice), we can observe the process multiple times (a sample), and take the average of the sample values.

As the sample size gets larger, the sample error usually gets smaller. **However, the noise of the original process does not change!**

```
> sample_2 <- replicate(2, sum(sample(6,3)))
9 6
> sample_5 <- replicate(5, sum(sample(6,3)))
11 10  9 11 12
> sample_10 <- replicate(10, sum(sample(6,3)))
9 11 12 11  7  9 14 11 13 10
> mean(sample_2)          #    7.5       > sd(sample_2)          # 2.12
> mean(sample_5)          #   10.6       > sd(sample_5)          # 1.14
> mean(sample_10)         #   10.7       > sd(sample_10)         # 2.05
```

Why take samples?

Larger sample sizes help us isolate the error related to the **noise of the process**. This has an influence on:

- Size of the confidence interval of the estimator;
- Confidence of statistical tests;
- Power of statistical tests;

This is why we are interested in having "big enough" sample sizes.

Is "as much as possible" the answer for sample size?

From what we talked until now, it is natural to think that "more observations is always better". In general, it is a good idea to have a large sample size, but there are other things that need to be considered:

- Many experiments have a cost associated with obtaining each observation. So large sample sizes result in large experimental cost.
- Some experiments require materials or conditions that are hard to obtain (for example, interviews);

The increase in sample size has diminishing returns in terms of confidence and power of tests. So, sometimes the cost of collecting more observations is higher than the benefit of those observations.

So, what is the appropriate sample size for a given experiment?

What is a good sample size?

For the choice of sample size, you usually take three things in consideration:

- The cost of the experiment;
- The desired confidence (α, Confidence interval size);
- The desired power (β)

Calculating Sample Size: When you don't have a choice

When an experiment is constrained by budget (not enough time, not enough money, etc), we might not have a choice of sample size.

But even if the sample size is fixed by the experimental conditions, it is still important to calculate the power of the experiment. The result of power calculation tells us how we should interpret the results of the experiment.

To calculate the power of the experiment, we need to first define the minimum interesting effect size δ^*.

Example of power calculation

Consider an one-sample experiment with the following parameters: Alternate hypothesis is one-sided, sample size is 10, $\delta^* = 0.5$, estimate of standard deviation is $\sigma = 1$, desired significance $\alpha = 0.01$.

What is the power of this experiment?

```
> power.t.test(n = 10, sd = 1, sig.level = 0.01, type = "one.sample",
+              alternative = "one.sided", delta = 0.5)

  One-sample t test power calculation

           n = 10
       delta = 0.5
          sd = 1
   sig.level = 0.01
       power = 0.1654013          <- Power = 0.16 - Very low!
 alternative = one.sided             High chance for false negatives
                                     *For this effect size*
```

Example of power calculation

Increasing the power by increasing δ^*

If we want to increase the power of the experiment, but we cannot increase the number of observations, we could increase the size of the "minimal difference" that this test detects. This calculation will tell us what is the minimum difference that our test will observe.

```
> power.t.test(n = 10, sd = 1, sig.level = 0.01, type = "one.sample",
+              alternative = "one.sided", power = 0.8)

   One-sample t test power calculation

            n = 10
        delta = 1.185308          <- When we fix the power and the sample size,
           sd = 1                    the calculation tells us that the minimum
    sig.level = 0.01                 effect size becomes "1.18". If the real
        power = 0.8                  effect is less than this, the probability
  alternative = one.sided            of a false negative increases.
```

Power calculations

Some considerations

Using power calculations, we can obtain the estimation of the Type-II error probability. This gives us a better understanding of the ability of an experiment to detect effects of interest.

The statistical test will have lower power for differences smaller than δ^*, but these differences are below the minimally interesting effect; any effect greater than δ^* will result in a higher power for the test;

This technique is most useful to compute the required sample size for the experiment.

Part II – Sample Size Calculation

"We repeat each experiment 30 times"

We see many papers use this sample size without much explanation.

Why 30 repetitions? Is it a good value?

Why 30 repetitions?

Remember that the **Central Limit Theorem** (CLT) states that the distribution of the "sample mean" estimator becomes closer to normal, as the sample n size increases.

When $n > 30$, the CLT holds (the sample mean follows a normal distribution) for most cases (except very extreme cases of non-normality). This is important, because many test statistics require the assumption of normality.

However! Other than (potentially) guaranteeing that the estimator is a normally distributed random variable, there is nothing special about using 30 repetitions.

In particular, $n = 30$ does not guarantee anything about the power or confidence of the experiment, so you still need to perform power calculations! (And you also need to guarantee the other assumptions of the experiment.)

Sample Size Calculation

One Sample Testing

The sample size calculation in related to the power calculation. If we fix the value of power and δ^*, the calculation will give us the minimum sample size:

```
> power.t.test(sd = 1, sig.level = 0.01, type = "one.sample",
               alternative = "one.sided",   power = 0.85,     delta = 0.5)

  One-sample t test power calculation
           n = 47.98044   # <-- Round this value up!
       delta = 0.5
          sd = 1
   sig.level = 0.01
       power = 0.85
 alternative = one.sided
```

We need at least 48 observations to detect a one-sided effect of $\delta^* = 0.5$ or more on the mean with a power level of 0.85.

Sample Size Calculation

Example for Two Means

Let's consider a more general example, where we are comparing two means with the following experimental characteristics:

- Desired significante $\alpha = 0.05$
- Desired power: $(1 - \beta) = 0.8$;
- Minimally relevant effect size (MRES): $\delta^* = 15$
- Variances of the samples: $\sigma_1, \sigma_2 =$?

From these specifications, we can obtain the required sample sizes.

Sample Size Calculation

Two means with equal variances

For the specific case of approximately equal variances, the optimal sample size ratio is $n_1 = n_2 = n$, with:

$$n \approxeq 2 \left(\frac{t_{\alpha/2}^{(2n-2)} + t_{\beta}^{(2n-2)}}{d^*} \right)^2$$

where $d^* = \delta^*/\sigma$ is the (standardized) minimally interesting effect size; and $t_{\alpha/2}^{(2n-2)}$ and $t_{\beta}^{(2n-2)}$ are the $\alpha/2$ and β quantiles of the $t^{(2n-2)}$ distribution.

Sample Size Calculation

Calculation Example

Back to the original example, with an estimation of standard deviation $= 15$, we would calculate the sample size as:

```
> ss.calc <- power.t.test(delta = 15,  sd = 15,
                          sig.level = 0.05,  power = 0.8,
                          type = "two.sample", alternative = "one.sided")

  Two-sample t test power calculation
           n = 13.09777       <- NOTE: n is the size of *EACH* sample.
       delta = 15
          sd = 15
   sig.level = 0.05
       power = 0.8
 alternative = one.sided
```

Sample Size Calculation

Obtaining the variance estimate

The Power calculation formulas and functions are convenient, but leave us with a problem: we need a variance estimate to calculate the sample size, but we need observations to estimate the variance!

There are a few ways to proceed in this case. The most practical are:

- Use process knowledge or historical data to obtain an (initial) estimate of the variance;
- Use a standardized MRES to calculate sample size;
- Perform a pilot study and collect samples to estimate the variance.

Each approach has its own advantages and drawbacks.

Comparison of two means – Paired design

Paired designs can require smaller sample sizes for equivalent power in cases where the **between-units variation** is relatively high, and the **in-unit variation** is relatively homogeneous.

More specifically, if the within-level variation is given by σ_ϵ and the between-units variation is σ_u, we have that, for large enough n
(e.g., $n \geq 10$),

$$\frac{n_{\text{unpaired}}}{n_{\text{paired}}} \approx \sqrt{2\left[\left(\frac{\sigma_u}{\sigma_\epsilon}\right)^2 + 1\right]}$$

Sample size formulas for ANOVA

If one is interested in calculating the required sample size for the ANOVA procedure (without worrying about the eventual post-hoc testing), the formulas are almost as simple as those used for the t tests.

Essentially, the power/sample size calculations for the ANOVA boil down to the equality:

$$F_{(1-\alpha)} = F_{\beta;\phi}$$

with both F distributions having $(a-1)$ degrees of freedom in the numerator and $a(n-1)$ in the denominator. The noncentrality parameter ϕ is given by:

$$\phi = \frac{n \sum_{i=1}^{a} \tau_i^2}{\hat{\sigma}^2}$$

Sample size formulas for ANOVA

To ilustrate the sample size calculation procedure, imagine an experimental design with $a = 4,\ \alpha = 0.05,\ \hat{\sigma} = 7$, and suppose that the researcher wants to be able to detect whether any two means present differences of magnitude $\delta^* = 12$ with power $(1 - \beta) = 0.8$.

Under these conditions, two scenarios tend to be of interest: the first is if we have two levels biased symmetrically about the grand mean, and all the others equal to zero:

$$\tau = \left\{ -\frac{\delta^*}{2}, \frac{\delta^*}{2}, 0, 0 \right\}$$

and the second is if we have one level biased in relation to all others:

$$\tau = \left\{ -\frac{(a-1)\delta^*}{a}, \frac{\delta^*}{a}, \frac{\delta^*}{a}, \frac{\delta^*}{a} \right\}$$

Sample size formulas for ANOVA

For the first case we have a noncentrality parameter of:

$$\phi = \frac{4\left(6^2 + 6^2 + 0 + 0\right)}{7^2} = 5.88$$

And we can calculate it as:

```
> a       <- 4                          > alpha   <- 0.05
> sigma   <- 7                          > delta   <- 12
> beta    <- 0.2
> tau <- c(-delta/2, delta/2, rep(0, a - 2))
> vartau <- var(tau)

> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 8.463358
```

Sample size formulas for ANOVA

The second case (one level biased in relation to all others) is also quite easy to calculate:

```
> tau <- c(-delta*(a - 1)/a, rep(delta/a, a - 1))
> vartau <- var(tau)
> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 6.018937
```

It is important to remember that these are the sample sizes required for the ANOVA only - any multiple comparisons procedure executed afterwards to pinpoint the significant differences will have smaller power for same-sized effects (unless more observations are added). This is one reason why it is common to design experiments calculating the sample sizes based on the multiple comparisons procedure, instead of using the ANOVA formulas.

More on sample size calculation for Computer Science experiments

These formulas and concepts only scratch the surface of the problem of sample size calculation.

By understanding the characteristics of the populations under study, we can identify a minimum sample size that gives us a test with desired confidence and power.

A more recent discussion of the calculation of sample sizes for the specific case of algorithm comparison is the paper by Felipe Campelo:
`https://link.springer.com/article/10.1007/s10732-018-9396-7`

I highly recommend reading this paper as a complement to this lecture.

Recommended Reads

- Felipe Campelo *"Sample size estimation for power and accuracy in the experimental comparison of algorithms"*, 2019
  `https://link.springer.com/article/10.1007/s10732-018-9396-7`
- Paul Mathews' *Sample Size Calculations*, MMB, 2010.
- Zhang (2003), J. Biopharm. Stat. 13(3):529-538.

About these Slides

These slides were made by Claus Aranha, 2021. You are welcome to copy, re-use and modify this material.

These slides are a modification of "Design and Analysis of Experiments (2018)" by Felipe Campelo, used with permission.

Individual images in some slides might have been made by other authors. Please see the following references for those cases.

Image Credits I