Experiment Design for Computer Sciences (0AL0400)

Topic 06 - Sample Size Calculations

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

Version 2023.1 (Updated June 1, 2023)

Part I – Introduction and Motivation

Outline

A important point to decide when preparing an experiment is **how many observations** are necessary. We call this number the **Sample Size**.

A large sample size can reduce the influence of random noise in the calculation of statistical indicators, such as the mean.

On the other hand, increasing the sample size also has a cost, in time, money or resources.

So how do we choose a proper sample size for an experiment?

Outline

In this lecture, we will discuss the following points:

- How to think about the sample size;
- What factors influence the choice of sample size;
- What factors are influenced by the choice of sample size;
- How to calculate the desired sample size for certain statistical models;

Why Repeat an experiment?

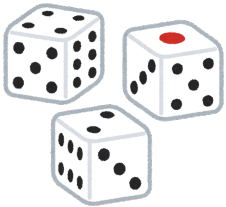
Noise Factors

The result of almost any experiment is affected by several factors. Some of these factors we can control and understand, **but many we cannot**.

Noise Factors are these uncontrollable, unknown factors. As much as possible, we want to reduce the influence of these factors in the result of our experiment.

Important: In most cases, random factors actually have a cause, we just don't know it. If you manage to identify the cause for some of the random noise, consider if it would be possible to remove that noise from your experiment.

Repeating an experiment to remove noise factors



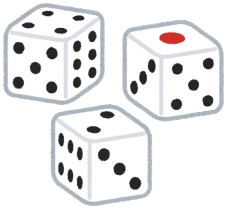
For example: When we throw a dice, the throw is affected by the air resistance, by small changes in our hand movements, by small imperfections in the dice, etc. All of these will affect the final result of the dice;

Some assumptions about random noise:

- The variance of random noise is smaller than our minimum interesting difference;
- Random noise is unbiased;
- Random noise is normally or uniformly distributed;

You should consider if these assumptions hold true, and maybe check them by analyzing the result of a **trial experiment**.

Example: Sample size and Sample Mean Error



Imagine the following simple experiment: We throw three dice, and take their sum. We want to estimate the *mean value* of this process.

So we repeat the process several times to reduce the effect of random noise. What happens when the sample size increase?

The mean and the standard error of the process (10.5 and 2.1), in general, will not change much. However, the standard error of the **sample mean** will change a lot!

```
# Three experiments with different sample sizes
sample_2_10 <- replicate(10, mean(replicate(2, sum(sample(6,3)))))
sample_20_10 <- replicate(10, mean(replicate(20, sum(sample(6,3)))))
sample_200_10 <- replicate(10, mean(replicate(200, sum(sample(6,3)))))
# The sample mean is about the same, but the SD of the sample mean changes!
sample size   2: mean of 10 samples: 11.50  sd of 10 samples: 1.90
sample size  20: mean of 10 samples: 10.48  sd of 10 samples: 0.37
sample size 200: mean of 10 samples: 10.56  sd of 10 samples: 0.20
```

Why should we use large sample sizes?

Larger sample sizes will reduce (average) the effect of random noise, and thus reduce the variance of the indicator.

This will have an effect on:

- Reduce the size of the confidence interval of the indicator.
- Reduce the Type-I error (increase the confidence) of statistical tests.
- Reduce the Type-II error (increase the power) of statistical tests.

Why should we use SMALL sample sizes?

Right now, you might be thinking: “I always want a sample size as big as possible!”

But there are other things to be considered. Experiments usually have a **cost** associated with them. Large sample sizes may result in large experimental cost:

- Money;
- Time;
- Resources;
- Special conditions;
- Ethical considerations;

Also, the increase of sample size has diminishing returns in terms of reducing experimental error. So we want to use the smallest possible sample size that satisfy our requirements.

What is a good sample size?

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

- The costs of the experiment; (hard limit to sample size)
- The desired confidence level (Probability of Type-I error, α);
- The desired experimental power (Probability of Type-II error, β);

When you can't choose the sample size, calculate the power

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

In this case, **It is important to calculate the confidence and power of the experiment**. The power of the experiment will let us know how reliable the result is.

For example, if the power is too low with the available budget, this information can be used as a justification to require more budget for the experiment.

Example of Power Calculation

To calculate the power of the experiment, we need the following variables:

- Sample size;
- Standard deviation;
- Significance Level;
- Minimum Interesting effect (Delta: δ^*)

The last one is important. δ^* has a large effect on the experiment power. In fact, we usually say something like "The experiment is powerful enough to detect a difference of at least δ^*".

Example Calculating the Power with R

Let's calculate the power of the dice experiment.

Consider a sample size of 20, our observed standard deviation of 2.1. Can we detect a change in the mean of at least $\delta^* = 0.5$, with 95% confidence?

```
> power.t.test(n = 20, sd = 2.1, sig.level = 0.05,
  + type = "one.sample", alternative = "two.sided", delta = 0.5)

     One-sample t test power calculation
              n = 20
          delta = 0.5
             sd = 2.1
      sig.level = 0.05
          power = 0.171485                    <-- Very low power
    alternative = two.sided
```

It is very likely that our experiment will not detect this difference, even if the difference exists!. Also note that no data was needed!

Example Calculating Experiment Power – changing δ^*

Let's repeat the calculation, but change δ^* to "1.5".
As you can see, the power of the experiment is now 0.85. This means that the experiment with 20 samples is sensitive enough to detect a difference of 1.5 in the mean, with 95% confidence.

```
> power.t.test(n = 20, sd = 2.1, sig.level = 0.05,
  + type = "one.sample", alternative = "two.sided",
  + delta = 1.5)                                  <-- Changing Delta!

     One-sample t test power calculation
              n = 20
          delta = 0.5
             sd = 2.1
      sig.level = 0.05
          power = 0.85755                         <-- Power increased!
    alternative = two.sided
```

Example Calculating Experiment Power – changing the sample size

What if we really want to detect a difference of 0.5, with power 0.85?

What is the necessary sample size?

```
> power.t.test(power = 0.85,           <-- Fixed power, sample size not given
  + sd = 2.1, sig.level = 0.05, type = "one.sample",
  + alternative = "two.sided", delta = 0.5)

     One-sample t test power calculation

              n = 160.311      <-- Necessary Sample Size
          delta = 0.5
             sd = 2.1
      sig.level = 0.05
          power = 0.85
    alternative = two.sided
```

Power Calculations: In Summary

The power calculation can be used to estimate one of the three below, given the other two:

- The sample size (n);
- The minimum interesting effect (δ^*)
- The experiment power (β);

Note that for the power calculation, **You don't need the data!**[1]. So It is possible to calculate the sample size before you run the experiment.

[1] except for an estimation of the SD

What about the magic number 30?

A lot of people have an instinct to use 30 as the “default” sample size.

This comes from earlier studies that showed that, for many distributions, $n = 30$ is enough for the **CLT** to apply, and for the **sample mean** to be roughly normally distributed.

This is a very important result! It helps us not worry about the assumption of normality. However, other than that, there is nothing special about $n = 30$

You should always calculate the power of your experiment. Sometimes you can use a smaller sample size. Sometimes you need a much bigger sample size.

Also, it is important to know the power level of your experiment.

What about the Variance?

The Power calculation function is very convenient, but it has a problem:

- To calculate the power / sample size, we need to input the standard deviation;
- To estimate the standard deviation, we need to do an experiment!

This is a bit of a “chicken and egg” problem. There are a few ways to solve this problem:

- Use knowledge about the problem domain, or historical data, to obtain an (initial) estimate;
- Perform a pilot study to collect a sample only for the SD estimation.

What about the Variance? – Pilot Study

A **Pilot Study** is a small, preliminary experiment. It is used to obtain a few estimates about the model under study:

- mean and standard deviation;
- possible sources of noise;
- possible difficulties in the experiment;

One way to calculate the pilot study sample size is:

$$n_{pilot} \approx 2 \left(\frac{z_{\alpha_n/2}}{e_n} \right)^2$$

where $(1 - \alpha_n)$ is the desired confidence level for the sample size estimate of the main study, and e_n is the maximum relative error allowed for the sample size.

This calculation can yield some scarily large sample sizes for a pilot study (much larger than would be actually required for the main study itself), so use this with caution.

Sample Size Calculations for Specific Models

Sample Size Calculation for several models

Until now, we saw how to calculate the power / sample size for the “1 sample” model, where we compare one sample against a fixed value.

As you may imagine, this calculation will change slighly as the statistical model used for the experiment changes. Let’s see some of these considerations.

Sample size for Two Means

Consider the situation where we are comparing the means of two samples. For example, the typical experiment where we compare two algorithms (A and B) on a single experiment.

While the basic idea is the same, there is a key difference to consider here: Should the two sample sizes be the same, or different?

Sample Size Calculation

Example for Two Means

Consider a case 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: $\sigma_1, \sigma_2 =$?

What are the required sample sizes for this case?

Sample Size Calculation

Case 1: Two means, equal variances

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

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

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

Case 1: Two means, equal variances

Let's say that we estimate the standard deviation (for both samples) to be $sd = 15$.

We can now calculate the sample size as:

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

Case 2: Two means, unequal variances

When the variance is not the same in both samples, we can use:

- A balanced design ($n_1 = n_2$) or;
- An unbalanced design ($n_1 \neq n_2$)

The unbalanced design usually has a lower total number of observations. The optimal allocation of sample size for this kind of design is to have the ratio of observations to be equal to the ratio of variance:

$$\frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$$

.

(provided that a good estimate of the ratio of variances is available, of course)

Sample Size Calculation

Case 2: Two means, unequal variances – Example unbalanced design

```
> MESS::power_t_test(n=NULL, sd=15, delta=15,
+                    ratio=2, sd.ratio=2,
+                    power=0.8, sig.level = 0.05,
+                    type="two.sample", alternative="one.sided")

Two-sample t test power calc with unequal sample sizes and unequal varianc

             n = 19.01525, 38.03050 <-- bigger sample size
         delta = 15
            sd = 15, 30             <-- total var is lager
     sig.level = 0.05
         power = 0.8
   alternative = one.sided

NOTE: n is vector of number in each group
```

Sample Size Calculation

Case 2: Two means, unequal variances – Example balanced design

```
> MESS::power_t_test(n=NULL, sd=15, delta=15,
+                    ratio=1, sd.ratio=2,
+                    power=0.8, sig.level = 0.05,
+                    type="two.sample", alternative="one.sided")
     Two-sample t test power calculation with unequal variances

              n = 31.8629, 31.8629 <-- larger total observations
          delta = 15
             sd = 15, 30
      sig.level = 0.05
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

Sample Size Calculation

Case 2: Two means, unequal variances – Example power calculation

```
> MESS::power_t_test(n=20, sd=15, delta=15,
+                    ratio=1.5, sd.ratio=2,
+                    power=NULL, sig.level = 0.05,
+                    type="two.sample", alternative="one.sided")

Two-sample t test power calc with unequal sample sizes and unequal varianc

              n = 20, 30
          delta = 15
             sd = 15, 30
      sig.level = 0.05
          power = 0.7438965 <-- lower power
    alternative = one.sided

NOTE: n is vector of number in each group
```

Sample Size Calculation

Two means, unequal variances considerations

- We can use a balanced design, but the total number of observations will be higher.
- Using an unbalanced design will result in a smaller total sample size.
- We should calculate the power if the sample size is fixed.

Comparison of two means – Paired design

The analysis of an experiment with paired design can require smaller sample sizes for the same power.

This happens when the **between-units variation (σ_U)** is relatively high, and the **in-unit variation (σ_ϵ)** is relatively homogeneous.

For large enough n (e.g., $n \geq 10$), we have that:

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

So we can initially calculate the unpaired sample size, and adjust it for the paired case.

Comparison of two means – Paired design

Alternatively, we can also use good old *power_t_test*:

```
> MESS::power_t_test(n=NULL, sd=15, delta=15,
+                    ratio=1, sd.ratio=1,
+                    power=0.8, sig.level = 0.05,
+                    type="paired", alternative="one.sided")

     Paired t test power calculation

              n = 7.727622       <-- paired tests are very powerful!
          delta = 15
             sd = 15             <-- don't forget you have to show this!
      sig.level = 0.05
          power = 0.8
    alternative = one.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Sample Size Calculation for multiple means

The analysis of multiple means, as discussed in a previous class, is composed of two steps:

- Overall analysis using ANOVA;
- Posthoc analysis between pairs of samples;

So when calculating sample size, we can use the calculation for the first part, or for the second part. Usually the second part implies a larger sample size, but sometimes we want to know the power for the first part too.

Power calculation for the ANOVA

Scenarios of Interest

The power of a test is the probability of the test not being able to detect the alternate hypothesis (type II error). So to calculate the power of ANOVA, we need to establish a **scenario of interest**.

For example, if we are comparing 4 samples, 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:

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

Power calculation for the ANOVA

General Formula

Given the scenario of interest, the power/sample size calculations 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:

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

Power calculation for the ANOVA

Example

To ilustrate, 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 $(1 - \beta) = 0.8$.

For the first scenario of interest we have a noncentrality parameter of:

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

Power calculation for the ANOVA

Example

And we can calculate the power/sample size of the first scenaro as:

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

Power calculation for the ANOVA

Example

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, 2022. 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