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

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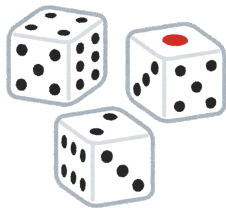
For example: When we throw a dice, the throw is affected by the air resistance, by small changes in our hand movements, by small imperfections in the dice, etc. All of these will affect the final result of the dice;

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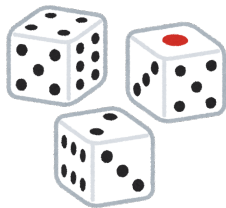
Imagine the following simple experiment: We throw three dice, and take their sum. We want to estimate the *mean value* of this process.

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
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:

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

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
```

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:

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

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, 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