Experiment Design for Computer Sciences (01CH740)

Topic 07 - Experiment Power and Sample Size

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

June 18th, 2020

Version 2020.1 (Updated June 29, 2020)

Class Schedule Reminder

The schedule for the final activities of this course are as follows:

- Today: Sample Size Calculation;
- 7/03 : Q&A Session 2;
- 7/20 : Report 3 Deadline;

Outline of the Lecture

Topic: Calculation of Sample Sizes

- Review of type II error;
- Approaches for choosing sample sizes;
- Sample size calculation for different tests;
  - Single sample
  - Comparison of Two samples
  - Paired Comparison
  - Equality Testing
  - ANOVA
- Other situations;

How do we choose sample size?

During this course, we talked many times about the importance of taking repeated observations of an experiment. The reasons for this included:

- Obtaining more precise point estimators for population parameters;
- Calculating confidence intervals for parameter estimators;
- Increasing the confidence or the power of a test;
- Reducing variance, detecting outliers, etc;

But until now, we left open the question: "How many observations do we need?"

Is "the more the merrier" the answer?

From the previous slide, we could think that "the more observations the better", but that is not quite right.

- As we saw in the first class on hypothesis testing, it is possible to reduce p to arbitrary values by largely increasing n.
- On the other hand, increasing the number of observations can be costly (financial costs, time costs, human costs, etc)
- The choice of n influences how the data will be selected and analyzed, so it is helpful to know this value during experiment design;

Because of this, we are interested in a more formal way to define the number of observations necessary in an experiment;

"We repeated each experiment 30 times"

In Computer Science papers, it is common to see papers stating that they repeated computational experiments "30 times" (or sometimes 20). Where does this rule of thumb comes from?

The Central Limit Theorem (CLT) states that the distribution of sample means tends to follow a normal distribution. This effect can be observed even on means of much smaller samples for well-behaved distributions. However, in general the CLT will hold for $n > 30$, except for extreme cases of non-normality;

As many testing procedures require the assumption of normality from the underlying distribution, an "oral tradition" of using $n = 30$ began in some CS groups. But it is much better to try to actively identify the correct sampling size for your experiment.

When is sample size 30 not appropriate?

- The underlying distribution of the population is very well behaved;
- Your planned analysis does not require assumption of normality;
- You are comparing samples with very different variants;
- Your budget does not allow for sample size equal to 30;
- Your experiment is dangerous (example, drug tests), and you want to minimize the number of experiments;

All of these cases are strong reasons for a more explicit sample size calculation;

Experiment Power

Budget Constrained case

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

However, even in this cases the calculation of sample size is important, because it can be used to give us information about **experiment power**.

The power of an experiment is an expression of the probability of **Type II errors** (False Negatives). It tells us **how sensitive our experiment is to actually find the effect that we are looking for**.

If our experiment is budget-constrained, a power analysis may tell us if the experiment is not useful at all, or if we need to settle for a larger minimum effect.

Sample size and Type-II error

The probability of Type-II error can be easily (and often wrongly) evaluated *a posteriori*, but its definition *a priori* requires some care;

Given a desired test, its power is essentially a function of 4 elements:

- Actual size of the difference;
- Variability of the observations;
- Significance level;
- Sample size.

The experimenter generally has very little control over the first two. But they can be estimated.

Sample size and Type-II error

Estimating Experimental Power

A strategy for estimating an effective lower bound for the power of a test includes a definition of an *minimally interesting effect* δ^*.

This value must be derived from technical and scientific knowledge about the phenomenon or system under experimentation.

It is essential to have a good understanding of the field in which the experiment will be conducted.

Once δ^* is defined, the experimenter can obtain an estimate of the variability of observations (e.g., by a pilot study), which can then be used to obtain an approximate power value for the experiment;

Sample size and Type-II error

Some considerations

Having obtained this estimation of the Type-II error probability, one can run his/her experiment with a better understanding of its ability to detect effects of interest.

The 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.

Sample size and Type-II error

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, delta = 0.5, sd = 1, sig.level = 0.01,
+         type = "one.sample", alternative = "one.sided")

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*
```

Sample size and Type-II error

Example of sample size calculation

With only $n = 10$ samples, the experiment is **underpowered**. What is the smallest sample size needed to obtain a desired power of 0.85?

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

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 deviation of 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.

Can we do this calculation? Or is there something missing?

Sample Size Calculation

How to obtain the variance estimate

These formulas are very convenient, but leave us with a riddle: we need variance estimate in order to calculate the sample size, but we need observations to be able 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.

Sample Size Calculation

Pilot Study

If no information is available to estimate the variance, a pilot study must be performed to obtain this value. The sample size required for this pilot study is given by:

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

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.

Calculation of Sample Sizes

Case of Known Standard Deviation

Suppose that the engineer uses data available from the system manuals, as well as historical measurements, to estimate a reasonable upper bound for the common standard deviation as $\sigma \approxeq 15$.

Assuming that equal sample sizes are desired, we can simply use the formula:

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

Easy, right?

Calculation of Sample Sizes

Case of Known Standard Deviation

The last problem we have to solve is that the values of $t_{\alpha/2}^{(2n-2)}$ and $t_{\beta}^{(2n-2)}$ are also dependent of n, which makes the sample size equation transcendental in n.

We can solve that by using an initial estimate of $t_{\kappa}^{(2n-2)} \approx z_{\kappa}$, and iterating until we find the smallest n that satisfies:

$$n \geq 2 \left(\frac{\hat{\sigma}}{\delta^*} \right)^2 \left(t_{\alpha/2} + t_{\beta} \right)^2$$

Calculation of Sample Sizes

Case of Known Standard Deviation

In practice, there are many sample size calculators in statistical software. But it is important to known the idea behind the calculation, for when we find ourselves in special situations.

```
> 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 number in *each* group
delta = 15
sd = 15
sig.level = 0.05
power = 0.8
alternative = one.sided
```

Case of Two Means – Unequal Variance

The two-sample Welch t-test for considering unequal variances is usually the first test of choice, since it drops one (often inconvenient) assumption, at a very small cost in terms of power.

Calculating sample sizes for the general case (unequal variances, unequal sample sizes) is not particularly difficult, and can be done for either a *balanced* case (i.e., $n_1 = n_2 = n$) or an optimal, *unbalanced* case (in which $n_1 \neq n_2$).

For the unbalanced case, it is not particularly difficult to prove that the optimal allocation of observation is to keep:

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

.
(if a good estimate of the ratio of variances is available, of course).

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 for Equivalence of a single mean

Sample sizes for testing equivalence of a single mean can be derived using essentially the same considerations used for the usual tests. In the case of a single sample:

$$n \geq \left(\frac{(t_\alpha + t_\beta)\,\hat{\sigma}}{\delta^* - \Delta\mu} \right)^2$$

As in the previous cases, iteration is needed to solve for n (since the quantiles of the t distribution depend on n). Use $t_x = z_x$ for the first iteration.

Sample size for Equivalence of two means

Sample size for the $n_1 = n_2 = n$ case can be approximated based on the Zhang formula.

$$n \geq \left(t_{\alpha;\nu} + t_{(1-c)\beta;\nu}\right)^2 \left(\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\delta^* - \Delta\mu^*}\right)^2$$

with $\Delta\mu^* < \delta^*$ as the maximum real difference between the two means for which a power of $(1 - \beta)$ is desired, and:

$$c = \frac{1}{2} \exp\left(-7.06 \frac{\Delta\mu^*}{\delta^*}\right)$$

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:

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

Which allows us to calculate the required sample size by iterating on n until:

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

Sample size formulas for ANOVA

Doing it manually:

```
> a       <- 4                                  > alpha   <- 0.05
> sigma   <- 7                                  > delta   <- 12
> beta    <- 0.2
>
> tau <- c(-delta/2, delta/2, rep(0, a - 2)) # define tau vector
> n   <- 2
> while (qf(1 - alpha, a - 1, a*(n - 1)) >
+         qf(beta, a - 1, a*(n - 1), n*sum(tau^2)/sigma^2))
+     n <- n + 1
> print(n)
[1] 9
```

Using `power.anova.test()`:

```
> 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 manually, but lets keep it simple:

```
> 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, 2020. 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 references in each slide for those cases.

Image Credits I