HPRB 8900, Guest Lecture — March 27, 2004

Power Analysis and Sample Size Determination

Power analysis and sample size determination are aspects of the **design** of experiments and other research studies in which data are collected.

- Typically, a random sample is drawn which is believed to be representative of some population of interest.

- (Random) **variables** (characteristics) are measured on the sample members, with the goal of learning about the **distribution** of those variables in the population of interest.

While we'd like to know everything we can about the population distribution, we are typically most interested in the population **mean,** and possibly some other **parameters** (fixed, unknown constants) of that distribution such as the **variance** (or s.d.), median, range, etc.

- Population parameters may be estimated from the sample, but these estimates will almost certainly be wrong.

Statistical inference is concerned with

(1) quantifying the error in our estimates (forming **confidence intervals**), and

(2) answering questions about the values of the parameters (performing **statistical hypothesis tests**).

(1) and (2) are really flip-sides of the same coin, and power analysis/sample size determination can be approached from either perspective. Initially, we will concentrate on hypothesis testing.

Background on Hypothesis Tests:

Statistical hypothesis testing is set up very similarly to the American criminal justice system.

- The investigator trying to prove a scientific proposition is like a prosecutor trying to establish guilt.

- In the same way that a defendant is assumed innocent until proven guilty, the scientific proposition is assumed false until "proven" true.

- The threshold of proof (conviction) in the courts is "guilt beyond a reasonable doubt". In statistics, the threshold is the α-level, the probability of convicting an innocent man, which we are free to choose, but which is often take as $\alpha = .05$, or $\alpha = .01$.

The state of innocence, or the contrary of the scientific proposition under investigation, is assumed true. This is called the **null hypothesis** and is written as H_0.

The **alternative hypothesis** is what we're typically trying to establish. This is usuall written as H_A (or sometimes H_1).

- Classically, H_0 and H_A are always mutually exclusive. That is, H_A is always of the form

  $$H_A : H_0 \text{ is not true,}$$

  so that rejecting H_0 leaves us no alternative but to conclude that H_A is true.

- It is important to realize that, just as we can never prove innocence, we can never establish the truth of H_0. Thus, we never talk of accepting H_0. Instead we say things like "there is insufficient evidence to reject H_0."

Error Types:

There are two types of errors that can be made in choosing between H_0 and H_A:

I. **Type I Error:** Reject H_0 when H_0 is true.

II. **Type II Error:** Fail to reject H_0 when H_0 is false.

- The probability of I is called α, and it is usually fixed at $\alpha = .05$ or $\alpha = .01$ by the investigator.

- The probability of II is called β and it cannot be fixed (set) by the investigator.

  - Note that the **power** of a hypothesis test is

$$\begin{aligned} \text{power} &= \Pr(\text{reject } H_0 | H_0 \text{ is false}) \\ &= 1 - \beta, \end{aligned}$$

    or the probability of establishing our scientific proposition given that it is true.

- Want power to be high (want to have high probability of detecting the effect we're looking for) and β to be low.

  - However, power depends upon lots of things we can't control and a few we can.

Example — Two sample t test to compare means:

Suppose we have an active treatment (a new drug say) and a control treatment (placebo, say). Suppose we have a randomly sample of $2n$ subjects, n randomly assigned to active treatment, and n to control group.

Let μ_1, μ_2 be population means under the two treatments, and let σ_1, σ_2 be the corresponding population standard deviations.

- For simplicity, suppose the distribution of the response in the two populations is normal, with same s.d.: $\sigma_1 = \sigma_2 = \sigma$.

Picture:

Power depends upon:

(A) Difference in the means, $|\mu_1 - \mu_2|$. (Unknown)

- A big difference is easier to detect than a small one.

(B) σ. (Unknown)

- If there's lots of variability in the population, it will be hard to detect a difference in the means.

- For power/sample size calculations, (A) and (B) are usually combined into a single quantity called **effect size.**

(C) Sample size n. (Can be chosen, so known)

- Sample size is a measure of the amount of information available about the population. The more information we have, the more powerful our inferences about the population will be.

- Sometimes, sample size is fixed by constraints. Then question is how much power will we have for that sample size.

- More often, we fix power (we want to design a study with a certain level of power - 80%, say), calculate the power for each of several increasing values of n, and then choose the smallest n that gives us the power that has been chosen.

(D) α, the Type I error rate. (Can be chosen, but in practice is usually fixed by convention)

- High α values make it easier to reject H_0, which will certainly increase power, but at the expense of Type I error.

Typical Power Analysis for Two-sample t Test:

1. Fix $\alpha = .05$ (or .01, or some other value).

2. Assume a value for $\mu_1 - \mu_2$ and σ based upon estimated values from previous research (yours or from literature), if available, or take educated guesses at these values if not.

3. Select a desired level of power, 80%, say. This may be dictated by the funding agency, if power analysis is part of a grant proposal, or chosen by the investigator (represents risk of not being able to detect true effect, so how much risk are you comfortable with?).

4. Compute power for each of several values of n. Select smallest n that gives you power $\geq$ the level selected in 3.

Typically, step 2 is the hardest part. However, it can be made easier by noting that all that is really important in this step is the ratio

$$\frac{|\mu_1 - \mu_2|}{\sigma} = \text{effect size for } t \text{ test.}$$

The effect size measures the treatment effect relative to the variability in the population. It is a unitless quantity, with a scale for which there is some frame of reference, depending upon the general field of scientific investigation. Therefore, step 2 can be replaced by

2′. Choose an effect size.

An effect size can often be chosen based upon the general type of intervention under investigation and the expected magnitude of effect. (See table from Murphy and Myors, 1998.)

How is power calculated?

Remember, power is the probability of rejecting H_0 given that it is false. Whether or not H_0 is rejected is determined by a comparison of a test statistic with a critical value.

- Therefore, power computations are based upon the statistical test employed in the analysis!

In the two-sample t test situation, the null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0, \quad \text{vs.} \quad H_A : \mu_1 - \mu_2 \neq 0,$$

(assuming a two-tailed alternative) and the rejection rule is: reject H_0 if

$$t = \frac{|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2}(n_1 + n_2 - 2),$$

where

$\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$	are the sample means in groups 1 and 2
s_p	is the pooled sample standard deviation
n_1, n_2	are the sample sizes in the two treatment groups

- We can allow $n_1 \neq n_2$, but typically it is best to design **balanced** experiments, so we allocate equal sample sizes $n_1 = n_2 = n$ to the two groups. Thus, the rule becomes: reject H_0 if

$$t = \frac{|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|}{s_p\sqrt{\frac{2}{n}}} > t_{\alpha/2}(2n - 2).$$

Now that we have the test statistic, it is clear from the definition of power that

$$\text{power} = \Pr\left\{ \frac{|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|}{s_p\sqrt{\frac{2}{n}}} > t_{\alpha/2}(2n-2) \middle| \mu_1 - \mu_2 \neq 0 \right\}. \qquad (*)$$

- Now it is clear why power depends upon $|\mu_1 - \mu_2|$ (affects size of $|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|$), σ (affects size of s_p), n and α.

- Actually, note that $|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|$ and s_p only figure into the rejection rule through their ratio, which is the sample version of the effect size in this context:

$$\text{effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}.$$

How is the probability in (*) calculated?

Under the classical assumptions of a t test (normal samples, equal variances, independence), statistical distribution theory can be used to show that under H_A, the t statistic in (*) follows a **non-central t distribution.**

- This is a parametric distribution (like the normal distribution) with two parameters: the degrees of freedom ($2n-2$ here), and the non-centrality parameter (a function of the effect size).

Therefore, just like we use tables of the normal distribution to figure out the probability that a normal random variable exceeds some given value, we can figure out the probability in (*) from tables or computer programs that give non-central t probabilities.

Example — An Actual Power Analysis:

Suppose we intend to conduct a study to investigate whether a new drug reduces anemia in elderly women following hip fracture.

Response: Change in hematocrit over a two week period of treatment.

Design: A two-group, randomized, parallel, double-blind study. Patients will be randomized to two equal sized groups receiving the drug or placebo.

Statistical Analysis: Two-sample t test.

Previous Research to Determine Effect Size:

1. In a pilot study of 6 elderly women with hip fracture, the mean hematocrit was 32.3%. From the same institution, the mean hematocrit among 32 healthy elderly women was 33.5%.

2. Two previous studies tested different doses of the new drug in other patient groups. In those studies, the placebo group showed no change in hematocrit. The treated group showed changes of betwen 2.5 and 5.0%. The standard deviation in the change of hematocrit over a similar period of time was 2.0% in one of these studies.

- In the absence of such information, one might use a best guess of size of effect (small=.2, medium=.5, large=.8).

- In NQuery Advisor, you can specify effect size; or μ_1, μ_2, and σ; or $|\mu_1 - \mu_2|$ and σ.

- You can also specify power to generate n, or n to generate power.

Sample Size Determination for Confidence Intervals:

As mentioned before, hypothesis tests and interval estimation are really flip-sides of the same coin.

- E.g., for comparing two means μ_1 and μ_2, a 95% confidence interval for $\mu_1 - \mu_2$ is given by all of those values of $\mu_1 - \mu_2$ so that the test statistic

$$t = \frac{|(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\mu_1 - \mu_2)|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  would not be significant at the $\alpha = .05$ level.

Therefore, determining the sample size necessary to achieve a certain power in a hypothesis test, is very closely related to determining the sample size necessary to achieve a certain width (narrowness, really) for a confidence interval.

Example:

Suppose instead of wanting a certain power, I want to put a confidence interval around the difference between the mean change in hematocrit of the two groups: $\mu_1 - \mu_2$.

- Suppose I want that confidence interval to have a certain half-width. What sample size is necessary to achieve that precision?

  - Power and precision refer to the sample idea in the testing and interval estimation contexts, respectively.

- Alternatively, suppose I have a certain sample size. What will be the resulting width of the confidence interval?

Sample Size for Other Types of Designs/Statistical Analyses:

The two-sample design analyzed with a t test is among the simplest settings for a power analysis. Often, however, the design and statistical test will be more complex.

- E.g., Suppose we have g treatment groups to compare rather than just 2. This is a one-way layout (design), for which a one-way analysis of variance is appropriate.

The same principles underlying power analysis in the two-sample situation apply here as well, but the test statistic differs, and the issue of effect size is more complex.

In particular, for g groups, the probability of rejecting

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

depends not just on the difference between two means, but the difference between all possible pairs of means — i.e., on the spacing of the means.

In addition, the statistical test is now an F test, not a t test.

In fact, it can be shown that the t test is a special case of the F test corresponding to $g = 2$.

- So, power for a two sample comparison of means is based on the non-central t, and power for more general F tests based on the general linear model (e.g., the one-way anova model) is based on the non-central F distribution.

For the one-way ANOVA, the noncentrality parameter of the F distribution depends upon the spacing of the population means $\mu_1, \ldots, \mu_g$.

- So, power analysis for the one-way anova, can be done by specifying the values of all of the population means $\mu_1, \mu_2, \ldots, \mu_g$.

- Alternatively, one can take a conservative approach to power, and calculate the power assuming only a value for the difference

$$|\mu_{\max} - \mu_{\min}|$$

  together with the assumption that the rest of the means are spread out (spaced) in the least favorable (for rejecting H_0) configuration possible. This places all of the means except $\mu_{\min}$ and $\mu_{\max}$ together, halfway between $\mu_{\min}$ and $\mu_{\max}$:

- Power computed here will be less than or equal to the power under any other possible configuration of the means with the same value of $|\mu_{\max} - \mu_{\min}|$. Therefore, in practice, we can expect the study actually conducted to be more powerful.

Power Analysis for Other Types of Problems:

Power analysis for more complex designs/methods of analysis are available, but

1. additional assumptions/predictions about the data-generating mechanism are necessary;

2. the methods can be (much) more difficult to understand and to implement; and in some cases,

3. power analysis methods may not exist at all.

- Power analysis is easiest in the classical, normal-theory linear model for simple designs.

- Often, more complex designs can be simplified and thought of in terms of simpler designs for power/sample size purposes.

  - E.g., for a two-way layout with two factors with 2 and 3 levels, respectively, think of this as a one way layout with $2 \times 3 = 6$ treatments.

- Often, sample size is much more heavily determined by resource constraints and other practical matters than power. In many cases, the right sample size is "as many as you can afford".

  - In these cases, power analysis' role is to tell you whether the study is worth doing at all, or

  - to satisify a bureaucratic requirement (e.g., of a funding agency, or course instructor).