Power and Sample Size

STA 102: Introduction to Biostatistics

Yue Jiang

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Why calculate sample size and power?

- To show that under certain required conditions, a hypothesis test has a good chance of showing the anticipated difference, if it really exists
- To be more confident that a null result is not simply a sample of excessive variability
- To show a funding agency that the study has a reasonable chance of reaching a useful result
- To show that necessary resources (human, animal, financial, time, etc.) will be minimized

Example

Question: Does exercise affect body weight?

- Study design: participants will be randomized to two grups, exercise and control
- Outcome: change in weight from randomization to end of study
- Want to detect: weight change as small as one pound
- Known: SD of weight change is around 1.5 pounds, as given in exercise science literature
- Unknown: how many subjects do we need to show a difference, if one truly exists?

Study design

Proper study design depends on a variety of factors, including

- Measurement scale of variable of interest
- Hypothesis of interest
- Size of result that is clinically meaningful
- Cost of carrying out the study
- Time needed to carry out the study

Power

Showing adequate statistical power is usually necessary in order to get funding for research!

Power

Power is the probability of rejecting the null hypothesis when it is false (i.e., of avoiding a Type II error)

$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false})$$

and can be also thought of as the likelihood a planned study will detect a deviation from the null hypothesis if one really exists. Power is a function of

- Sample size n
- Deviation from the null one hopes to detect
- Standard deviation σ
- α, the Type I error rate

n, detectable difference, variance, and α

How do these four considerations affect power?

Power

When we design a study, it is not enough to know we have a small probability of rejecting H_0 when it is in fact true. We want to know we have a large probability of rejecting the null when it is false. Practically speaking, power less than 80% is typically considered insufficient to warrant a study.

Increasing power by tolerating more Type I errors is not acceptable. Therefore, we can increase power by

- Considering larger deviations from the null (need to think about clinical/practical importance)
- Increasing n (good, though not always affordable)

Interactive visualization

Let's examine a web visualization and see how this all works for a one-sample z-test.

What is this deviation?

When we calculate power, we need to know the minimum difference from the null mean μ_0 that we wish to detect. We may set up our hypotheses as

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

but need to specify a *minimum detectable difference*, often called $\delta = \mu_1 - \mu_0$ such that we reject H_0 with a certain power (usually 80% or 90%) when in fact $\mu = \mu_1$. We will need a bigger sample size when δ is small, and fewer subjects when δ is larger.

For example, in the weight change study, we would have $\delta = 1$ lb.

Sample size for two-sided one-sample test of mean

Power and sample size are interrelated. Ideally, when we plan a study, we have a prespecified idea of the minimum difference we want to detect, $\delta = \mu_1 - \mu_0$, the standard deviation, σ, and the power, $1 - \beta$, we'd like to have to detect it. In that case, it is simple to calculate the required sample size (here given for a one-sample test):

$$n = \left[\frac{\sigma \left(z^\star_{1-\alpha/2} + z^\star_{1-\beta} \right)}{\mu_1 - \mu_0} \right]^2$$

Sample size estimation based on CI width

Sometimes we're instead interested in estimating an effect with a given degree of precision. Suppose we wish to estimate the mean of a normal distribution with sample variance s^2 and require that the two-sided $100\% \times (1-\alpha)$ CI for μ be no wider than L.

How many subjects would we need to ensure that this is the case as a function of L, s^2, and α?

Hint: use what you know about how this confidence interval would be constructed.

Case study: ultra low dose contraception

Suppose you want to ensure a new manufacturer of birth control pills provides the correct dosage of 0.02 μg estrogen. How many pills do you need to sample in a shipment in order to ensure 80% power to detect a difference of 10% at $\alpha = 0.05$, assuming $\sigma = 0.008$? (10% of 0.02 is 0.002)

$$
\begin{aligned}
n &= \left[\frac{\sigma\left(z^\star_{1-\alpha/2} + z^\star_{1-\beta}\right)}{\mu_1 - \mu_0}\right]^2 \\
&= \left[\frac{0.008(1.96 + 0.84)}{0.002}\right]^2 \\
&= 125.44
\end{aligned}
$$

Fractional people

To be conservative, always round up for sample size calculations.

R uses a $t^\star$ instead of a $z^\star$ approximation. Do you expect the necessary sample size to be higher, lower, or the same?

Back to the hypothetical exercise study...

Question: Does exercise affect body weight?

- Study design: participants will be randomized to two grups, exercise and control
- Outcome: change in weight from randomization to end of study
- Want to detect: weight change as small as one pound
- Known: SD of weight change is around 1.5 pounds, as given in exercise science literature
- Unknown: how many subjects do we need to show a difference, if one truly exists?