

BIOS 600: Principles of Statistical Inference

Power and Sample Size

Fall 2012

Holiday Shopping Just Around the Corner...



Help a Starving Student!

Help! This enterprising SPH student has 2 years to finish her dissertation on attitudes about gun control and potential public health implications. She needs one year to finish collecting her data and one year for analysis and writing, and she thinks she can bag 50 subjects in a year. Are 50 subjects enough?



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

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 show a funding agency that the study has a reasonable chance of reaching a useful result

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 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?

Example

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control

Example

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control
- ▶ Outcome: change in weight from randomization to end of study

Example

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control
- ▶ Outcome: change in weight from randomization to end of study
- ▶ Want to detect: weight change as small as 1 pound

Example

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control
- ▶ Outcome: change in weight from randomization to end of study
- ▶ Want to detect: weight change as small as 1 pound
- ▶ Known: standard deviation of weight change around 1.5 pounds in the literature

Example

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control
- ▶ Outcome: change in weight from randomization to end of study
- ▶ Want to detect: weight change as small as 1 pound
- ▶ Known: standard deviation of weight change around 1.5 pounds in the literature
- ▶ Unknown: how many subjects needed to show a difference if it really exists?

Proper study design depends on a variety of factors, including

- ▶ Measurement scale of variable of interest

Proper study design depends on a variety of factors, including

- ▶ Measurement scale of variable of interest
- ▶ Hypothesis of interest

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

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

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 study

Power

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



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

$$\text{power} = \Pr(\text{reject } H_0 | H_0 \text{ is false})$$

and can also be 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 (larger n , higher power)
- ▶ Deviation from the null one hopes to detect (larger deviation, higher power)
- ▶ Standard deviation σ (smaller σ , higher power)
- ▶ α , the type I error rate (larger α , higher 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)

Let's return to the **applet** and see how this works for a 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 hypothesis as

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_A : \mu \neq \mu_0$$

but need to specify a *minimum detectable difference*, often called $\delta = \mu_A - \mu_0$, such that we reject H_0 with a certain power (usually 80 or 90%) when in fact $\mu = \mu_A$. 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 $\delta = 1\text{lb}$.

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 ($\delta = \mu_A - \mu_0$), 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)

$$n = \left[\frac{\sigma \left(z_{\frac{\alpha}{2}} + z_{\beta} \right)}{\mu_A - \mu_0} \right]^2$$

This can be easily derived from looking at the rejection regions when H_0 and H_A are true, though I'll spare you the details (unless you really want them!). You can also easily calculate power in Stata using the **sampsi** command with the **onesample** option. We'll play around with sample size a bit with our **applet**.

Sample Size Estimation Based on CI Width

Sometimes interest focuses instead on 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 2-sided $100\% \times (1 - \alpha)$ CI for μ be no wider than L . The number of subjects needed is approximately

$$n = \frac{4z_{\frac{\alpha}{2}}^2 s^2}{L^2}.$$

Sample Size Estimation Based on CI Width

This result follows from noting that the $100\% \times (1 - \alpha)$ CI for μ is $\bar{x} \pm t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ and thus the width of this interval is

$$\begin{aligned} 2 \times t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= L \\ 2 \times t_{n-1, \frac{\alpha}{2}} \frac{s}{L} &= \sqrt{n} \\ 4 \times t_{n-1, \frac{\alpha}{2}}^2 \frac{s^2}{L^2} &= n \end{aligned}$$

And the t is usually approximated by a z .

Case Study: Ultra Low Dose Contraception

Suppose you want to ensure the pills provide the correct dosage of $0.02 \mu\text{g}$ estrogen. How many pills do you need to sample in each 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.

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

```
. sampsi .02 .018, sd(.008) onesam power(.80)
```

Estimated sample size for one-sample comparison of mean to hypothesized value

Test Ho: m = .02, where m is the mean in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.8000
alternative m = .018
sd = .008
```

Estimated required sample size:

```
n = 126
```

Fathers with Heart Attacks

Suppose the mean fasting cholesterol of male teens in the US is $\mu = 175$ mg/dL with $\sigma = 50$ mg/dL. Investigators wish to know whether boys whose fathers had heart attacks have a different cholesterol level. What sample size would be needed to detect a difference of 5 mg/dL in mean fasting cholesterol with 80% power and $\alpha = 0.05$?

Fathers with Heart Attacks

Suppose the mean fasting cholesterol of male teens in the US is $\mu = 175$ mg/dL with $\sigma = 50$ mg/dL. Investigators wish to know whether boys whose fathers had heart attacks have a different cholesterol level. What sample size would be needed to detect a difference of 5 mg/dL in mean fasting cholesterol with 80% power and $\alpha = 0.05$ if we assume that 10% of the data collected will not be usable?

Fathers with Heart Attacks

Suppose the mean fasting cholesterol of male teens in the US is $\mu = 175$ mg/dL with $\sigma = 50$ mg/dL. Investigators wish to know whether boys whose fathers had heart attacks have a different cholesterol level. Investigators have \$5,000 in pilot funds and will need to pay \$100 for each expanded lipid profile. Assuming all data are complete, what is the minimum detectable difference in mean fasting cholesterol that investigators can afford on their budget with 80% power and $\alpha = 0.05$?

Fathers with Heart Attacks

Suppose the mean fasting cholesterol of male teens in the US is $\mu = 175$ mg/dL with $\sigma = 50$ mg/dL. Investigators wish to know whether boys whose fathers had heart attacks have a different cholesterol level. Assuming a fixed sample size of 1000 boys, what is their power for detecting a difference of 3 mg/dL in mean fasting cholesterol using $\alpha = 0.05$?

Exercise Study

- ▶ Question: does exercise affect body weight?
- ▶ Study Design: participants will be randomized to two groups, exercise (Insanity[®] videos) and control
- ▶ Outcome: change in weight from randomization to end of study
- ▶ Want to detect: weight change as small as 1 pound
- ▶ Known: standard deviation of weight change around 1.5 pounds in the literature
- ▶ Want type 1 error rate of 5% and power 80%.

“A recent study by the CDC ... found that Latinos ... outlive both whites and blacks, with a life expectancy of 80.6 years, compared with 77.7 for the nation as a whole. (People of Asian ancestry have even longer life spans, but because of their relatively high education and affluence levels, those findings are not considered surprising.) Latinos tend to be less educated than African Americans and their poverty rates are similar, yet Latinos outlive black people by nearly eight years.”

- ▶ H_0 : Latino culture is not advantageous to the population's longevity vs. H_1 : Latino culture is advantageous to the population's longevity
- ▶ Life expectancy at birth in US: 77.7 years
- ▶ Life expectancy is not normally distributed and has standard deviation approximately 4 years at birth
- ▶ How many subjects should be recruited if we want to detect a difference as small as 1 year between Latino and non-Latino residents of the US?

Formulate a Study

Suppose we wish to test the efficacy of the planned NYC ban on large ($>16\text{oz}$) sugar drinks, which will take effect on March 12, 2013. How could we design and carry out such a study?

Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 11
- ▶ North et al., “Examining a Comprehensive Model of Disaster-Related Posttraumatic Stress Disorder in Systematically Studied Survivors of 10 Disasters”