
Techniques for Sample Size

Russ Lenth
University of Iowa
russell-lenth@uiowa.edu

Software is available at
<http://www.stat.uiowa.edu/~rlenth/Power/>

Army Conference on Applied Statistics
October 22, 2004

Power basics

Power function

Given a test T of a parameter θ (scalar or vector)

$$\pi(\theta, n, \alpha, \phi) = \Pr(T \text{ is "significant"} \mid \theta, n, \alpha, \phi)$$

where α is the significance level, n is the sample size, and ϕ represents other parameters (e.g., σ^2)

Power basics

Sample-size determination

$$n(\theta^*) = \min\{n : \pi(\theta^*, n, \alpha, \phi) \geq \pi_0\}$$

with θ^* set at a clinically [scientifically] important value of θ .
Typically, people choose $\pi_0 = .8$, $\alpha = .05$.

Two-sample t test of significance

- * $\theta = \mu_T - \mu_C$, treatment vs. control
- * $n = \{n_T, n_C\}$ (often, constrain $n_T = n_C = n$)
- * $\phi = \{\sigma_T, \sigma_C\}$ (often, constrain $\sigma_T = \sigma_C = \sigma$)
- * Test statistic for $H_0^s : \mu_T = \mu_C$ vs. $H_1^s : \mu_T \neq \mu_C$:

$$U_s = \hat{\theta} / \widehat{SE}(\hat{\theta}) \sim t'(\nu, \theta / SE(\hat{\theta}))$$

with d.f. ν (may be estimated, approximate)

- * Power function:

$$\pi_s(\theta, n_T, n_C, \alpha, \sigma_T, \sigma_C) = \Pr(U_s < -t_{\alpha/2, \nu}) + \Pr(U_s > t_{\alpha/2, \nu})$$

Two-sample t test of equivalence

- * Same sampling situation
- * Let τ be a threshold for “smallness”
- * Test statistic for $H_0^e : |\mu_T - \mu_C| \geq \tau$ vs. $H_1^e : |\mu_T - \mu_C| < \tau$:

$$U_e = \frac{\min\{\hat{\theta} + \tau, \tau - \hat{\theta}\}}{\hat{S}\hat{E}(\hat{\theta})} = \frac{\tau - |\hat{\theta}|}{\hat{S}\hat{E}(\hat{\theta})}$$

- * Power function:

$$\pi_e(\theta, n_T, n_C, \alpha, \sigma_T, \sigma_C) = \Pr(U_e > t_{\alpha, \nu})$$

- * This is equivalent to two one-sided t tests of size α ; combined test is conservative.

Practical example

Strength of two materials

* Goals

- Want ability to detect a 15% difference ($\theta^* = \log_e 1.15 = 0.14$)
- A difference of less than 15% is negligible ($\tau = \log_e 1.15 = .14$)
- Tests with $\alpha = .05$, power goal of $\pi_0 = 80$

* Pilot data on $Y = \log_e$ strength: $\sigma \approx .20$ independent of mean.

* Using GUI. . .

- Sample size for each test
- Graphs
- Budget-based calculations

Just the FAQs

Most e-mail questions I get center on two issues

- * Sample size for a “medium” effect (per J. Cohen books)
- * Retrospective (observed) power

I have some opinions about these. . .

Retrospective power

Compute power based on...

- * Observed effect size
- * Observed SD(s)
- * Same sample size and significance level

Rationale: If result is nonsignificant, is it because...

- * Effect size is too small? ← high retrospective power
- * Sample size is too small? ← low retrospective power

Retrospective power

Compute power based on...

- * Observed effect size
- * Observed SD(s)
- * Same sample size and significance level

Rationale: If result is nonsignificant, is it because...

- * Effect size is too small?
- * Sample size is too small?
- * **Answer: The power is *always* small in this case (duh!)**

Retrospective power—another approach

Given. . .

- * Observed effect size
- * Observed SD(s)
- * Same sample size and significance level

Then the outcome of the test is also known

- * Recall that power = $\Pr\{\text{Reject } H_0\}$
- * GUI example

T-shirt sizes for effects

Power of a two-sample t test depends on $d = |\mu_1 - \mu_2| / \sigma$

* **Small:** $d = .15$

* **Medium:** $d = .25$

* **Large:** $d = .40$

T-shirt sizes for effects

Power of a two-sample t test depends on $d = |\mu_1 - \mu_2| / \sigma$

* **Small:** $d = .15$

* **Medium:** $d = .25$

* **Large:** $d = .40$

Q. Who is the T-shirt supposed to fit?

* **Human?** $\sigma = 1$, say Medium : $|\mu_1 - \mu_2| = .25$ in

T-shirt sizes for effects

Power of a two-sample t test depends on $d = |\mu_1 - \mu_2| / \sigma$

* **Small:** $d = .15$

* **Medium:** $d = .25$

* **Large:** $d = .40$

Q. Who is the T-shirt supposed to fit?

* **Human?** $\sigma = 1$ Medium : $|\mu_1 - \mu_2| = .25$ in

* **Hippo?** $\sigma = 25$ Medium : $|\mu_1 - \mu_2| = 6.25$ in

T-shirt sizes for effects

Power of a two-sample t test depends on $d = |\mu_1 - \mu_2| / \sigma$

* **Small:** $d = .15$

* **Medium:** $d = .25$

* **Large:** $d = .40$

Q. Who is the T-shirt supposed to fit?

* **Human?** $\sigma = 1$ Medium : $|\mu_1 - \mu_2| = .25$ in

* **Hippo?** $\sigma = 25$ Medium : $|\mu_1 - \mu_2| = 6.25$ in

* **Mouse?** $\sigma = .04$ Medium : $|\mu_1 - \mu_2| = .01$ in

A definitive study

... is not based on generic criteria

- * Specify effect size on the actual scale of measurement, based on well-considered scientific goals
- * Need to know σ , approximately *
- * If you can't do these things, reaching the bottom line is a matter of luck
- * * Except possibly in cases where σ defines population norms

Planning a pilot study

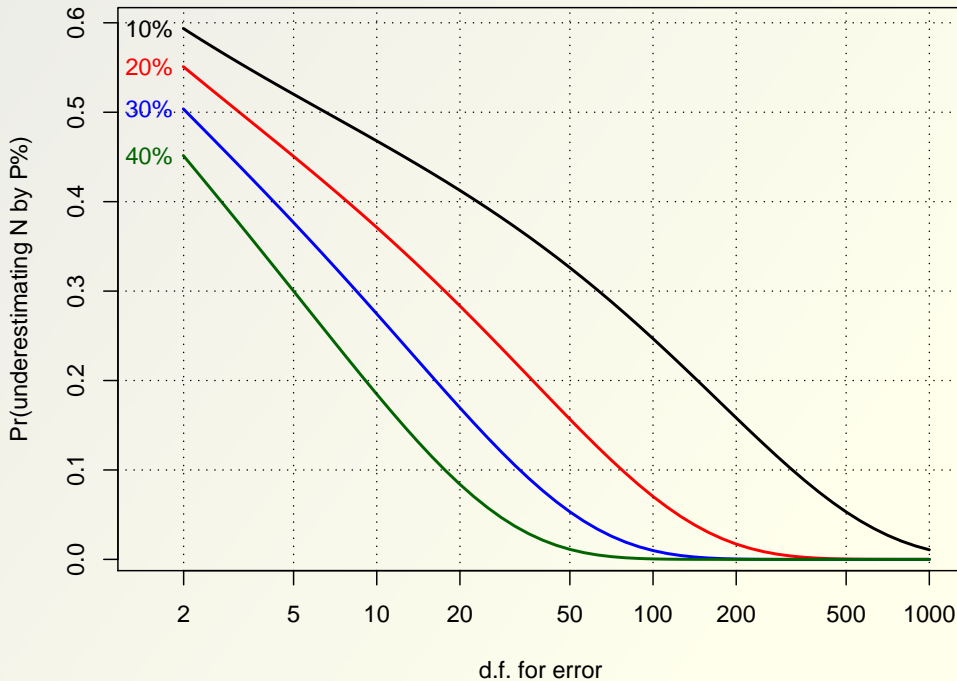
One simple approach

- * Control the probability of under-estimating N by a specified percentage P
- * Percentage by which N is underestimated = percentage by which σ^2 is underestimated
- * In normal case,

$$\begin{aligned}\Pr(S^2 \leq (1 - P)\sigma^2) &= \Pr(\nu S^2/\sigma^2 \leq (1 - P)\nu) \\ &= \Pr(Q \leq (1 - P)\nu)\end{aligned}$$

where S^2 has ν d.f. and $Q \sim \chi^2_\nu$

Chart for planning (or fudging)



Example: Semiconductor experiment

Structure

- * Response measure: Oxide thickness of silicon wafers
- * Three whole-wafer treatments
- * n lots of three wafers each: one wafer per treatment
- * Three sites per wafer

Target effect sizes (for power .80, .05 sig. level)

- * Difference of ± 10 between two treatments
- * Difference of ± 5 between two site means
- * Difference of ± 15 between two treatment*site means

Available data

(From R package nlme) 4 lots of 3 wafers each from each of two sources; 3 sites/wafer

```
source × site
|
LOT
|
WAFER

Error: Lot
      Df Sum Sq Mean Sq F value Pr(>F)
Source  1 1830.1  1830.1   1.5261 0.2629
Residuals 6 7195.2  1199.2

Error: Lot:Wafer
      Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 16 1922.67  120.17

Error: Within
      Df Sum Sq Mean Sq F value Pr(>F)
Site    2  15.44    7.72   0.6416 0.5313
Source:Site  2  58.33   29.17   2.4234 0.1004
Residuals 44 529.56   12.04
```

SD estimates

- * **SD(LOT)** $\approx \sqrt{(1200 - 120)/9} \approx 11.0$ (not really needed)
- * **SD(WAFER in LOT)** $\approx \sqrt{(120 - 12)/3} = 6.0 \rightarrow$ **SD(LOT \times treat)**
- * **SD(ERROR)** $\approx \sqrt{12} \approx 3.5$

Summary

- * Power/sample size is technically messy
- * Often have multiple objectives
- * Sometimes need to re-define goals
- * A flexible user interface can help