BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 12: Sample Size and Power

Version: 03/11/2021

Outline

Topics:

- Sample size and power
- Power and sample size for the mean difference
- Power and sample size for simple linear regression
- Power and sample size for survival outcomes
- Simulation-based power calculations in R
- Type-I error and false discovery error rate control

Outline

Topics:

- **Sample size and power**
- Power and sample size for the mean difference
- Power and sample size for simple linear regression
- Power and sample size for survival outcomes
- Simulation-based power calculations in R
- Type-I error and false discovery error rate control

Sample size and power

Ideas:

- At the study planning stage, we need to have some idea of what kind of sample size we'll be looking for.
  - If study is too small, we'll be setting ourselves up for a study that is unable to detect an association, if one exists.
  - If study is too big, we're being wasteful (could have gotten the answer with fewer patients—less time, monetary resources).
- Mistake to avoid: Thinking that a fixed power corresponds to one study (every study has, for instance, 80% power to detect *some* association).

Sample size and power

Terminology: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

- Significance level, α: Threshold at which H_0 will be rejected.
  - In the cleanest of examples, this also corresponds to the type-I error rate of the test being performed.
  - For this reason, terms "level" and "type-I error rate" often used interchangeably, although they technically shouldn't be.
  - Typically, we design a test to control this quantity.
- Type-II error rate, $\beta(\theta^*)$: P(Reject $H_0 | \theta = \theta^*$).
  - Power$(\theta^*) = 1 - \beta(\theta^*)$.
  - Power is not a fixed quantity and depends upon the truth.
- We want a low α and a low β (high power). In particular,
  - If H_0 is true, we don't want there to be too large a probability of inadvertently rejecting it.
  - If θ is meaningfully different from θ_0, we want to reject H_0 with high probability.

Sample size and power

Factors influencing power:

- In general, power to detect a fixed alternative depends upon:
  - The level of the test.
    - Setting the threshold to declare significance at a higher value obviously increases the probability of rejecting the null.
  - Variability in the outcome.
    - The lower the variability, the higher the power.
  - Sample size.
    - The higher the sample size, the higher the power.
  - "Distance" between true value of θ and the null value, θ_0.
    - The larger the distance, the higher the power.
- Some wacky examples exist that serve as exceptions to the above "rules."

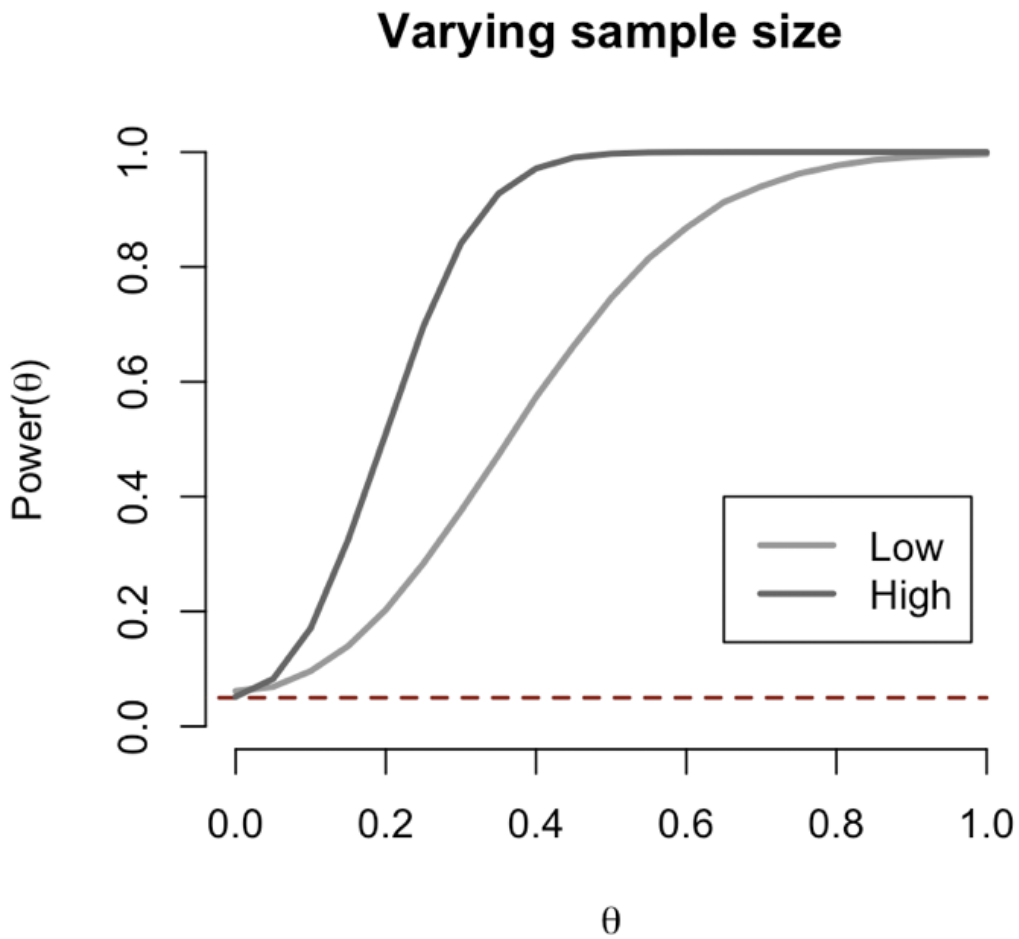
Sample size and power

Example: Linear regression

- Consider the following setup for some sample size, N:
  - $X \sim \mathcal{N}(0, 1)$
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$
  - $Y = \theta X + \epsilon$
- Let us unpack how power varies across θ while varying:
  - Significance level, α.
  - Error variance, σ^2.
  - Sample size, N.

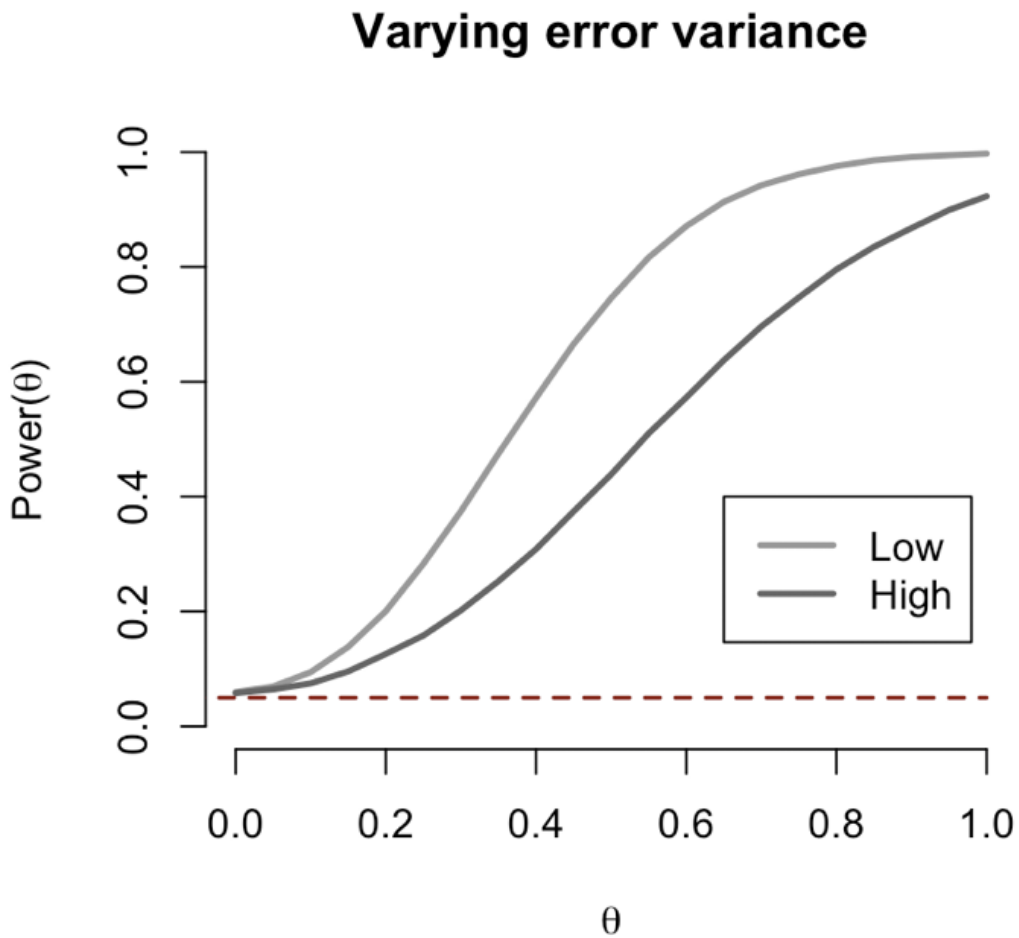
Sample size and power

Example: Linear regression



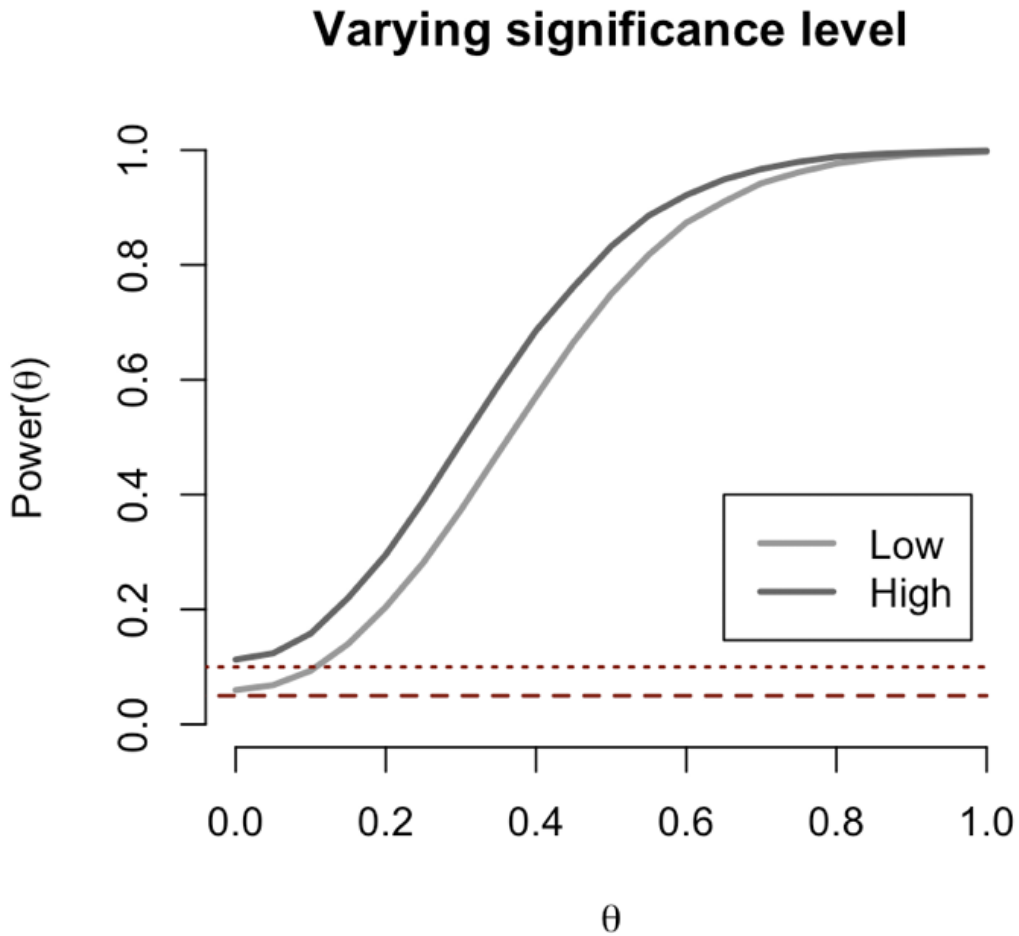
Sample size and power

Example: Linear regression



Sample size and power

Example: Linear regression



Sample size and power

Methodology:

- In an ideal world, the general line of logic is to find a sample size that has sufficiently high power to detect a difference that is scientifically/clinically relevant.
- Your daily dose of cynicism: Here is the most common sample size formula you will use:

$$N = \frac{\$\$\$\text{available} - \$\text{overhead} - \$\text{salary support} - \$\text{lab costs}}{\$\text{per study subject}}$$

- But just for fun, let's look at some others!

Outline

Topics:

- *Sample size and power*
- **Power and sample size for mean differences**
- Power and sample size for simple linear regression
- Power and sample size for survival outcomes
- Simulation-based power calculations in R
- Type-I error and false discovery error rate control

Mean differences

Setup and sample size formula:

- Test:
  - Null hypothesis: $H_0 : \mu_1 - \mu_0 = 0$.
  - Design alternative: $H_1 : \mu_1 - \mu_0 = \delta$.
- Operating characteristics (two-sided test):
  - Significance level: α.
  - Power: $1 - \beta$.
- Presumed variance in each group: σ^2.

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} \text{ per group.}$$

- Stata: `power twomeans` (very flexible).

Outline

Topics:

- *Sample size and power*
- *Power and sample size for mean differences*
- **Power and sample size for simple linear regression**
- Power and sample size for survival outcomes
- Simulation-based power calculations in R
- Type-I error and false discovery error rate control

Simple linear regression

Setup and sample size formula:

- Test:
  - Null hypothesis: $H_0 : \beta_1 = 0$.
  - Design alternative: $H_1 : \beta_1 = \beta_1^*$.
  - Association seeking to detect: β_1^* (alternatively, $\rho^*_{X,Y}$).
- Operating characteristics (two-sided test):
  - Significance level: α.
  - Power: $1 - \beta$.
- Moving between β_1^* and ρ^*_{XY}:
  - Presumed outcome variance: σ^2_Y.
  - Presumed exposure variance: σ^2_X.
  - Note: $\rho^*_{X,Y} = \beta_1^*(\sigma_X/\sigma_Y)$.

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{[\rho^*_{X,Y}]^2}.$$

- Stata: `power onecorrelation`.

Outline

Topics:

- *Sample size and power*
- *Power and sample size for mean differences*
- *Power and sample size for simple linear regression*
- **Power and sample size for survival outcomes**
- Simulation-based power calculations in R
- Type-I error and false discovery error rate control

Survival outcomes

Sample size: Two groups

- Hazard function in control group: $\lambda_0(t)$.
- Hazard function in experimental group: $\lambda_1(t)$.
- Proportional hazards assumption: $\lambda_1(t) = \lambda_0(t)\exp(\beta)$.
- $H_0 : \beta = 0 \Rightarrow \exp(\beta) = 1 \Rightarrow \lambda_0(t) = \lambda_1(t)$.
- Key fact: $\widehat{\beta} \dot{\sim} \mathcal{N}(\beta, 4/L)$, where L denotes the number of events needed in the pooled sample. Solving backwards:

$$L = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\left[\frac{1}{2}\beta_{\text{Alt. Hypoth.}}\right]^2}.$$

Survival outcomes

Sample size: Two groups

- Formula for number of events:

$$L = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\left[\frac{1}{2}\beta_{\text{Alt. Hypoth.}}\right]^2}.$$

- If we want 80% power and a type I error rate of 0.05 (and as long as $H_0 : \beta = 0$ is your true null hypothesis, then the formula reduces: $L = 31.4/\beta^2_{\text{Alt. Hypoth.}}$.

Survival outcomes

Sample size: Two groups

- Formula for number of events:

$$L = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\left[\frac{1}{2}\beta_{\text{Alt. Hypoth.}}\right]^2}.$$

- If we want 80% power and a type I error rate of 0.05 (and as long as $H_0 : \beta = 0$ is your true null hypothesis), then the formula reduces: $L = 31.4/\beta^2_{\text{Alt. Hypoth.}}$.
- If we suppose an event rate of f, then our total sample size should be approximately $N = L/2f$ per group.

Outline

Topics:

- *Sample size and power*
- *Power and sample size for mean differences*
- *Power and sample size for linear regression*
- *Power and sample size for survival outcomes*
- **Simulation-based power calculations in R**
- Type-I error and false discovery error rate control

Simulations

Basic ideas:

- When in doubt, can use simulation-based methods.
- Accounts for complexities such as:
  - Precision variables.
  - Confounding.
  - More sophisticated models (competing risks, ordinal models, longitudinal data).
- Need to specify parameters of simulation (always a good idea to vary across reasonable range).
- Simulations can also inform you about the *actual* type-I error rate under a specific level.

Simulations

Example: Type-I error with log-normal errors

```
set.seed(1)
nsim <- 500000
res <- matrix(0, nrow = nsim, ncol = 1)
for (j in 1:nsim)
{
    X <- matrix(cbind(1, rnorm(n = 30, mean = 0, sd = 1)), nrow = 30)
    Y <- exp(rnorm(n = 30, mean = 0, sd = 0.2))
    bht <- (solve(t(X) %*% X) %*% t(X) %*% Y)
    sebht <- sqrt((solve(t(X) %*% X)[2, 2] * sum((Y - X %*% bht)^2)/(30 - 2)))
    res[j,1] <- as.numeric(2 * pnorm(-abs(bht[2]/sebht)) < 0.05)
    if (round(j/50000) == (j/50000)) {print(j)}
}
print(mean(res))
```

- Nominal level: $\alpha = 0.05$
- Setting seed (1): type-I error rate: 0.0598.
- Resetting seed (2): type-I error rate: 0.0607
- Resetting seed (3): type-I error rate: 0.0601

Outline

Topics:

- *Sample size and power*
- *Power and sample size for mean differences*
- *Power and sample size for linear regression*
- *Power and sample size for survival outcomes*
- *Simulation-based power calculations in R*
- **Type-I error and false discovery error rate control**

Error rate control

Multiple comparisons: Error rate inflation

- The more level-α tests you conduct in a single data set, the higher the chance you'll declare at least one of those tests statistically significant.
  - The *family-wise* type-I error rate will be inflated past α.
- The type-I is related to, but (importantly) *distinct from* a false discovery rate.
  - Type-I error rate: P(Declare significance$|H_0$ true).
  - False discovery rate: Expected proportion of rejections among all rejections.
- Some methods aim to control the former (the most well known of which is probably the Bonferroni correction). Others aim to control the latter, and they are *not* equivalent.

Error rate control

Type-I error rate control: Bonferroni

- Let $H_1, \ldots, H_K$ denote a family of K hypotheses, and $p_1, \ldots, p_K$ their respective p-values.
- Let $K_0 \leq K$ denote the number of these null hypotheses that are actually true (generally unknowable).
- Bonferroni procedure:
  - Rejects null hypotheses in which p-values do not exceed α/K, thereby controlling the probability that of at least *one* type-I error at under α.
- Often seen as conservative (stricter than necessary).
  - But not always (e.g., group-sequential monitoring).
  - Holm–Bonferroni method is less conservative.

Error rate control

False discovery rate control: Benjamini–Hochberg

- Let $H_1, \ldots, H_K$ denote a family of K hypotheses, and $p_1, \ldots, p_K$ their respective p-values, *in ascending order*.
- Benjamini–Hochberg procedure:
  - Find the largest k such that $p_k \leq k\alpha/K$.
  - Reject the null hypotheses for all H_i with $i = 1, \ldots, k$.
- If tests are independent, false discovery rate is controlled at α, but the type-I error rate is *not* (less conservative).
  - Benjamini–Yekutieli serves as alternative for dependent tests.

Error rate control

Notes:

- The idea of controlling the rate of "mistakes" is not (and should not be) controversial.
- What lacks consensus the degree to which and the circumstances under which these procedures are appropriate.
  - Thousands of genes?
    - Certainly.
  - Two pre-specified primary hypotheses in a Phase III trial?
    - Probably not.
  - Other things? Gray area.
- Advice: Pre-specify your hypotheses, analyses, and testing procedures and power accordingly (easier said than done).

Error rate control

Notes:

- Jeffrey Blume's second-generation p-value was designed to translate point null hypotheses to *interval null* hypotheses of clinically relevant thresholds.
  - Loosely speaking, characterizes extent of overlap between an interval estimate and the interval null.
- Interestingly, in the context of multiple tests, the second-generation p-value tends to have better properties in terms of error rates as compared to traditional p-values.