Accelerating Impact: Immersive Summer Bootcamp in Implementation Science and Biostatistics

Georgian Implementation Science Fogarty Training (GIFT) Program

Ilia State University & Yale University

Yale Center for Analytical Sciences

Sample Size and Power

Effects of a Weight Management Program on Body Composition and Metabolic Parameters in Overweight Children

A Randomized Controlled Trial

Mary Savoye, RD, CD-N, CDE
Melissa Shaw, BS
James Dziura, PhD
William V. Tamborlane, MD
Paulina Rose, RD, CD-N, CDE
Cindy Guandalini, APRN
Rachel Goldberg-Gell, APRN
Tania S. Burgert, MD
Anna M. C. Cali, MD
Ram Weiss, MD, PhD
Sonia Caprio, MD

Context Pediatric obesity has escalated to epidemic proportions, leading to an array of comorbidities, including type 2 diabetes in youth. Since most overweight children become overweight adults, this chronic condition results in serious metabolic complications by early adulthood. To curtail this major health issue, effective pediatric interventions are essential.

Objective To compare effects of a weight management program, Bright Bodies, on adiposity and metabolic complications of overweight children with a control group.

Design One-year randomized controlled trial conducted May 2002-September 2005.

Setting Recruitment and follow-up conducted at Yale Pediatric Obesity Clinic in New Haven, Conn, and intervention at nearby school.

Participants Random sample of 209 overweight children (body mass index [BMI] >95th percentile for age and sex), ages 8 to 16 years of mixed ethnic groups were recruited. A total of 135 participants (60%) completed 6 months of study, 119 (53%) completed 12 months.

Intervention Participants were randomly assigned to either a control or weight management group. The control group (n=69) received traditional clinical weight management counseling every 6 months, and the weight management group (n=105) received an intensive family-based program including exercise, nutrition, and behavior modification. Intervention occurred biweekly the first 6 months, bimonthly thereafter. The second randomization within the weight management group assigned participants (n=35) to a structured meal plan approach (dieting), but this arm of the study was discontinued while enrollment was ongoing due to a high dropout rate.

Main Outcome Measures Change in weight, BMI, body fat, and homeostasis model assessment of insulin resistance (HOMA-IR) at 6 and 12 months.

Results Six-month improvements were sustained at 12 months in weight management vs control, including changes in the following (mean [95% confidence interval]): weight (+0.3 kg [−1.4 to 2.0] vs +7.7 kg [5.3 to 10.0]); BMI (−1.7 [−2.3 to −1.1] vs +1.6 [0.8 to 2.3]); body fat (−3.7 kg [−5.4 to −2.1] vs +5.5 kg [3.2 to 7.8]); and HOMA-IR (−1.52 [−1.93 to −1.01] vs +0.90 [−0.07 to 2.05]).

Conclusion The Bright Bodies weight management program had beneficial effects on body composition and insulin resistance in overweight children that were sustained up to 12 months.

Trial Registration clinicaltrials.gov Identifier: NCT00409422

JAMA. 2007;297:2697-2704 www.jama.com

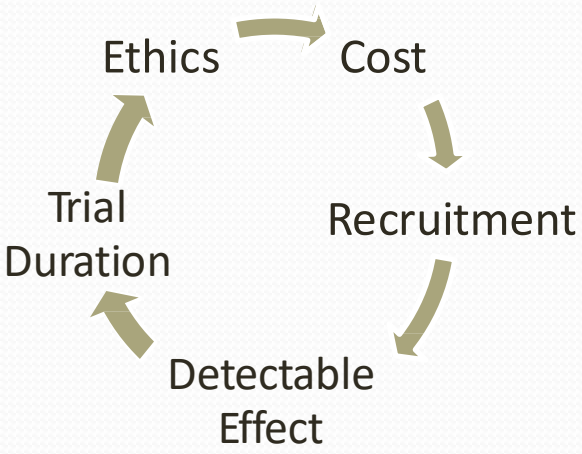
THE PREVALENCE OF OVERweight among children and adolescents increased significantly from 1999-2004, with 17% overall defined as overweight and with an even higher prevalence among African American youth (18%-26%).[1] Attention has focused on the increase of type 2 diabetes in adolescents that has accompanied the epidemic of childhood obesity, which is due to a combination of severe insulin resistance and progressive beta-cell failure, and is more common in African American and Hispanic youth.[2,3] However, an even greater proportion of obese children have impaired glucose tolerance,[4] as well as hypertension, dyslipidemia, and other components of the metabolic syn-

Sample Size and Statistical Analysis

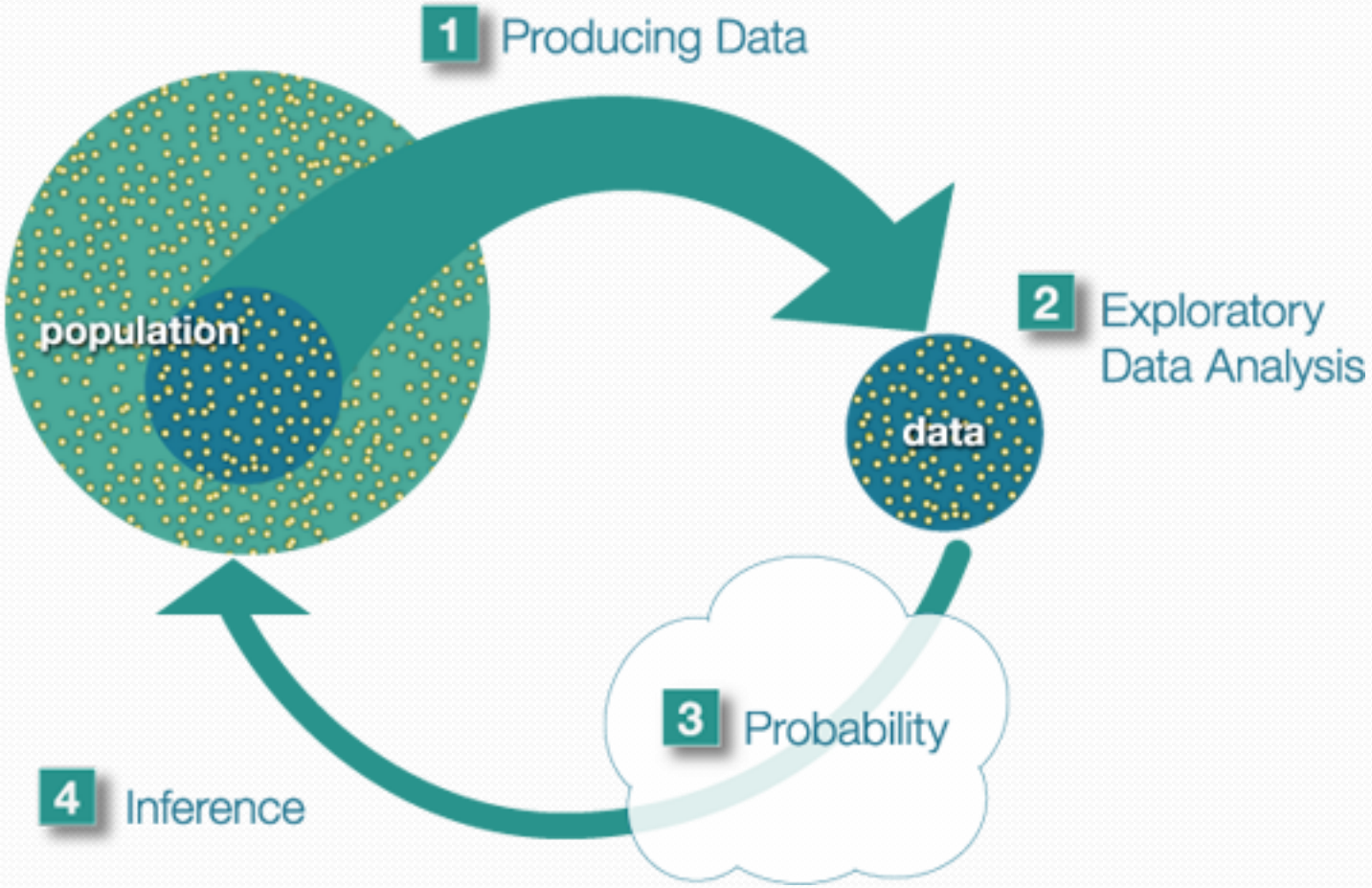
For the purpose of sample size estimation, the primary outcome of this study was change in BMI between baseline and 6 months. Our database of over 600 overweight children between the ages of 8 and 16 years estimates an SD for BMI of 6.8. Conservatively assuming a correlation of 0.75 between BMI at baseline and 6 months, a 2-sided .017 significance level and an SD of 7, a sample size of 58 participants per group would provide 80% power to detect a treatment difference of 3. We planned to enroll 70 participants per group to account for a 20% dropout. Dropping the

- Sample size determination is one of the most important steps in study design
  - Inadequate sample size reason for many inconclusive studies

E.g. Goal for Clinical Trial: Determine the sample size needed to detect a pre-specified clinically relevant difference between treatment groups with high probability



The Big Picture



http://bolt.mph.ufl.edu/6050-6052/preliminaries/the-big-picture/

Hypothesis Tests

- Statistical way of evaluating association
- Assume null: H_0: $\mu_1 = \mu_2$ or $\pi_1 = \pi_2$
  - eg 1. Mean cholesterol is same for obese adults following treatment with either statin or placebo
  - eg 2. The proportion of individuals with hypertension is the same for those on treatment A as compared to treatment B.
- Calculate a statistic that tells you how far your observed data are from the null hypothesis
  - t, F or X^2 --- p-value under the null

Outcomes of a Hypothesis Test

- Correctly identify an association
  - "Convict the guilty"
- Correctly identify no association
  - "Acquit the innocent"
- Type I error (α): probability of finding an association when there is none
  - False positive / "Convict the innocent"
- Type II error (β): probability of missing an association if there really is one.
  - False negative / "Acquit the guilty"

What does Sample Size Depend On?

- Z-statistic for comparison of 2 pop. means

$$z = \frac{\mu_A - \mu_B}{\sigma_{(\mu_A - \mu_B)}} \longrightarrow \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

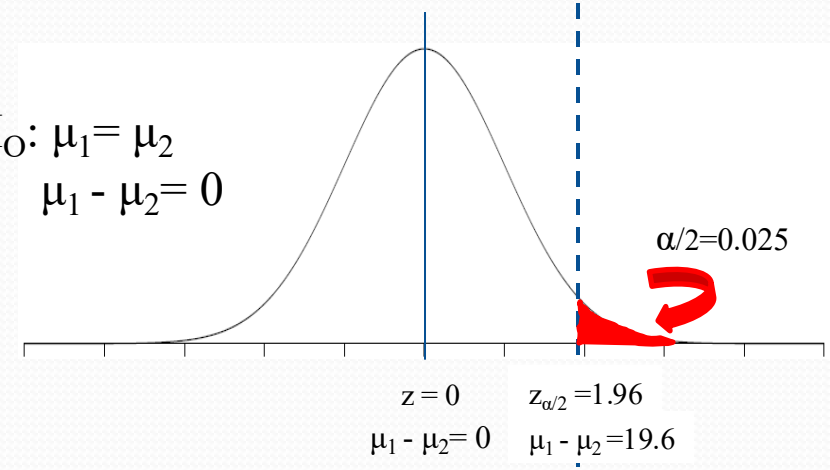
$$n_{\text{per group}} = \frac{2\sigma^2 z^2}{(\mu_A - \mu_B)^2}$$

Power

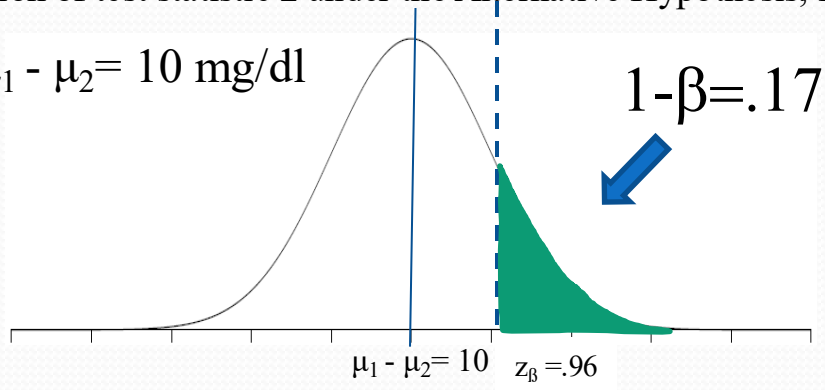
- 1-β: The ability to detect a difference or association with a hypothesis test when a true difference or association exists in the population(s).
  - Probability of finding statistical significance when a true difference exists.

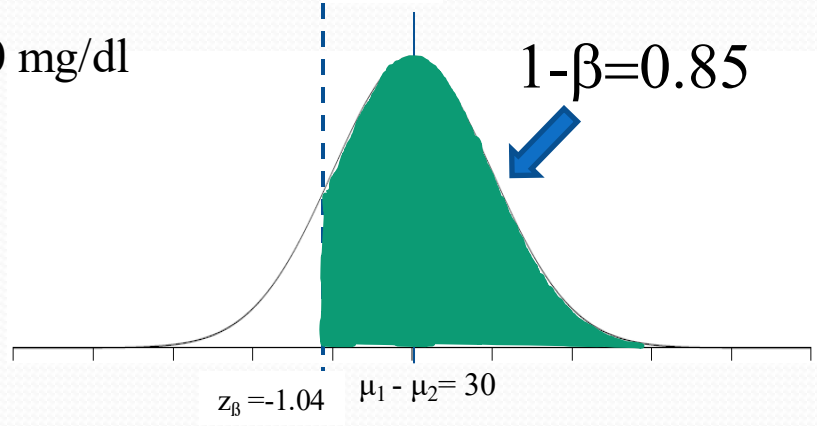
- How do we quantify power?
  - Based on the distribution of a test statistic (e.g, z or t) under the Null and the Alternative hypotheses

Distribution of test statistic **z** under the Null Hypothesis, H_0:



Distribution of test statistic **z** under the Alternative Hypothesis, H_A:





Compare the Cholesterol Levels following receipt of either Drug 1 or Drug 2, using:

- 2-sided test, z- or t-test
- α=0.05
- $s_1 = s_2 = 44.7\ mg/dl$
- $n_1 = n_2 = 40$

Estimate population standard error for difference of means:

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{44.7^2\left(\frac{1}{40} + \frac{1}{40}\right)} = 10\ mg/dl$$

Why Bother Estimating Sample Size?

- Freiman JA (NEJM, 1979) - reviewed 71 randomized trials with non-significant results - only 20% included enough subjects to detect a 25% improvement in outcome with 50% power.

- Moher D et al (JAMA, 1994) – reviewed 383 RCT – 32% of studies with negative results reported sample size calculations; these increased from 0% in 1975 to 43% in 1990.

- Charles P et al (BMJ, 2009) – reviewed 215 RCT published between Jan 2005 and Dec 2006; 5% did not report any sample size and 43% did not report all required parameters; 31% mismatched assumptions (>30% difference for control group assumptions).

Ideal Sample Size

- Big Enough – scientifically or clinically meaningful effects will be declared significant - POWER
- Not Too Big – so resources aren't wasted on finding an effect of no scientific/clinical importance

Requirements for Sample Size Planning

- Clearly defined, simple and specific hypothesis
  - Best - one primary hypothesis and one primary outcome
  - What is the outcome?
  - What is the comparison? (superiority, noninferiority, equivalence, association)
- Acceptable type I and II error rates
  - $\alpha = 0.05$
  - $\beta = 0.20$ or 0.10 (more conservative)
- Estimate of variability in outcome (i.e., variance)
- Estimate of clinically meaningful effect size (i.e. difference in means or proportions, hazard ratio)
  - absolute vs. relative
  - This is not a statistical decision

Components of Sample Size

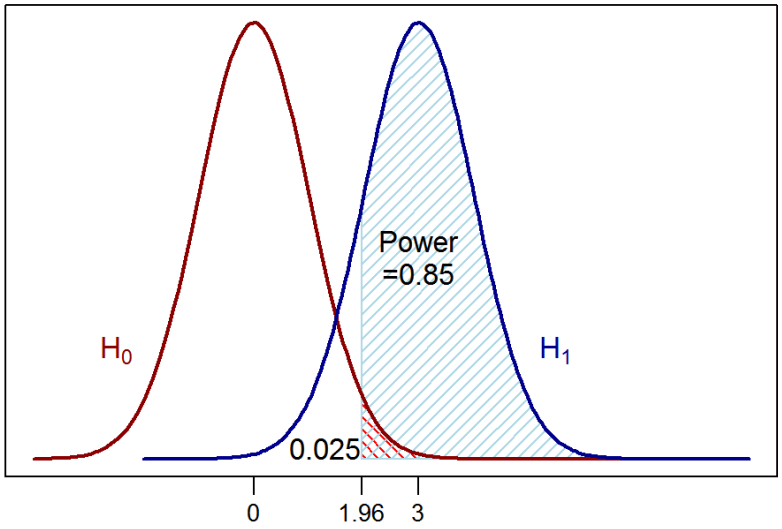
General Per-Group Sample Size Formula

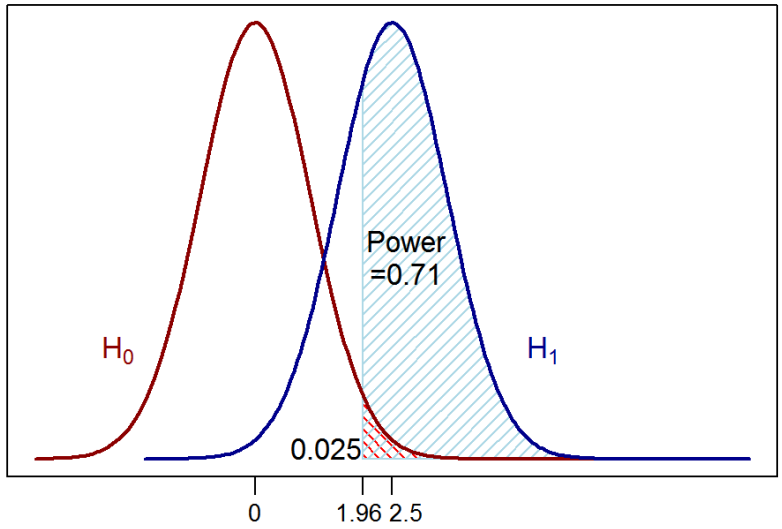
$$n\ per\ group = \frac{(z_{\alpha/2} + z_{\beta})^2 * 2 * (variability)^2}{\delta^2}$$

- **Size of difference** considered important (δ)
  - δ down, *n* **up**
- **Type I error** or significance level
  - Type I error down ($z_{\alpha/2}$ up) , *n* **up**
- **Type II error** or power
  - Type II error down (power up, z_{β} up), *n* **up**
- **Variability** of response
  - Variability up, *n* **up**

Factors that Affect Power and Sample Size

1. **Size of difference** considered important (δ)
   - δ smaller, power smaller

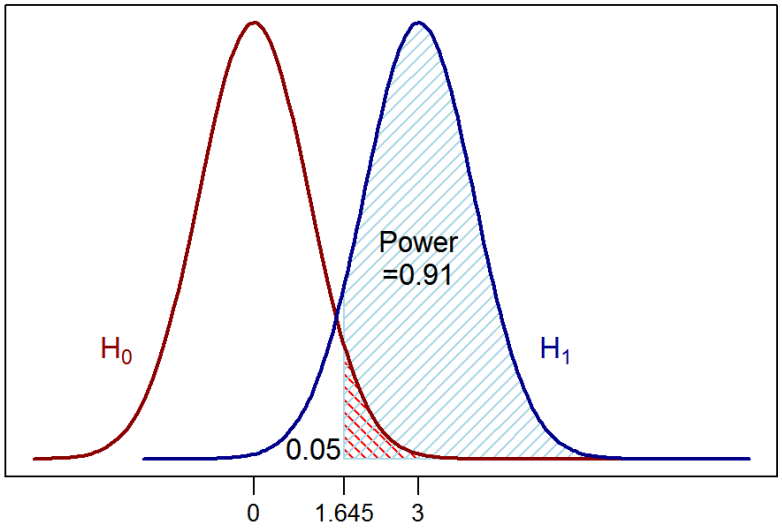


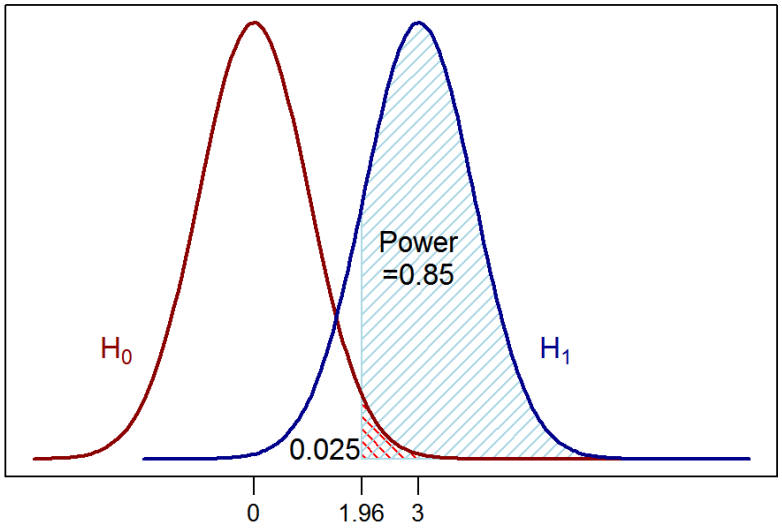


We will need a **larger sample size** to detect a smaller δ at the same level of power

Factors that Affect Power and Sample Size

2. **Type I error** or significance level (α)
   - α smaller, power smaller

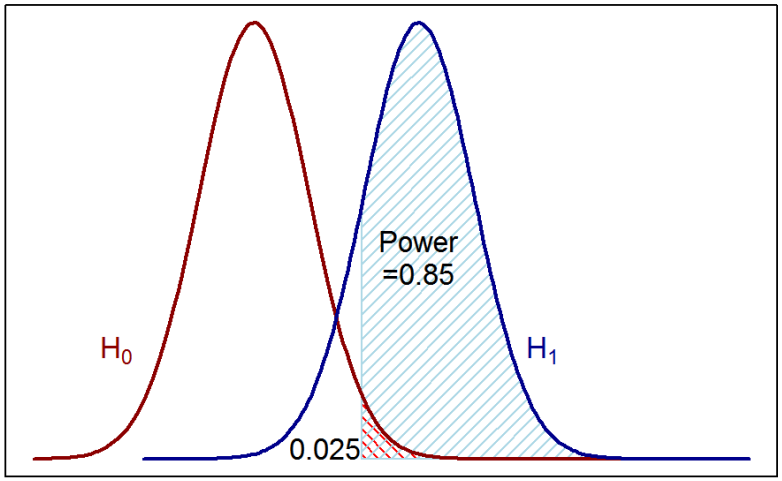


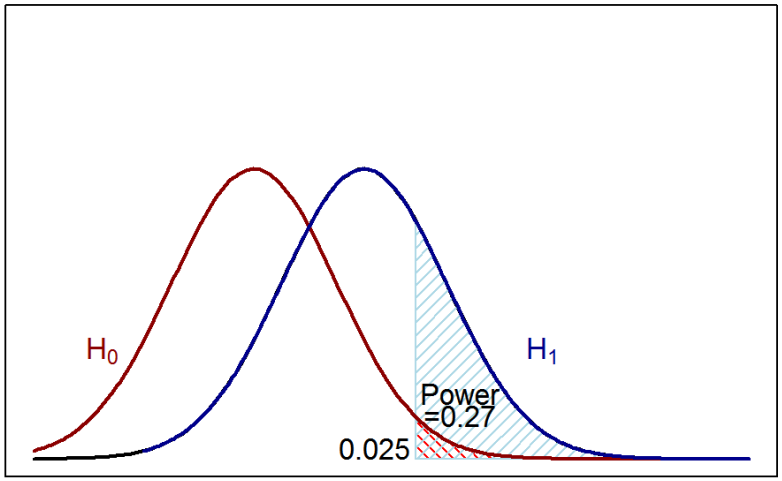


We will need a **larger sample size** to perform the test at a lower α and maintain the same level of power

Factors that Affect Power and Sample Size

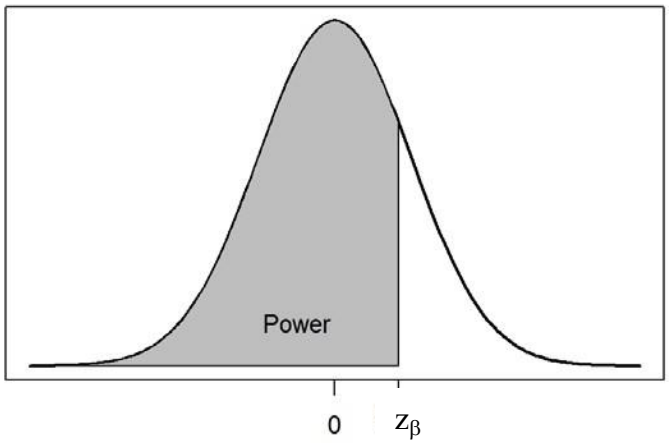
3. **Variability** of the response
   - Variability larger, power smaller





We will need a **larger sample size** to maintain the same level of power if the variability is larger

General Sample Size Formula



$$n\ per\ group = \frac{(z_{\alpha/2} + z_\beta)^2 * 2 * (variability)^2}{\delta^2}$$

1-β	z_β	α (2-sided)	$z_{\alpha/2}$	$(z_\beta + z_{\alpha/2})^2$
0.80	0.84	0.05	1.96	7.84
	0.84	0.025	2.24	9.49
	0.84	0.01	2.58	11.70
0.90	1.282	0.05	1.96	10.51
	1.282	0.025	2.24	12.40
	1.282	0.01	2.58	14.92

1/3 larger for 90% vs. 80% power

Sample Size Key Point

Sample size calculation should be based on the same statistical test planned to analyze the primary outcome at study's completion.

- Types of outcomes
  - **Binary** (yes/no, success/failure)
  - **Categorical** (excellent, good, poor; below, within, above normal range)
  - **Continuous** (number of days of smoking abstinence due to behavioral intervention; change in weight between baseline and week 6)
  - **Time to event** (time to HIV progression; time to smoking cessation)

Sample Size Formulas

- t-test for independent samples
  - n per group =

$$\frac{(z_{\alpha/2} + z_{\beta})^2 * 2 * \sigma^2}{(\mu_2 - \mu_1)^2}$$

- test for proportions, using unpooled variance
  - n per group =

[or use formula on p.155]

$$\frac{(z_{\alpha/2} + z_{\beta})^2 * (p_1 q_1 + p_2 q_2)}{(p_2 - p_1)^2}$$

- Survival
  - Number of events =

$$\frac{4 * (z_{\alpha/2} + z_{\beta})^2}{(\ln HR)^2}$$

Effect Size

- The smallest difference that is considered to be scientifically or clinically meaningful
  - Not a statistical decision

- An upper bound might be what results you expect to see based on:
  - Preliminary data
  - Literature
  - Experience

Variance

- Experience
- Literature
- Pilot study
- For proportions and hazard rates, not a problem since variance is built in

Other Issues

- Loss to follow-up/Withdrawal/Crossovers
- Multiple Comparisons
- Interim Analyses
- Transformations
- Clustered Randomization
  - inflation factor: $1+(m-1)\rho$
- Longitudinal Repeated Measures
- Unequal Allocation
  - Need more information about safety and event rates
  - Expensive treatment
  - Increase participation

Other Issues

- Post-hoc (observed) power – don't bother
  - Conceptually power only describes pre-study probability
  - Provides information that is redundant to p-values and confidence intervals

Equivalence and Non-inferiority

- Rather than investigating whether two (or more groups) are different, we might be interested in showing that the groups are not different (equivalence) or that one group is no worse than another (non-inferiority)

- Traditional null hypotheses of no difference don't examine the hypothesis we're interested in

- Need to set limits of equivalence – narrow boundaries that indicate acceptable, clinically irrelevant differences
  - Calculate 95% confidence interval and make sure the interval lies fully within the boundaries
  - Conduct one-sided tests – declare equivalence if reject both
    - Ho: $\theta = \theta_L$ vs. Ha: $\theta > \theta_L$
    - Ho: $\theta = \theta_U$ vs. Ha: $\theta < \theta_U$

Power and Sample Size Software

- Commercial
  - nQuery
  - PASS
  - SAS (proc power)

- Freeware
  - http://www.math.uiowa.edu/~rlenth/Power/
  - Write your own code in R-script: