Sample Size Calculations: Theory and Practice

JEFF GILL
Distinguished Professor
Departments of Government and Mathematics & Statistics
American University

Motivation

- The goal is to achieve desired: significance levels, effect sizes, subsetting ability, and power (one minus the probability of a Type II error) by collecting a sufficiently large sample.

- In some settings the statistician does not participate in the design of the study.

- Therefore others do the required sample size calculations.

- This is all really about stipulating reasonable and defensible assumptions on the future data-generating process.

- It is important to realize that this is a *fictional enterprise*.

- Complexities: cluster sampling, stratified sampling, multilevel data, dropouts, outcome variable measurement, multiple studies, and stopping rules.

Power Difference: Aspirin For Non-Fatal Myocardial Infarction

- Treatment: 300 mg aspirin three times daily.

- Elwood and Sweetnam (1979) studied 1239 patients just after a myocardial infarction, and found the mortality difference at one year to be:

$$p_{aspirin} = 0.080 \qquad p_{placebo} = 0.107 \qquad CI_{0.95} = [-0.005 : 0.06].$$

- Later a study by the *Aspirin Reinfarction Research Study Group* (1980) of 6292 patients under almost exactly the same circumstances found:

$$p_{aspirin} = 0.092 \qquad p_{placebo} = 0.115 \qquad CI_{0.95} = [0.008 : 0.038].$$

- So the first confidence interval has length 0.065, and the second confidence interval has length 0.03.

- What is going on here?

Hypothesis Testing Review: Physiotherapy for Lung Cancer Patients

- Because sample size calculations are tied to hypothesis testing and decision-making in general, return to the foundational setup.

- Griffiths *et al.* (2000) conducted a randomized clinical trial to compare post-operative lung cancer patients under:
  - Treatment: a new pulmonary rehabilitation program, $n_T = 93$.
  - Control: standard care, $n_C = 91$.

- The outcome of interest is *distance walked in meters* six weeks after randomization, which is assumed to be normally distributed, and observed to have:
  - $\bar{x}_T = 211$, $s_T = 118$.
  - $\bar{x}_C = 123$, $s_C = 99$.

- Therefore the difference is characterized by the observed effect size and its standard error:

$$\bar{d} = \bar{x}_T - \bar{x}_C = 88, \qquad SE(\bar{d}) = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}} = \sqrt{\frac{118^2}{93} + \frac{99^2}{91}} = 16.04.$$

where $SE(\bar{d})$ assumes different underlying population standard deviations.

Hypothesis Testing Review: Physiotherapy for Lung Cancer Patients

- Hypotheses:

$$H_0 : \delta = \mu_T - \mu_C = 0 \qquad\qquad H_A : \delta = \mu_T - \mu_C > 0$$

where δ is the *true population effect size*, estimated by the observed effect size, $\bar{d}$.

- We know that $\bar{d} = \bar{x}_T - \bar{x}_C = 88$, which is clearly not zero, but the question is whether it is sufficiently far from zero *in units of* $SE(\bar{d})$ such that we conclude for H_A.

- The test statistic is:

$$z_d = \frac{\bar{d} - \delta_{\text{null}}}{s_{\text{pooled}}} = \frac{88 - 0}{16.04} = 5.4863 > 1.645$$

for a one-sided test.

- We therefore reject the Null Hypothesis of no difference and find evidence for the Alternative Hypothesis that the physiotherapy helps lung cancer patients at the $\alpha = 0.05$ level for a one-sided hypothesis test.

- This represents a completely typical hypothesis testing process, and serves as reminder before we go on.

Living With Errors

- Whenever you make a decision like this based on statistical analysis you run the risk of making an error because these decisions are based on probabilistic not deterministic statements.

- With hypothesis testing we care principally about two types of errors:

  - Type I Error: the probability that the Null Hypothesis is true and we reject it anyway. This is labeled α.

  - Type II Error: the probability that the Null Hypothesis is false and we fail to reject it. This is labeled β.

- Often we care more about $1 - \beta$, which is called Power, rather than β.

- Key problem: these errors are traded-off by determination of α, δ, β, n, and can never be fully eliminated.

- Frequent strategy: fix $\alpha = 0.05$ and $1 - \beta = 0.8$, determine a useful threshold for δ, and calculate the required n to make these happen.

Multiple Objectives

- So if you require $1-\beta > 0.8$ and $\alpha < 0.05$, then you only have sample size and minimal detectable (hypothesized) effect size to manipulate.

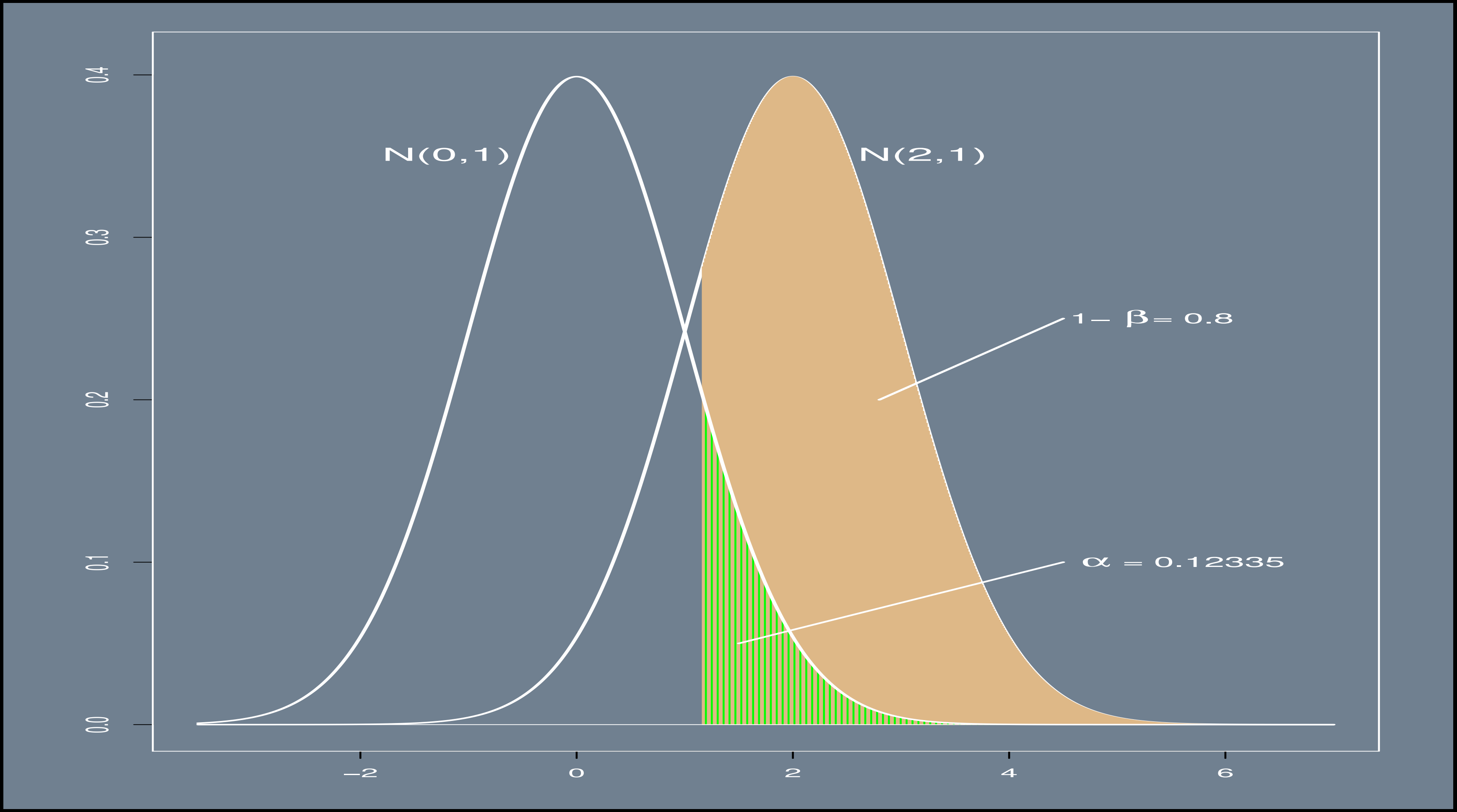
- Effect Size is usually the key quantity of interest here (the true magnitude of some treatment, procedure, exposure, etc.).

- Smaller samples are achieved in studies designed to maximize potential effect size, subject to other constraints.

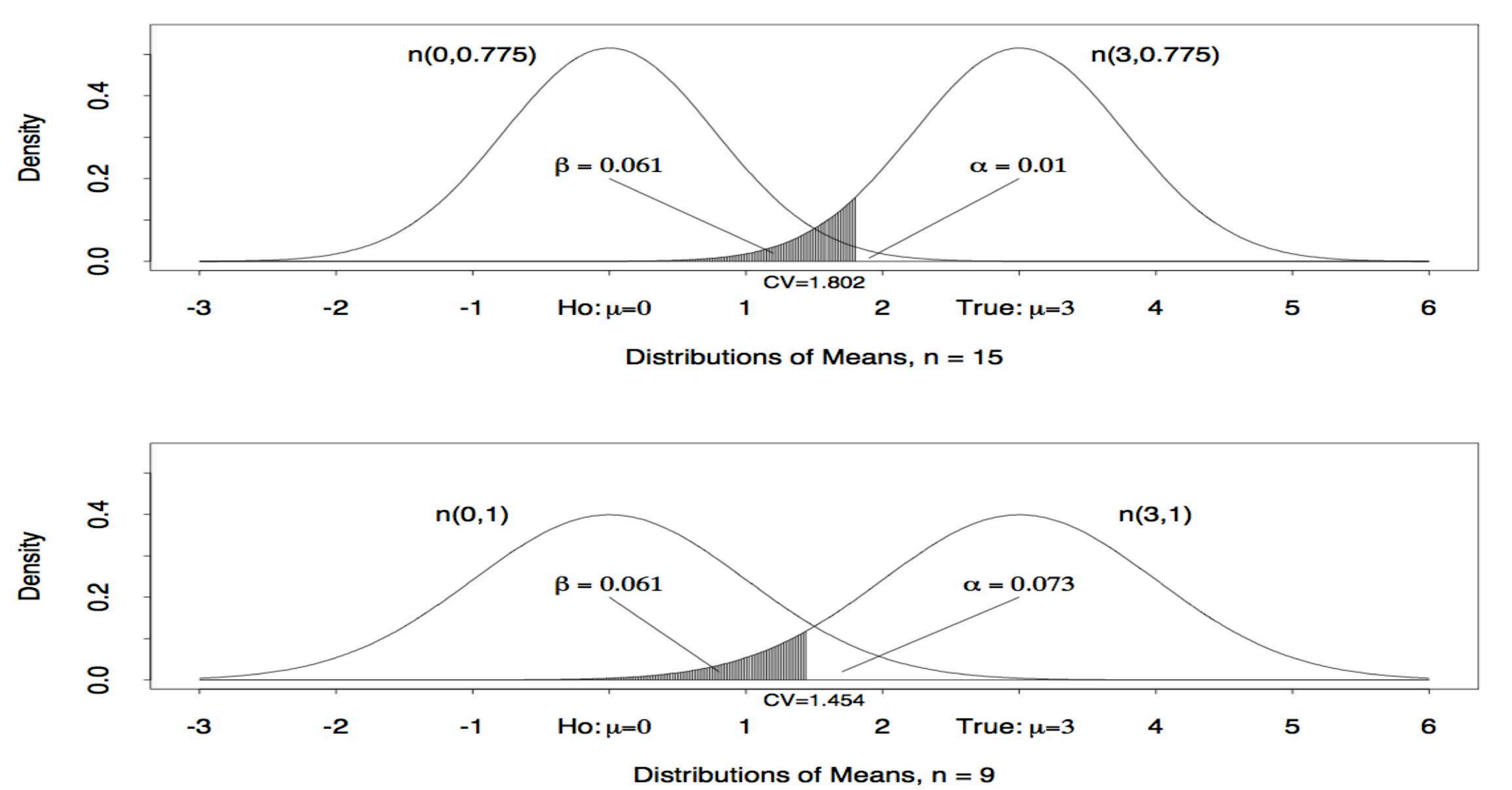
- This is why many drug studies give the lowest possible level to the control group and the highest safe level to the treatment group.

- So how do we *visualize* these trade-offs between α, δ, β, and n?

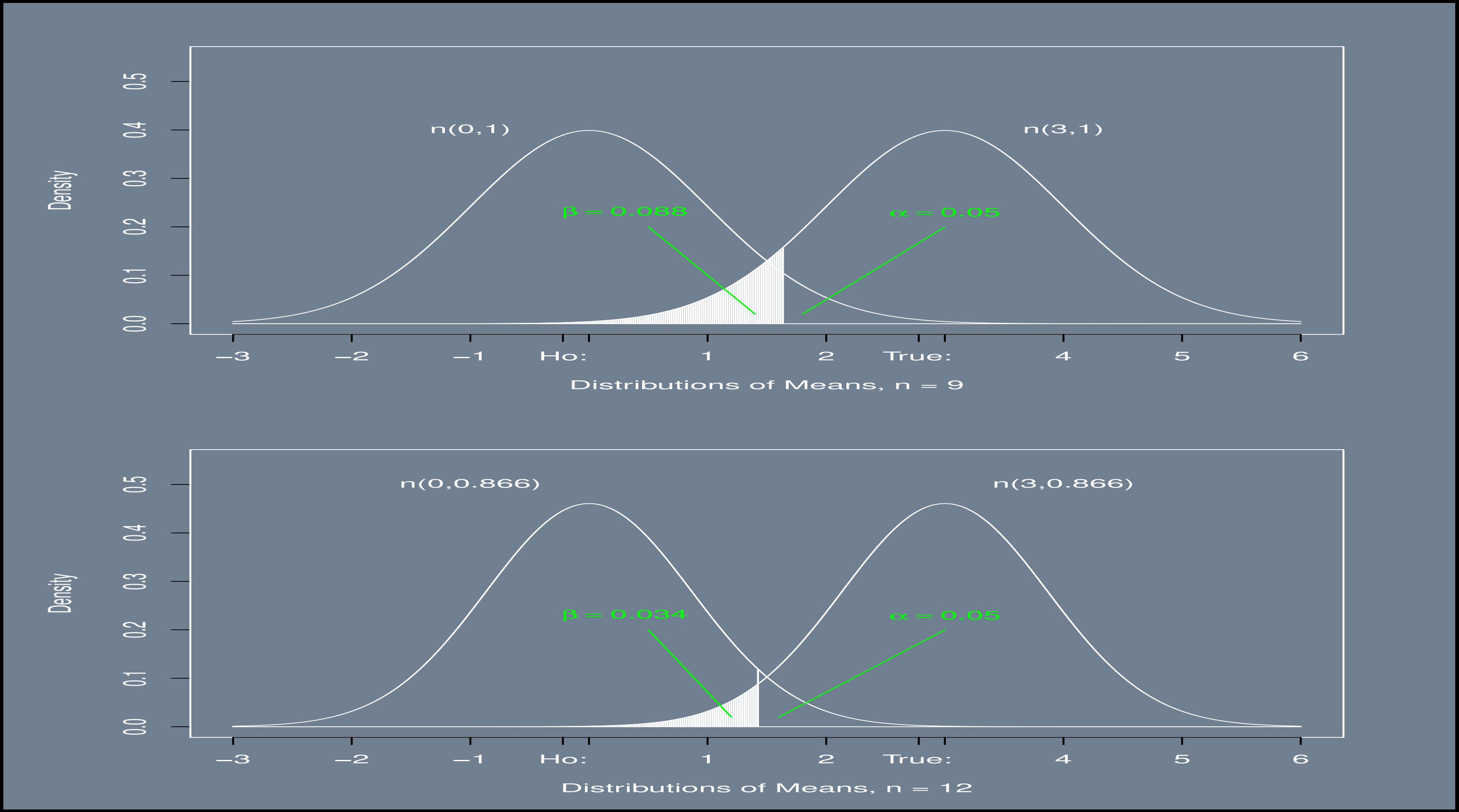
General Illustration of Trade-Offs



Implications of Sample Size on Realized Significance For Means



Implications of Sample Size on Power For Means



A Key Principle

- We are always interested in testing some *effect size* as shown in the figures as a difference in hypothesized means using a mean estimate ($\bar{y}$), or possibly an observed proportion ($\hat{p}$).

- More generally denoted as $\delta = \theta - \theta_0$ for the difference in the true value and the hypothesized value that motivates the study (in medicine often the "minimal detectable effect size").

- If the green area from the first plot equaled the tan area then there would be no difference, $\delta = 0$.

- This means that for this standard normal case:

$$\delta = 0 \;\longrightarrow\; \theta_0 + \alpha = \theta - (1 - \beta).$$

- If we generalize this statement to all normals it becomes:

$$\theta_0 + \alpha \times SE(\bar{y}) = \theta - (1 - \beta) \times SE(\bar{y}).$$

- Restated more generally and using $\alpha = 0.5$ and $1 - \beta = 0.8$:

$$\text{Threshold} + 95\%\ \text{CV} \times \text{Standard Error} = \text{Assumed Mean} - \Phi(0.8) \times \text{Standard Error}$$

  where 95% CV $= 1.96$ and $\Phi(0.8) = 0.8416$, so findings occur when Threshold and Assumed Mean are statistically different.

A Key Principle

- Substituting in the definitions of these quantities:

$$\text{Threshold} + 95\%\ \text{CV} \times \text{Standard Error} = \text{Assumed Mean} - \Phi(0.8) \times \text{Standard Error}$$

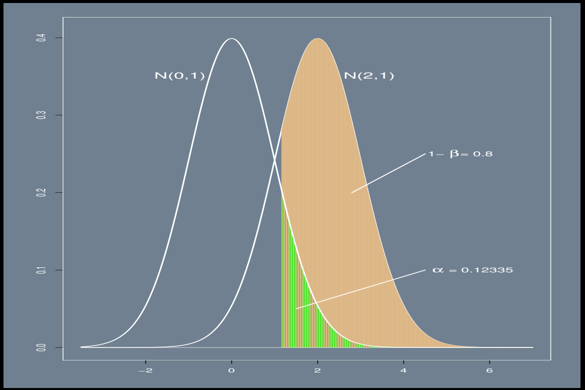
$$\theta_0 + 1.96 \times SE(\bar{y}) = \theta - 0.84 \times SE(\bar{y})$$

(with some rounding).

- And using JHSA:

$$2.8 = 1.96 + 0.84 = \frac{\theta - \theta_0}{SE(\hat{y})}$$

which we will use repeatedly do to atheoretical but overwhelming conventions.



A Key Issue

- *Increasing the sample size decreases the standard error of statistics in proportion to* $1/\sqrt{n}$.
- We usually do not know the underlying variance from the *future* data generating process *for certain.*
- Media survey researchers always have a good estimate of the variance due to repetition.
- Often we have a very good idea from previous work (ours or others).
- If we are concerned with proportions we can exploit the properties of the *Bernoulli PMF* to get a variance estimate.

Sample Size Calculation To Get a Specified Standard Error, Proportions

- Objective # 1: we want to estimate the population proportion, $\hat{p}$, who are technically overweight or obese (BMI >25) specifying a standard error (accuracy) that is no worse than $\sigma_{\hat{p}} = 0.05$.

- Suppose we suspect that the population percentage is around 60%, so the tested effect size is $\hat{p} - p_0$ where $p_0 = 0.6$.

- The standard error of the estimated proportion is

$$\sigma_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/n},$$

  with a mathematical upper bound of $0.5/\sqrt{n}$.

- Using our suspected value (specified effect size), $p_0 = 0.6$,

$$\sigma_{\hat{p}} = 0.05 = \sqrt{0.6(1 - 0.6)/n} = 0.4899/\sqrt{n}.$$

- Rewriting this algebraically means: $n = (0.4899/0.05)^2 = 96$.

- Or we could more conservatively use: $n = (0.5/0.05)^2 = 100$ from the upper bound.

Sample Size Calculations For Power, Proportions

- Objective # 2: we want evidence that more than half the population is overweight or obese ($p \geq 0.5$) using $\hat{p} = y/n$.

- The classical approach using the conservative standard error,

$$\sigma_{\hat{p}} = \sqrt{(0.5)(0.5)/n} = \frac{0.5}{\sqrt{n}}$$

  gives the 95% confidence interval (CI):

$$\left[\hat{p} \pm (1.96)\frac{(0.5)}{\sqrt{n}}\right].$$

- We really care that the lower bound of this CI, L_α, is greater than one-half, so rearranging gives:

$$L = \hat{p} - (1.96)\frac{0.5}{\sqrt{n}} \quad \longrightarrow \quad n = \left(\frac{0.98}{L - \hat{p}}\right)^2,$$

  so at $\hat{p} = 0.55$ we need $n = 384$, and at $\hat{p} = 0.65$ we need only $n = 43$

- This hightlights a big principle: the bigger the tested effect size the small the sample needed.

Sample Size Calculations For Power, Proportions

- What is the *power* of the test that the 95% CI will be completely above the comparison point of 0.5?

- Using simple Monte Carlo simulation in `R` with 10,000,000 draws, hypothesizing $p_0 = 0.6$, and using $n = 96$, we calculate:

```
# SET THE SIMULATION SAMPLE SIZE
m     <- 10000000
# GENERATE m NORMALS WITH MEAN 0.6 AND STANDARD DEVIATION sqrt(0.6*(1-0.6)/96)
p.hat <- rnorm(m,0.6,sqrt(0.6*(1-0.6)/96))
# CREATE A CONFIDENCE INTERVAL MATRIX THAT IS m * 2 BIG
p.ci  <- cbind(p.hat - 1.96*0.5/sqrt(96), p.hat + 1.96*0.5/sqrt(96))
# GET THE PROPORTION OF LOWER BOUNDS GREATER THAN ONE-HALF
sum(p.ci[,1] > 0.5)/m
[1] 0.4997
```

  showing that the probability that the complete CI is greater than 0.5 is 0.4997.

- $\boxed{\textit{Note}}$ that we fixed the sample size (96), fixed the effect size ($0.6 - 0.5$), fixed the significance level ($\alpha = 0.05$), got the standard error by assumption, but let power be realized.

Sample Size Calculations 80% Power, Proportions

- Objective # 3: we want n such that 0.8 of the 95% CIs do not cover 0.5 (80% power).

- First calculate n by solving the equation:

$$\text{Threshold} + 95\%\ \text{CV} \times \text{Standard Error} = \text{Assumed Mean} - \Phi(0.8) \times \text{Standard Error}$$
$$0.5 + 1.96(0.5/\sqrt{n}) = 0.6 - 0.84162(0.5/\sqrt{n})$$

  meaning that (from JHSA) $n = 196.22$ (but rounding up to 197), using the conservative $\sigma_{\hat{p}} = 0.5$, with `qnorm(0.8) [1] 0.84162`, equivalent to $\Phi(0.8) = \int_{-\infty}^{0.84162} f_N(x)dx$.

- R code to check this probability calculation:

```
n     <- 197
m     <- 10000000
p.hat <- rnorm(m,0.6,sqrt(0.5*0.5/n))
p.ci  <- cbind(p.hat - 1.96*0.5/sqrt(n),p.hat + 1.96*0.5/sqrt(n))
sum(p.ci[,1] > 0.5)/m
 [1] 0.8013  # POWER IS A LITTLE BETTER THAN 0.8 BECAUSE WE ROUNDED-UP
```

- $\boxed{\textit{Note}}$ that we fixed the power level ($1 - \beta = 0.8$), fixed the effect size (using 0.6), fixed the significance level ($\alpha = 0.05$), got the standard error by assumption, but let the sample size be realized.

Wait! Where Did You Get That Last Equation?

- The 0.8 CDF of a standard normal is:

  ```
  qnorm(0.8)
  [1] 0.8416212
  ```

  but we can pick other levels obviously.

- We want the scaled difference of the lower bound and the threshold to be equal to the 0.8 CDF:

  $$0.84162 = \frac{L - 0.5}{\sigma/\sqrt{n}}.$$

  which means that there 80% of separation on between these values in terms of a standard normal distribution for the difference.

- Rewriting this gives:

  $$L = 0.5 + 0.84162(\sigma/\sqrt{n}).$$

- Since $L = \hat{p} - z_{\alpha/2}(\sigma/\sqrt{n})$ by definition of a confidence interval for the mean, then:

  $$\begin{aligned} \hat{p} - z_{\alpha/2}(\sigma/\sqrt{n}) &= p_0 + 0.84162(\sigma/\sqrt{n}) \\ 0.6 - 1.96(\sigma/\sqrt{n}) &= 0.5 + 0.84162(\sigma/\sqrt{n}) \\ 0.5 + 1.96(0.5/\sqrt{n}) &= 0.6 - 0.84162(0.5/\sqrt{n}) \end{aligned}$$

Sample Size Calculations With Proportions, *Summary*

- Proportions allow us to exploit the *Bernoulli* variance calculation: $\hat{p}(1-\hat{p})/n$.
- If the goal is a specific *standard error*:
  - ▷ Conservative Approach: $\sigma_{\hat{p}} = 0.5/\sqrt{n} \;\longrightarrow\; n = \left(\frac{0.5}{\sigma_{\hat{p}}}\right)^2$.
  - ▷ If you have more information: $n = \left(\frac{p(1-p)}{\sigma^2_{\hat{p}}}\right)$.
- If the goal is *80% power* to distinguish p from a specified value p_0:
  - ▷ Conservative Approach: $n = \left(\frac{(2.8)(0.5)}{p-p_0}\right)^2$.
  - ▷ If you have more information: $n = p(1-p)\left(\frac{2.8}{p-p_0}\right)^2$.
- Repeated use of 2.8 comes from rounding $1.96_{(z_{\alpha/2=0.025})} + 0.84_{(z_{1-\beta=0.8})}$ in the calculation:

$$\begin{aligned}
0.5 + 1.96(0.5/\sqrt{n}) &= 0.6 - 0.84(0.5/\sqrt{n}) \\
1.96(0.5/\sqrt{n}) + 0.84(0.5/\sqrt{n}) &= 0.6 - 0.5 \\
2.8(0.5/\sqrt{n}) &= 0.1
\end{aligned}$$

Sample Size Calculation for Comparing Two Proportions

- Now the effect size of interest is the difference between two population proportions.

- From Stat-101, the standard error for the difference of proportions is:

$$\sigma_{prop} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \quad \text{or conservatively} \quad \sigma_{prop} = 0.5\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

- Restricting the sample sizes to be equal, $n_1 = n_2 = n$, gives:

$$\sigma_{prop} = 0.5\sqrt{\frac{1}{n/2} + \frac{1}{n/2}} = 1/\sqrt{n},$$

  meaning that $n = 1/\sigma^2_{prop}$.

- Then for $\alpha = 0.05$ and $1 - \beta = 0.8$, with effect size $p_1 - p_2$:
  - ▷ very conservative Approach: $n = [2.8/(p_1 - p_2)]^2$,
  - ▷ or if we have more information: $n = 2[p_1(1-p_1) + p_2(1-p_2)][2.8/(p_1 - p_2)]^2$

  (see G&H page 442).

Sample Size Calculation for Comparing Two Proportions, Example

- We suspect that the proportion of overweight and obese is 10% higher in the US than in Sweden.

- If the surveys are equal size, $n/2$, how big must the total sample size be such that there is 80% power and significance at 0.05, if the true difference in proportions is hypothesized to be 10%.

- So for the 10% to be 2.8 standard errors from zero we need: $n > (2.8/0.10)^2 = 784$.

- What if the true difference in proportions is hypothesized to be 20%?

- Now for the 20% to be 2.8 standard errors from zero we need: $n > (2.8/0.20)^2 = 196$.

- Going the other way, what about a hypothesized 5%?

- Then $n > (2.8/0.05)^2 = 3136$

- $\boxed{Note}$ the important principle that as the hypothesized percent difference (effect size) goes down, the required sample size goes up.

What About Unequal Sample Sizes?

- We have a study where 20% are in the treatment group and 80% are in the control group.

- Using:

$$\sigma_{prop} = \sqrt{\frac{p_1(1-p_1)}{0.2n} + \frac{p_2(1-p_2)}{0.8n}},$$

  which has an upper bound of

$$0.5\sqrt{1/0.2n + 1/0.8n} = 1.25/\sqrt{n},$$

  giving

$$n = (1.25/\sigma_{prop})^2$$

  (from JHSA).

- What sample size is needed to get 80% power to find a 10% group difference?

- Using the upper bound for the standard error, for 10% to be 2.8 standard errors from zero we need:

$$n > [(2.8)(1.25)/0.10]^2 = 1225,$$

  giving $n_1 = 245$, and $n_2 = 980$.

What About Unequal Sample Sizes, More Generally?

- More generally, suppose we have arbitrary proportional sample sizes, q and $(1-q)$.
- Then

$$\sigma_{prop} = \sqrt{\frac{p_1(1-p_1)}{qn} + \frac{p_2(1-p_2)}{(1-q)n}},$$

  which has an upper bound of

$$\sigma_{prop,max} = 0.5[q(1-q)]^{-\frac{1}{2}}/\sqrt{n}.$$

- With a little rearranging, we get:

$$\sqrt{n} = \frac{0.5[q(1-q)]^{-\frac{1}{2}}}{\sigma_{prop,max}}$$

$$n = \left(\frac{[q(1-q)]^{-\frac{1}{2}}/2}{\sigma_{prop,max}}\right)^2.$$

What About Unequal Sample Sizes, More Generally?

- If we consider information on p_1 and p_2 as accurate:

$$n = \left(\frac{[q(1-q)]^{-\frac{1}{2}} / [p_1(1-p_1)(1-q) + p_2(1-p_2)q]^{\frac{1}{2}}}{\sigma_{prop}} \right)^2 .$$

- But this has σ_{prop} on the RHS, which would rely on some information about sample size besides proportional difference.

- So use the *Fleiss Equation* (1981) as an estimate:

$$n \approx \frac{1}{\delta^2} \left[z_{1-\alpha/2} \sqrt{(p_1+p_2)\left(1 - \frac{1}{2}(p_1+p_2)\right)} + z_{1-\beta} \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right]^2 ,$$

where δ is the desired effect size.

Sample Size Calculations With a Single Continuous Outcome

- Continuous outcomes are more challenging than dichotomous outcomes.

- Suppose we want to determine if a population mean θ is statistically distinct from some fixed value θ_0, giving $\delta = \theta - \theta_0$.

- We will use the mean of a sample, $\bar{y}$, to estimate θ for this test.

- Objective #4: estimate the sample size for an effect size of $\delta = \theta - \theta_0$, requiring 80% power and $\alpha = 0.05$ to distinguish δ from zero, using $\bar{y}$ as an estimate of θ.

- Since we do not have the *Bernoulli* trick available anymore with continuous data, we now need one more piece of information: an estimate of the population standard deviation: σ.

- Getting this estimate is often more "art" than sharp procedure, but looking at previous work that you have done, similar data analyzed in your literature, pre-testing, and even just intuition.

Sample Size Calculations With a Single Continuous Outcome

- We can use the fact that $SE(\bar{y}) = \sigma/\sqrt{n}$ (σ a population parameter), to get $n = (\sigma/SE(\bar{y}))^2$.

- For 80% power and $\alpha = 0.05$, a conservative estimate based on a normal approximations is:

$$SE(\bar{y}) = \frac{\theta - \theta_0}{2.8},$$

again using $2.8 = 1.96_{[z_{\alpha/2=0.025}]} + 0.84_{[z_{1-\beta=0.8}]}$ from a normal approximation.

- A more conservative students-t version (say with 5 degrees of freedom) can be calculated by:

```
qt(0.975,df=5) + qt(0.8,df=5)
[1] 3.490126
```

- Therefore:

$$n = \left[\frac{\sigma}{(\theta - \theta_0)/2.8} \right]^2.$$

- So if we have some reasonable estimate for σ, then this is quite simple.

- Notice again that bigger hypothesized $\delta = \theta - \theta_0$ requires smaller samples.

Sample Size Calculations With a Single Continuous Outcome, Example

- The NOV gene, encoding a cysteine-rich glycoprotein NOVH, is suspected to be associated with certain cancers (in renal carcinomas for human prostate cell lines and prostatic tumors, the NOVH gene is overexpressed: the switching on of genes in aging cells).

- For a mouse model we have NOVH concentration from a lean mouse mom's umbilical cord blood as 354 ng/ml.

- The research hypothesis is that mouse moms on a high-fat diet will have a NOVH concentration that is 20% higher.

- So the question is how many mice do we need in the study for power 80% power and $\alpha = 0.05$?

- We need an estimate of σ, which comes from a study by Thibout *et al.* (*Characterization of Human NOV in Biological Fluids: An Enzyme Immunoassay for the Quantification of Human NOV in Sera from Patients with Diseases of the Adrenal Gland and of the Nervous System*, 2003): $\sigma^2 = 18357$.

- Simple R code:

```
( delta <- 354*1.20 - 354 )    [1] 70.8
( sigma <- sqrt(18357*1.5) )   [1] 165.94 #1.5 MULTIPLIER TO BE MORE CONSERVATIVE
( n <- (sigma/(delta/2.8))^2 ) [1] 43.067
```

Sample Size Calculations for Comparison of Two Means, Unequal and Equal Variances

- The standard error of the *difference* of two means is:

$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

- With equal sample sizes, $n_1 = n_2 = n/2$, this simplifies to

$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{2(\sigma_1^2 + \sigma_2^2)/n}.$$

- If we are willing to assume $\sigma_1 = \sigma_2 = \sigma$, then

$$SE_{\bar{y}_1-\bar{y}_2} = 2\sigma/\sqrt{n}$$

  and the total required sample size is:

$$n = (2\sigma/SE_{\bar{y}_1-\bar{y}_2})^2.$$

- If we want 80% power and $\alpha = 0.05$ to detect a difference of δ with equal sample sizes, then the total sample size required is:

$$n = 2(\sigma_1^2 + \sigma_2^2)(2.8/\delta)^2,$$

  which simplifies to $n = (5.6\sigma/\delta)^2$ with equal variances.

Difference of Means, Example Calculations

- From the paper: Muller, O., Becher, H., van Zweeden, A. B., Ye, Y., Diallo, D. A., Konate, A. T., Gbangou, A., Kouate, B, and Gareene, M. (2001). Effect of Zinc Supplementation On Malaria and Other Causes of Morbidity In West African Children: Randomized Double Blind Placebo Controlled Trial. *British Medical Journal* **322**, 1567-1573. (G&H pages 249, 277, 449.)

- The authors provide evidence that adding zinc to the diet of HIV-positive children in West Africa helps to control diarrhea.

- They also find that zinc can improve growth by 0.5 standard deviations ($\delta = 0.5\sigma$), over the control group.

- Suppose we wanted to design a similar study, what is the total required sample size to get 80% power at $\alpha = 0.05$:

$$n = (5.6\sigma/(\delta))^2 = (5.6\sigma/(0.5\sigma))^2 = (5.6/0.5)^2 = 125.44 \approx 126$$

  assuming equal sample variances.

- Using the `samplesize` package in `R`:

```
n.ttest(power = 0.8, alpha = 0.05, mean.diff = 0.5, sd1 = 1.0, sd2 = 1.0)
```

Case Study: Blood Conservation in STL Children's Hospital

- This is an example where the observational data already exists but it is necessary to show that there is enough of it.

- Katherine Steffen, Allan Doctor, Julie Hoerr, Jeff Gill, Chris Markham, Sarah M. Brown, Daniel Cohen, Rose Hansen, Emily Kryzer, Jessica Richards, Sara Small, Stacey Valentine, Jennifer L. York, Enola K. Proctor, Philip C. Spinella. "Controlling Phlebotomy Volume Diminishes PICU Transfusion: Implementation Processes and Impact." *Pediatrics*, Vol. 140(2), 2017.

- We calculate sample estimates separately using the two intended outcome variables: (1) the sum of overdrawn blood and (2) proportion transfused, requiring a minimal effect size of 20% lower levels of these measures than the pre-intervention period in both cases and two-tailed tests.

- For the interval-measured overdrawn outcome (θ in cc/kg/day) we stipulate a conservative estimate of the sample size based on a Student's-t approximation with small degrees of freedom for a 20% difference: $\theta - \theta_0$.

Case Study: Blood Conservation in STL Children's Hospital

- We seek 80% power and $\alpha = 0.05$, so that the standard error is $SE(\bar{y}) = \frac{\theta - \theta_0}{3.08}$ (using $3.08 = 2.23_{[t_{\alpha/2=0.975,df=10}]} + 0.86_{[t_{1-\beta=0.8,df=10}]}$), rearranged to get:
$$n = \left[\frac{\sigma}{(\theta - \theta_0)/3.08}\right]^2$$
using the fact that $SE(\bar{y}) = \sigma/\sqrt{n}$, meaning that $n = (\sigma/SE(\bar{y}))^2$.

- With $\hat{\sigma} = 0.22$ cc/kg/day as found in Valentine and Bateman (2013) and $\theta_0 = 0.498$ is 20% below the preintervention overdraw mean of $\theta = 0.623$ cc/kg/day, the required sample size is:
$$n = \left[\frac{0.22}{(0.623 - 0.498)/3.08}\right]^2 = 29.58 \approx 30.$$

- For the dichotomous outcome of transfusion yes/no we start with the proportion of pre-intervention transfusions as a population estimate, $p = 0.321$, where we want a value that is 20% lower: $\hat{p} = 0.257$, with a standard error of:
$$SE = \sigma_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.191/\sqrt{n}.$$

Case Study: Blood Conservation in STL Children's Hospital

- Again we use conservative Student's-t approximation with small degrees of freedom (10).

- Power of 0.8 and a 95% confidence interval bounded below $p = 0.321$ means that:

$$\int_{0.8}^{+\infty} f(t)dt = \frac{p - H}{SE},$$

where the CI upper bound is $H = \hat{p} + t_{\alpha/2,10}SE$ by definition.

- We can easily get the integral and tail values from:

```
1-qt(0.8,df=10)
[1] 0.12094
pt(0.95,df=10,lower.tail=FALSE)
[1] 0.18225
```

- Substituting into:

$$SE = \sigma_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.191/\sqrt{n}.$$

and solving gives $n = 85.945 \approx 86$.

Ordered Outcomes

- Usually ordered outcomes need an approximation, such as the Wilcoxon-Mann-Whitney test.

- We use the two-sided Wilcoxon test comparing two independent samples where the outcomes are ordered categories.

- The outcome for `R` is a pair of vectors with the proportions, testing if they are different:

```
p <- c(0.25,0.20,0.10,0.45);  q <- c(0.20,0.15,0.15,0.50)
```

- Parameters: $t = n/N$, where n is the size of the second group and N is the total size.

- Running the function for 55% in the first group and 45% in the second group:

```
n.wilcox.ord(power = 0.8, alpha = 0.05, t = 0.45, p, q)
[1] 1446
$m
[1] 795
$n
[1] 651
```

from the `samplesize` package, returning N, n_1, and n_2.

Ordered Outcomes

```
n.wilcox.ord(power=0.8,alpha=0.05,t=0.5,p=rep(0.2,5),q=rep(0.2,5))
$‘total sample size‘          $m                     $n
[1] 2.038e+32                 [1] 1.019e+32          [1] 1.019e+32

p <- c(0.1,0.1,0.1,0.1,0.6)
q <- c(0.6,0.1,0.1,0.1,0.1)
n.wilcox.ord(power=0.8,alpha=0.05,t=0.5,p,q)
$‘total sample size‘          $m                     $n
[1] 23                        [1] 12                 [1] 12

p <- c(0.1,0.1,0.1,0.1,0.4,0.1,0.1)
q <- c(0.4,0.1,0.1,0.1,0.1,0.1,0.1)
n.wilcox.ord(power=0.8,alpha=0.05,t=0.5,p,q)
$‘total sample size‘          $m                     $n
[1] 93                        [1] 46                 [1] 46
```

Required Adjustments for *Multilevel* Power Calculations

- Consider data from cluster sampling: estimating μ by $\bar{y}$ with J equally sized (m) clusters selected at random creating a total sample size of $n = J \times m$.

- If the number of actual clusters in the population is large relative to J and the number of population units in these clusters is large relative m, then:

$$SE_{\bar{y}} = \sqrt{\frac{\sigma_y^2}{n} + \frac{\sigma_\alpha^2}{J}}$$

  where we can get σ_y and σ_α from regression output (such as `lmer` in `R`).

- This can be rewritten using the *intraclass correlation*:

$$SE_{\bar{y}} = \sqrt{\frac{\sigma_{total}^2}{n}[1 + (m-1)\zeta]}, \qquad \zeta = ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}$$

  where $\sigma_{total}^2 = \sigma_\alpha^2 + \sigma_y^2$.

- $\boxed{Note}$ that ICC gives the proportion of total variance that comes from between-group variation and is therefore a rough measure of how group distinctiveness affects the reliability of the mean estimate.

Multilevel Variance Calculations

- Snyders (2005) gives the following structure for J groups/clusters with fixed cluster/group size m, so the total data size is $J \times m$.
- As usual data variance is denoted σ_y^2 and group level variance is denoted σ_α^2.
- Model 1:

$$Y_{ij} = \mu + \beta_{0j} + \epsilon_{ij}$$

(an overall mean, a group-level random effect, and a residual) which gives the following variance:

$$\text{Var}(\hat{\mu}) = \frac{m\sigma_\alpha^2 + \sigma_y^2}{Jm}.$$

Multilevel Variance Calculations

- Model 2:

$$Y_{ij} = \beta_0 + \beta_1\mathbf{X}_{1i} + \beta_2\mathbf{X}_{2i} + \cdots + \beta_k\mathbf{X}_{ki} + U_{ij} + \epsilon_{ij}$$

  (level one variables, a vague definition of a group effect U_{ij}, and a residual) with the assumptions:

  - X_1 is an individual level effect with no group, and is uncorrelated with other level-one variables,
  - σ^2_y is constant in groups (e.g. `lmer`).

  Then:

$$\text{Var}(\beta_\ell) = \frac{\sigma^2_y}{Jm\text{Var}(\mathbf{X}_\ell)}.$$

Multilevel Variance Calculations

- Model 3:
$$Y_{ij} = \beta_0 + (\beta_{01} + \beta_{11}\mathbf{Z}_j)\mathbf{X}_{1ij} + \beta_2\mathbf{X}_{2i} + \cdots + \beta_k\mathbf{X}_{ki} + U_{ij} + \epsilon_{ij}$$
(same as before except now there is a hierarchy on X_1) where the variance of the random slope is τ_1^2. Then:
$$\text{Var}(\beta_{01}) = \frac{m\tau_1^2\text{Var}(\mathbf{X}_1) + \sigma_y^2}{Jm\text{Var}(\mathbf{X}_1)}.$$
and:
$$\text{Var}(\beta_{11}) = \frac{m\tau_1^2 + \sigma_y^2}{Jm\text{Var}(\mathbf{X}_1)}.$$

- Recall that with this is calculated with J equally sized (m) clusters.

Incontinence After Radical Prostatectomy, Survival Modeling

- Objective #5: Use a previous study to help inform a sample size for a survival model.

- The studies look at urinary incontinence as one of the most commonly reported and distressing side effects of radical prostatectomy for prostate carcinoma.

- The key explanatory variables of interest are exercise and obesity status.

- See Kathleen Y. Wolin, Jason Luly, Siobhan Sutcliffe, Gerald L. Andriole and Adam S. Kibel. (February 2010). Risk of Urinary Incontinence Following Prostatectomy: The Role of Physical Activity and Obesity. *The Journal of Urology* **183**, 629-633.

- This is related to some of the 2011-2016 Washington University TREC work, see `http://www.obesity-cancer.wustl.edu/`.

- How can we use this study to calculate sample sizes to go into a grant proposal to perform a similar study?

Incontinence After Radical Prostatectomy, the Data in R

- Recreating the relevant data using R:

```
( obs <- data.frame("freq"=c(8,6,6,10,8,23,21,59),
             "PtDry.12mos"=gl(n=2,k=4,labels=c("no","yes")),
             "Status"=c("obese.inactive","obese.active",
                        "nonobese.inactive","nonobese.active")) )

  freq PtDry.12mos            Status
1    8          no    obese.inactive
2    6          no      obese.active
3    6          no nonobese.inactive
4   10          no   nonobese.active
5    8         yes    obese.inactive
6   23         yes      obese.active
7   21         yes nonobese.inactive
8   59         yes   nonobese.active
```

Incontinence After Radical Prostatectomy, Tabular Summary in R

```
xtabs(freq ~ PtDry.12mos + Status, data=obs)

            Status
PtDry.12mos nonobese.active nonobese.inactive obese.active obese.inactive
        no               10                 6            6              8
        yes              59                21           23              8

summary(xtabs(freq ~ PtDry.12mos + Status, data=obs))

Call: xtabs(formula = freq ~ PtDry.12mos + Status, data = obs)
Number of cases in table: 141
Number of factors: 2
Test for independence of all factors:
        Chisq = 9.8, df = 3, p-value = 0.020
```

Incontinence After Radical Prostatectomy, Subset of interest

- Let's concentrate on the group that is obese and active since it is the most difficult category and therefore more conservative to analyze for sample size calculations, giving the 2×2 table:

Dry at 12 Months	Obese + Active	Obese + Inactive
No	6	8
Yes	23	8

- We will use a *proportional hazards model* for "not dry" with the goal of obtaining $\alpha = 0.05$ and $1 - \beta = 0.8$.

- Recall that the *hazard function* gives the instantaneous hazard (probability) of death at time t given survival until time t.

- Here we model the hazard of being "not dry" instead of the typical context of death.

Incontinence After Radical Prostatectomy, Hazard Models Basics

- From a PDF for the event over time, $f(t)$, define the cumulative distribution function (CDF) of time t and the corresponding *survival function*:

$$F(t) = \int_0^t f(t)dt = p(T < t) \qquad\qquad S(t) = p(T \geq t) = 1 - F(t).$$

The hazard rate is created by:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{p(t \leq T < t + \delta t | T \geq t)}{\delta t} \right]$$

(also called the *hazard function*, the *instantaneous death rate*, the *intensity rate*, and the *force of mortality*).

- These terms are further related by:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)) \qquad\qquad S(t) = \exp(-H(t)), \quad \text{where } H(t) = \int_0^t h(u)du$$

- The model focus is on the ratio of hazards under two different exposures/treatments:

$$HR = \frac{h(t_1)}{h(t_0)}.$$

Incontinence After Radical Prostatectomy, Test Values

- Our null hypothesis to test is $H_0\colon\ p(\text{not dry}|\text{active and obese}) = p(\text{not dry}|\text{inactive and obese})$ at 12 months.

- Looking down the columns of the two-by-two table we get $p_1 = \frac{a}{a+b} = \frac{6}{6+23} = 0.22$ and $p_2 = \frac{c}{c+d} = \frac{8}{8+8} = 0.50$ as empirical evidence of these two quantities.

- Use these to build the *log hazard ratio* under the proportional hazards assumption: $\log(HR) = \log\left[\frac{\log(p_1)}{\log(p_2)}\right]$ as our testable effect size.

- The calculation of n in this context is given by the scaled difference of z-scores that we care about:

$$n = (z_{1-\beta} - z_{\alpha/2})^2 \frac{\hat{\sigma}^2}{\log(HR)^2}$$

  from rearranging:

$$\log(HR) = (z_{1-\beta} - z_{\alpha/2})(\hat{\sigma}/\sqrt{n}),$$

  which is the definitional form for the test.

- We can later scale n for the 45 obese cases to the 96 non-obese cases for the full sample; this provides the "conservative" part of the inference.

Incontinence After Radical Prostatectomy, Getting a Variance

- What we have so far...
  - significance level: $\alpha = 0.05$
  - power level: $1 - \beta = 0.8$
  - parametric form for the test statistic: $\log(HR) = \log\left[\frac{\log(p_1)}{\log(p_2)}\right] = 0.82$ (effect size).
- What we lack... an estimate of σ^2.
- Using the standard approach, $\hat{\sigma}^2 = np(1-p)$, we can plug-in values from the group of interest:

$$\hat{\sigma}^2 = (n_{\text{obese+active}}) \times (p_{\text{obese+active}}) \times (1 - p_{\text{obese+active}}).$$

- Or we could use the conservative estimate:

$$\hat{\sigma}^2 = n_{\text{group of interest}} 0.5(1 - 0.5).$$

- There are many other methods in the survival modeling literature using different assumptions/conditions.

Incontinence After Radical Prostatectomy, Calculations

- Using the functional relationship between sample size and detectable log hazard rate,

$$\log(HR) = (z_{1-\beta} - z_{\alpha/2})(\hat{\sigma}/\sqrt{n}),$$

we can perform these operations in R with some data.

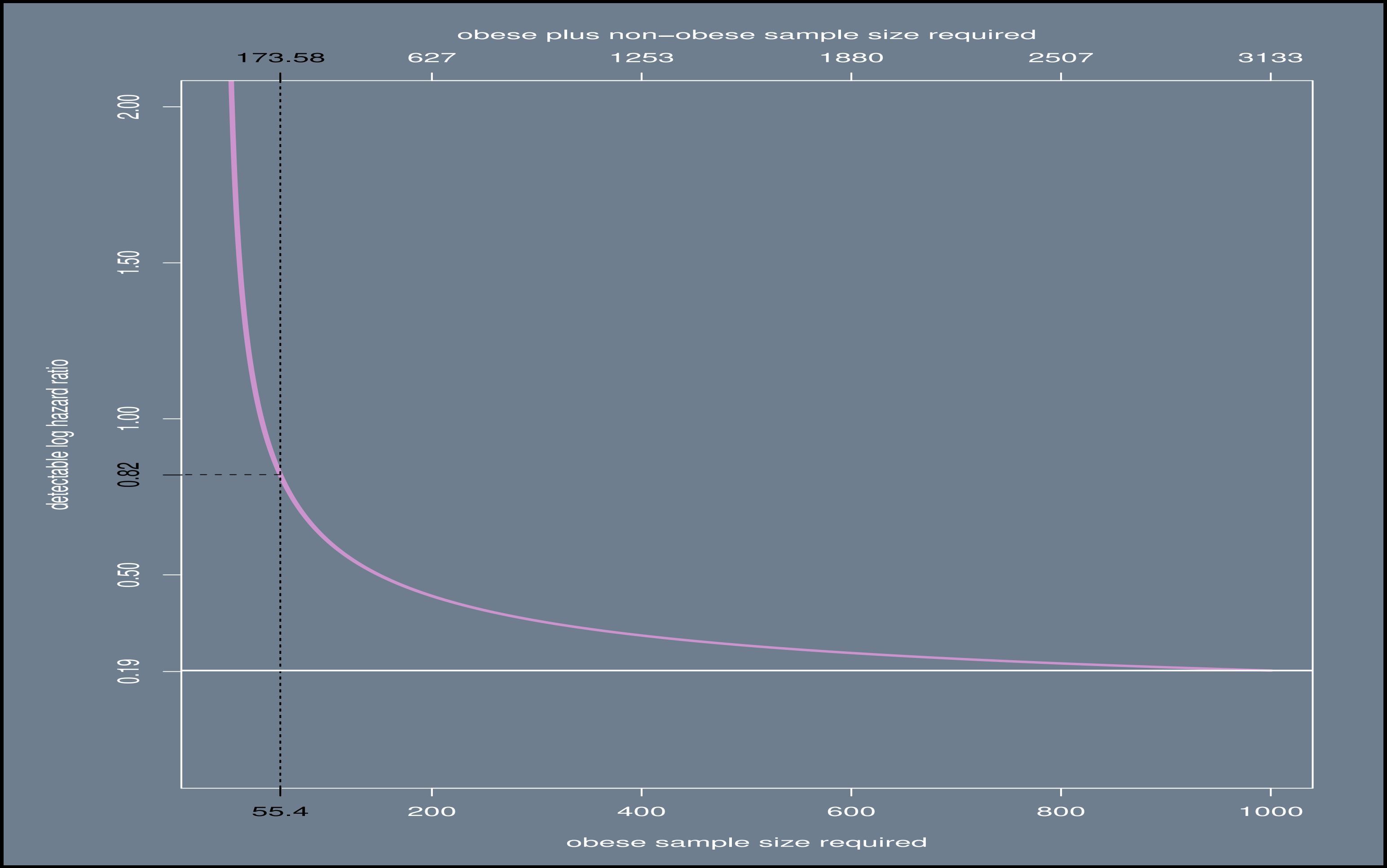
- If we use the previous data/paper:

```
p.1 <- 6/(6+23);   p.2 <- 8/(8+8)
( sigma.sq <- (6+23)*(6/(6+23))*(23/(6+23)) )
    [1] 4.7586
( log.HR <- log(log(p.1)/log(p.2)) )
    [1] 0.82111
( n.1 <- (qnorm(0.8)-qnorm(0.025))^2 * sigma.sq/(log.HR^2) )
    [1] 55.397
( n.needed.obese.nonobese <- n.1*(1+96/45) )
    [1] 173.58
```

where the last statement scales n for the 45 obese cases to the 96 non-obese cases in the full sample

- But what if we want other values? What does the relationship look like graphically?

Incontinence After Radical Prostatectomy



Sample Size Calculation for Superiority, Setup

- The objective is calculate a sample size for the test of superiority of the blood substitute (ErythroMer, EM) over colloid.
- The experiment is setup to hemorrhage 40% blood loss then perform resuscitation by returning the volume back with either of the substitutes.
- The test can be generalized to 50% or 70% blood loss as well until one half of the sample dies to establish a benchmark.
- When this occurs it is equivalent to a "lethal dose 50" (LD50) experiment.
- To obtain a benchmark clinical level we calculate from the output of a logit regression model:

$$\begin{aligned} p(y=1|x) = \frac{1}{2} = [1+\exp(-\beta_0-\beta_1 x)]^{-1} \;&\longrightarrow\; 2 = 1+\exp(-\beta_0-\beta_1 x) \\ &\longrightarrow\; 0 = -\beta_0-\beta_1 x \\ &\longrightarrow\; \widehat{\text{LD50}} = -\hat{\beta}_0/\hat{\beta}_1 \end{aligned}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and coefficient estimate for some explanatory variable.

Sample Size Calculation for Superiority, Setup

- We require a pilot study or similar work to obtain these estimates.

- Using logit regression model 2 from Lee *et al.* (2009, page 126),

$$\hat{\beta}_0 = -15.143, \text{ and } \hat{\beta}_{\text{systolic blood pressure}} = 0.326$$

  (SBP measured blood pressure waves in mmHg).

- This provides $\widehat{\text{LD50}} = 46.451$, which is positive implying an upward sloping logit curve for lower to higher SBP values and therefore supporting the principled comparison of different survival rates based on blood volume status.

- The test is specified by comparing $\lambda_A(t) = 0.5 =$ the death rate with colloid only to $\lambda_B(t) = 0.25 =$ the death rate with EM, for $H_0\!:\!\mu_A < \mu_B$ versus $H_1\!:\!\mu_A \geq \mu_B$ at $t = 48$ hours, meaning that the death rate is lower with EM under the alternative hypothesis.

Sample Size Calculation for Superiority, Survival Model

- Specify random survival time, T, as exponentially distributed such that the survival function is $S(t) = p(T \geq t) = \exp[-\lambda(t)t\exp(X\beta)]$, for some time t.

- We are interested in incorporating the explanatory effects such as the systolic blood pressure into the sample size calculations to make the comparison, but there are no suitable mice findings in the literature to get directly comparable results for inclusion here.

- From Choi *et al.* (2015, p.364) we see that in their similar study using rats the systolic blood pressure for postbleeding is $X = 32.3$ (with a reliable difference from resting in their model; page 364, Table 1).

- Using the logistic regression model for survival on rats with hemorrhagic shock from Lee *et al.* (2009, p.1276) we get a coefficient estimate for this variable, $\hat{\beta} = 0.326$, providing $\exp[X\hat{\beta}]$ to express the effect of SBP directly into our calculation of the hazard ratio (HR).

Sample Size Calculation for Superiority, Basic Calculations

- Assuming such exponential survival times over $t = 48$ hours (time affecting both groups equally), the *hazard rate* for EM versus colloid only is given by:

$$\text{HR} = \frac{\lambda_A(t)t\exp[X\hat{\beta}]}{\lambda_B(t)t\exp[X\hat{\beta}]} = \frac{0.5(48)\exp[10.53]}{0.25(48)\exp[10.53]} = 2$$

  meaning that equal application of covariate contributions does not change the effect of the death rate in the hazard ratio calculation.

- If the effect of other $X\hat{\beta}$ contributions are unequal then this calculation will change, but there is no evidence from rat studies applied here to make this claim.

Sample Size Calculation for Superiority, Final Calculations

- The expected number of events required in each arm of the experiment is then given by:

$$E = (2(Z_{1-\alpha} + Z_{1-\beta})^2)/((\log(\text{HR}))^2)$$

(Machin *et al.* 2011).

- Setting the significance level to $\alpha = 0.05$ for a one-tail test and power to $1 - \beta = 0.8$ gives:

$$E = (2(1.645 + 0.84)^2)/(\log(1.5)^2) = 17.818.$$

- Therefore the sample size required in arm B is:

$$n_B = \frac{2E}{2 - \exp[\log(\lambda_B(2)/\text{HR})] - \lambda_B(2)} = 21.93.$$

- Rounding up, 22 animals are required in the EM group with an additional 22 mice in the colloid only group for balance for a total of 44.

Sample Size Calculation for Non-Inferiority, Mean Comparison

- Now consider the hemorrhage then insertion of the blood substitute ErythroMer (EM) group.

- Using the second results from Terajima *etal.* (2006) we calculate our sample size estimates by comparing the HbV/rHSA (hemoglobin volume to recombinant human serum albumin) value from the mean baseline of 146 (with a standard error of 64) to the final observed time 2 hours later.

- Our objective is to recover the baseline mean across the n animals such that $\mu_A = \mu_B$, where the first term is the baseline and the second term is the goal.

- We perform the same test of non-inferiority with d as 20% of baseline for $d = 29.2$ HbV/rHSA.

- From Terajima *etal.* we set $SE(\mu_A) = 64$ for the baseline and $SE(\mu_B) = 31$ using their study endpoint.

Sample Size Calculation for Non-Inferiority, Mean Comparison

- The sample size for the treatment group is then given by:

$$n_A = \frac{[1.96 \cdot 64 + 0.84 \cdot 31]^2}{d^2} = 26.91193 \approx 27,$$

  again setting the significance level to $\alpha = 0.05$ and power to $1 - \beta = 0.8$.

- Therefore for a balanced experimental design 54 mice are required for 27 in *each* group.

- If we were restricted to 20 $n_A = n_B = 20$, algebraically rearranging the equation above means that power drops dramatically to 0.17.

Calculation of Sample Size for Clinical Trials

- Sample size calculation is important in this context since samples are expensive and time-consuming.
- There may also be ethical issues that affect the calculation by putting *bounds* on the size that are not statistical in nature.
- Other objectives that can affect sample size: superiority, non-inferiority, equivalence, bioequivalence, or precision.
- Endpoint assumptions may be: normal, binary, ordinal, time-to-event.
- See also the Bayesian stopping rules literature.

Calculation of Sample Size for a Single Clinical Trial

- Under the simplest scenario with two groups, treatment and control, we test an *hypothesized* null difference against an *expected* difference to be observed:

$$f(\mu) = \mu_T - \mu_C \qquad S = \bar{X}_T - \bar{X}_C$$

with standard conditions the difference of sample means, S, is normal with variance $\text{Var}(S)$, and:

$$\frac{S - f(\mu)}{\sqrt{\text{Var}(S)}} \sim N(0,1).$$

under the null hypothesis.

- Thus the two-tailed, alpha-level critical region for testing $f(\mu) = 0$ is:

$$|S| > z_{1-\alpha/2}\sqrt{\text{Var}(S)}.$$

- For this critical region to have power, $1 - \beta$, against a specific alternative $f(\mu)$ we need:

$$S - z_{1-\beta}\sqrt{\text{Var}(S)} = z_{1-\alpha/2}\sqrt{\text{Var}(S)}$$

- Rearranging gives:

$$\text{Var}(S) = \frac{S^2}{(z_{1-\beta} + z_{1-\alpha/2})^2}$$

Sample Size Calculation When Population Variance Known

- Knowing the variance, σ^2, (a seemingly rare event) allows non-controversial use of the normal distribution.

- Again S is the observed difference of sample means, scaled by our α and $1-\beta$ criteria:

$$\text{Var}(S) = \frac{S^2}{(z_{1-\beta} + z_{1-\alpha/2})^2},$$

  but this expression is not a function of n.

- Assuming equal variance in the two groups, we get:

$$\text{Var}(S) = \frac{\sigma^2}{n_T} + \frac{\sigma^2}{n_C} = \frac{r+1}{r}\frac{\sigma^2}{n_T},$$

  where $r = n_C/n_T$.

- Combining these statements together and rearranging for the sample size gives:

$$n_T = \frac{(r+1)(z_{1-\beta} + z_{1-\alpha/2})^2\sigma^2}{rS^2}.$$

Sample Size Calculation When Population Variance *Not* Known

- This is the more typical case replacing σ^2 with s^2, and changing the normal to a students-t gives:

$$n_T = \frac{(r+1)(z_{1-\beta} + t_{1-\alpha/2, n_T(r+1)-2})^2 s^2}{rS^2}.$$

  but n_T is on both sides of the equation and cannot be algebraically combined since it is in the degrees of freedom calculation for the students-t.

- Also don't confuse s^2 (the sample standard error) with S^2 (the squared observed mean difference).

- Power is confirmed with redefining in terms of cumulative non-central t:

$$1 - \beta = CDF_t\left(level = 1 - \alpha/2, df = n_T(r+1) - 2, ncp = \sqrt{\frac{rn_T S^2}{(r+1)s^2}} \right)$$

  which can be obtained easily in `R`.

- One approach is to iterate between these forms from some historical starting point.

Sample Size Calculation When Population Variance *Not* Known

- For example with $s^2 = 1$, $D = S = 0.3$, and $r = 1$, we can iterate for a sample size that gets 80% power, with two-tailed alpha at 0.05:

```
s.2 <- 1; d <- 0.3; r <- 1

# NORMAL APPROXIMATION
sample.size <- function(r,alpha,beta,d,S2)
        ((r+1)*(qnorm(beta)+qnorm(1-alpha/2))^2*S2)/(r*d^2)+qnorm(1-alpha/2)
( n <- sample.size(r,0.05,0.8,d,s.2) )
[1] 176.38

power.size <- function(r,alpha,beta,d,S2,N)  {
        DF <- N*(r+1)-2
        NCP <- sqrt((r*N*d^2)/((r+1)*s.2))
        TVAL <- qt(1-alpha/2,DF)
        1-pt(q=TVAL, df=DF, ncp=NCP)
}
power.size(r,0.05,0.8,d,s.2,n)
[1] 0.80138
```

Sample Size Calculation When Population Variance *Not* Known

- Let’s make the conditions less favorable: larger variance ($s^2 = 3$), smaller required effect size ($\delta = 0.1$), and differing sample sizes ($r = 1.5$), then see the effects.

- This makes the sample size blow-up:

```
s.2 <- 3; d <- 0.1; r <- 1.5
( n <- sample.size(r,0.05,0.8,d,s.2) )
[1] 3926.4
power.size(r,0.05,0.8,d,s.2,n)
[1] 0.80012
```

Calculation of Sample Size for a Multiple Clinical Trials

- We have k clinical investigations and we need an estimate of the eventual treatment effect.
- The *overall response across the studies* is estimated by:

$$d_s = \frac{\sum_{i=1}^k w_i s_i}{\sum_{i=1}^k w_i},$$

  where w_i is the precision (inverse of variance) and s_i is the effect size, from study i.

- We need a distributional assumption, such as $s_i \sim N(d_s, w_i^{-1})$, so that:

$$\sum_{i=1}^{k} w_i s_i = N\left(d_s \sum_{i=1}^{k} w_i, \sum_{i=1}^{k} w_i^{-1} \right),$$

  which uses the average within-study variance.

- This allows simple calculation of a CI:

$$d_s \pm z_{1-\alpha/2} \left[\sum_{i=1}^{k} w_i \right]^{-\frac{1}{2}}.$$

Calculation of Sample Size for a Multiple Clinical Trials

- But this does not account for variation between studies, which is given by:

$$\tau^2 = \frac{\sum_{i=1}^{k} w_i(s_i - d_s)^2 - (k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}.$$

- So the corrected precision is:

$$w_i^* = (w_i^{-1} + \tau^2)^{-1}$$

  which is obviously w_i when $\tau^2 = 0$.

- The corrected CI is:

$$d_s \pm z_{1-\alpha/2} \left[\sum_{i=1}^{k} w_i^* \right]^{-\frac{1}{2}}.$$

- This now gives us a reasonable range of effect sizes for sample size calculations.

Calculation of Sample Size for a Multiple Clinical Trials

- For 80% power and $\alpha = 0.05$ to detect a difference of

$$\delta \in d_s \pm z_{1-\alpha/2} \left[\sum_{i=1}^{k} w_i^* \right]^{-\frac{1}{2}}$$

with equal sample sizes, the total sample size required is:

$$n = k \left(\sum_{i=1}^{k} w_i^* \right)^{-1} (2.8/\delta)^2,$$

- Complications: changes due to unequal drop-out rates across studies, heterogeneity of treatment, adaptive designs, varying equipoise, and so on.

Bartlett's Test For Unwanted Heterogeneity In the Studies

- Another estimate of the overall variance across studies is:

$$s_p^2 \approx \frac{\sum_{i=1}^{k} df_i s_i^2}{\sum_{i=1}^{k} df_i}$$

  where $df_i = n_i - 1$.

- Define:

$$M = \left(\sum_{i=1}^{k} df_i\right) \log(s_p^2) - \sum_{i=1}^{k} df_i \log(s_i^2)$$

  and:

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{df_i}\right) - \frac{1}{\sum_{i=1}^{k} df_i}\right].$$

- The test for heterogeneity is then:

$$\frac{M}{C} \sim \chi^2_{k-1}$$

  where tail values indicate differences.

Models with Covariates

- Most models of course have covariates and this affects sample size calculation.

- Often this is done in the context of *baseline measures*: demographic and clinical variables designed to affect some outcome.

- Frison and Pocock (1992) give an estimated variance correction:

$$\sigma^2_{bl} = \sigma^2 \left[1 - \frac{k\rho^2}{1 + (k-1)\rho} \right],$$

  where ρ is the conventional Pearson's product moment correlation coefficient between observations, and k is the number of baseline measures taken on each case.

- Higher levels of correlation between cases mean lower variance.

- For fixed correlation the variance reduction tapers off after three.

Effects Over Time

- Drop-rates generally increase over time in longitudinal studies.

- If we have an estimated of the drop-out rate from previous studies, then we can degenerate all of our sample size calculations to the final wave, as a conservative approach.

- But new post-treatment measures increase knowledge and decrease variance.

- Suppose we are testing a difference of means between two groups post-treatment.

- The new corrected variance estimate is:

  $$\sigma^2_{pd} = \frac{\sigma^2[1 + (r - 1)\rho]}{r},$$

  where r is the number of post-treatment measures, and ρ is the correlation between these new measures.

- Higher levels of correlation between variables mean higher variance.

Clustering

- Cluster sampling leads naturally to multilevel models as the groups come to the analyst immediately identified.

- *One Stage* cluster sampling is when all units in a cluster are measured.

- *Two Stage* cluster sampling is when a sample is taken within the clusters.

- Treatments can be applied to an entire cluster or to individuals within clusters, where the application to individuals gives more accurate estimates.

Cluster Randomized Trials

- It is often the case that patients are clustered in hospitals, clinics, or primary care practices.

- This is a motivation for multilevel models of course.

- If there is between cluster variation, it reduces the effective sample size and gives lower power than intended.

- The key is getting an estimate of the *intracluster correlation coefficient* between observations at the same level:

  $$\zeta = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_y}.$$

  which is obviously another name for the intraclass correlation coefficient where clusters are classes.

- Here ζ gives the proportion of total variance that comes from between-group variation and is therefore a rough measure of how group distinctiveness affects the reliability of the mean estimate.

Cluster Randomized Trials

- The cluster corrected variance estimate is:

$$\sigma^2_{icc} = \frac{\sigma^2[1+(m-1)\zeta]}{cm},$$

  where ζ is the intracluster correlation coefficient, c is the number of clusters, and m is the average sample size per cluster.

- Now the sample size per arm goes from:

$$n_T = \frac{2(z_{1-\beta}+z_{1-\alpha/2})^2\sigma^2}{rS^2}.$$

  to:

$$n_T = \frac{2(z_{1-\beta}+z_{1-\alpha/2})^2\sigma^2[1+(m-1)\zeta]}{rS^2cm}.$$

- The numerator difference, $[1+(m-1)\zeta]$, is called the *variance inflation factor*, and when cluster sizes are unequal this is $[1+((cv^2+1)\tilde{m}-1)\zeta]$, where cv is the coefficient of variation (standard deviation over the mean), and $\tilde{m}$ is the average cluster size.

Cluster Randomized Trials

- Also to estimate the number of needed clusters per arm can be calculated:

$$c = \frac{n_T([1 + (m - 1)\zeta]}{m}$$

  for equal cluster sizes, and

$$c = \frac{n_T([1 + ((cv^2 + 1)\tilde{m} - 1)\zeta]}{m}$$

  for unequal cluster sizes.

Recap

- The goal is to achieve desired: significance levels, effect sizes, subsetting ability, and power by collecting a sufficiently large sample.

- These criteria compete with each other.

- In some settings the statistician does not participate in the design of the study, so others need to be able to perform efficacious sample size calculations.

- Automated sample size calculators are okay, if you understand the theoretical underpinning, but don't *rely* upon them.

- This is all really about stipulating reasonable and defensible assumptions on the *future* data-generating process, and is therefore fictional..