BST 210 Lab: Week 14
Sample Size and Power Calculations

So far, we've spent a significant amount of time focusing on and grappling with how we analyze data from observational studies or clinical trials that have already been planned and conducted. But how we design these studies—from how many patients we decide to recruit, to how long we follow those patients, to what covariate information we collect, to what analyses we conduct at the end of the study—can have profound impacts on what effects we are able to detect, and what conclusions we are able to draw. When we design our study, we need to think carefully about issues of **sample size** and **power**!

Some important ideas and concepts include:

- **The Type I error rate (α):** Probability of rejecting the null hypothesis, given that the null hypothesis is in fact true. In other words, it's the probability of a false positive.
- **The Type II error rate (β):** Probability of failing to reject the null hypothesis, given that the alternative hypothesis is in fact true. In other words, it's the probability of a false negative.
- **Power ($1-\beta$):** Probability that we correctly reject the null hypothesis when the alternative hypothesis is true.

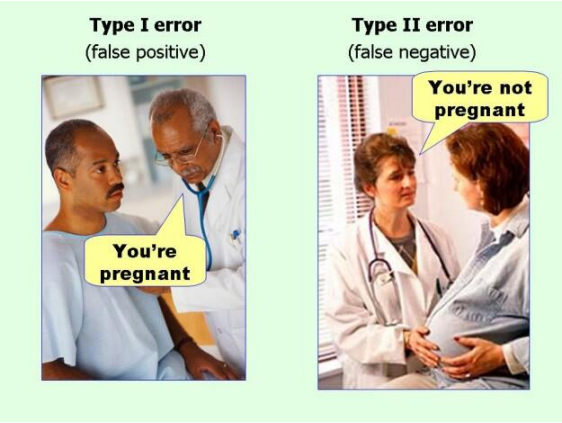

Image Credit: Paul D Ellis, "Effect Size FAQs." http://www.effectsizefaqs.com

Ideally, we'd like to design a study with a small, well-controlled type I error rate and a large power to detect our effect of interest. This is why we usually conduct sample size calculations: to determine how large of a study we need in order to guarantee a certain power!

What are some of the consequences of designing a study with too small a sample size?

Given this, why don't we just design our studies to include as many subjects as possible?

In practice, studies are constrained by their budgets, the size of the support and clinic staff helping to run them, and time (among other things) and so we sometimes only have a fixed sample size to work with. In that case, we can perform power calculations to determine what we can reasonably expect from our study.

A Brief Mathematical Interlude

In order to conduct power and sample size calculations, we typically need information about (1) our type I error rate, (2) our desired power (for sample size calculations) or sample size (for power calculations), and (3) the effect size under the alternative hypothesis that we'd like to detect. The exact method of calculating the sample size and power changes depending on the type of data we're working with, as well as the analysis we plan to conduct at the end of the study:

One-Sample Test of Binomial Proportions

Suppose that we want to test the null hypothesis $H_0 : p = p_0$ versus the alternative hypothesis $H_1 : p \neq p_0$, and that we're particularly interested in detecting whether $p = p_1$ under the alternative. Let α be the type I error rate that we've fixed for our study. Then:

Power Calculation

If we have a fixed sample size of n, then our power to detect $p = p_1$ is given by

$$\text{Power} = 1 - \beta = \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\left(z_{1-\alpha/2} + \frac{\sqrt{n}|p_0 - p_1|}{\sqrt{p_0(1-p_0)}}\right)\right).$$

The function $\Phi(\cdot)$ refers to the standard normal cumulative distribution function, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile from a standard normal distribution!

Sample Size Calculation

Now suppose that we've fixed our desired power at $1 - \beta$. Then in order to detect $p = p_1$ with this power, our necessary sample size is

$$n = \frac{p_0(1-p_0)\left(z_{1-\alpha/2} + z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{(p_1 - p_0)^2}.$$

How do you think the required sample size changes (if it all) when our null proportion and alternative effect of interest are close to 0.5 (ex: $p_0 = 0.4$ and $p_1 = 0.6$) versus when they are farther away from 0.5 (ex: $p_0 = 0.1$ and $p_1 = 0.3$)? Note that the variance for an estimated binomial proportion is $\hat{p}(1-\hat{p})/n$, and that this variance is largest for $\hat{p}$ close to 0.5.

The power and sample size formulas for two-sample tests of proportions are given in the lecture slides.

Log-Rank Test for Survival Outcomes

Suppose that we want to compare two survival curves, one corresponding to patients given some intervention of interest, and the other to patients given the standard-of-care. We plan on conducting a two-sided log-rank test with type I error α and power $1-\beta$, and we want to be able to detect a hazard ratio of size Δ. Finally, suppose we want to design our study so that there are k times as many patients in the intervention arm as in the standard-of-care arm. Then the total number of observed events we need is

$$d = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (k\Delta + 1)^2}{k(\Delta - 1)^2}.$$

Some General Patterns in Sample Size and Power Calculations

In general, what do you think will happen to the necessary sample size if we:

- *Increase our alternative effect size of interest? Decrease our alternative effect size of interest?*

- *Increase our desired power? Decrease our desired power?*

Similarly, what do you think will happen to the power of our study if we:

- *Increase our alternative effect size of interest? Decrease our alternative effect size of interest?*

- *Increase our sample size? Decrease our sample size?*

Once we start to consider issues such as study drop-out and non-adherence, or more involved analysis methods such as regression modeling, the actual process of calculating power and sample size can become quite complicated. But the motivation behind the calculations is always the same: we just want to make sure that—if there really is a clinically meaningful association between our outcome and our predictor of interest—we are able to design a study that can actually detect this association!

Note: we determine the sample size needed and assess the power of our trial design *before* actually running the trial. Once our study has been conducted, our sample size either is or is not sufficient, and we either will or will not reject our final hypothesis test. In other words, it doesn't make sense to discuss the power of a trial that has already been conducted!

An Example: Designing a Study for Depression Interventions

Suppose that we're piloting a new intervention that provides at-risk teenagers with free mental health counseling, and that we want to assess to what extent this program reduces the risk of depression. We design

a study with two groups: one group that receives the free counseling, and one group that does not. Our primary outcome of interest is whether or not a teenager reports having depressive symptoms one year after the start of the study (yes/no). Suppose that the prevalence of depressive symptoms is 60% in our population of interest, and that we will consider our program to be a success if it reduces this proportion to 40%.

Suppose 100 teenagers volunteer to participate in your study, and that you plan to assign 50 of them to each of the two study arms. *Given the output below, what is the power to detect a difference of 60% versus 40%, assuming we conduct a two-sided test at the* $\alpha = 0.05$*-level?*

```
# Determining power to detect p0 = 0.6 versus p1=0.4
power.prop.test(n = 50, p1 = 0.6, p2 = 0.4, sig.level = 0.05)
```

```
     Two-sample comparison of proportions power calculation

              n = 50
             p1 = 0.6
             p2 = 0.4
      sig.level = 0.05
          power = 0.5162969
    alternative = two.sided

NOTE: n is number in *each* group
```

Often times, researchers want to design their studies to achieve powers of 80%-90%. *Using the output below, what sample size would we need per group in order to obtain 90% power?*

```
# Determining the sample size needed to achieve 90% power
power.prop.test(p1 = 0.6, p2 = 0.4, sig.level = 0.05, power = .90)
```

```
     Two-sample comparison of proportions power calculation

              n = 129.2529
             p1 = 0.6
             p2 = 0.4
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Since we're primarily interested in determining whether or not the free mental health counseling actually *improves* overall mental health, we might consider using a one-sided test instead. *What are the differences*

between using a one-sided and a two-sided test? And how do our previous two answers change?

```
# Determining power and sample size for a one-sided test
power.prop.test(n = 50, p1 = 0.6, p2 = 0.4, sig.level = 0.05,
                alternative = "one.sided")
power.prop.test(p1 = 0.6, p2 = 0.4, sig.level = 0.05, power = .90,
                alternative = "one.sided")
```

```
     Two-sample comparison of proportions power calculation

              n = 50
             p1 = 0.6
             p2 = 0.4
      sig.level = 0.05
          power = 0.6414995
    alternative = one.sided

NOTE: n is number in *each* group
```

(a) Power with $n = 50$ in each group.

```
     Two-sample comparison of proportions power calculation

              n = 105.1622
             p1 = 0.6
             p2 = 0.4
      sig.level = 0.05
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group
```

(b) Sample size needed for 90% power.

When designing a study involving at-risk teenagers, we would reasonably expect there to be some loss to follow-up: participants may not follow through with the counseling appointments, or may not return for their one-year assessment. If this is the case, then our previous sample size calculations underestimate the true sample sizes we need!

Suppose that 20% of the participants in each study arm are expected to drop out, but that the probability of dropping out is assumed to be unrelated to the outcome of interest (depressive symptoms). *In that case, how would we perform our primary analysis, and what sample size would we need per group in order to obtain 90% power (assuming a two-sided test)?*

Do you think the assumption that we just made (that drop-out is unrelated to the outcome of interest) is reasonable? Why or why not?

Finally, suppose that another member of the research team suggests that—rather than randomizing the same

number of participants to both arms—we allocate twice as many patients to the counseling arm than to the no intervention arm. *How might this affect our power, assuming that our sample size is fixed at 261 (the sample size we previously needed to achieve 90% power, modified so that it's divisible by 3)?*

```
# We'll need a new package in order to specify an unequal allocation ratio
library(Hmisc)

# Determining power given a 1:2 allocation ratio and a sample size of 261
bpower(p1 = 0.4, p2 = 0.6, alpha = 0.05, n1 = 174, n2 = 87)

# What would the sample size need to be now if we want to maintain 90% power?
bsamsize(p1 = 0.6, p2 = 0.4, fraction = 1/3, alpha = 0.05, power = .90)
```

```
    Power
0.8671882
```

(a) Power assuming 1:2 allocation ratio with $n = 261$.

```
      n1       n2
96.67695 193.35389
```

(b) Sample size needed for 90% power.

A Second Example: Power Calculations for Logistic Regression

Let's return to the Global Longitudinal Study for Osteoporosis in Women (GLOW), which we first saw all those weeks ago in lab! (You can find the GLOW data on Canvas in the file `glow.csv`.) In particular, let's suppose that we want to design a companion study: the Global Longitudinal Study for Osteoporosis in Older Men (GLOOM). Our primary outcome of interest is still whether or not a participant has a fracture within the first year of follow-up, and suppose that our primary exposure of interest is age at baseline. We also plan to collect information on BMI and prior history of fracture.

Suppose that our analysis plan is to fit the following logistic regression

$$\text{logit}(p_{fracture}) = \beta_0 + \beta_1 \cdot \texttt{age} + \beta_2 \cdot \texttt{BMI} + \beta_3 \cdot \texttt{priorfrac},$$

and to then test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. We want to make sure that we design GLOOM to have a reasonable power (say, 80%) to detect a clinically meaningful association between age and risk of bone fracture.

What information would we need to specify and/or calculate in order to determine our necessary sample size?

Luckily, we can use the results of the previous GLOW study to guide our choices for our effect size of interest, and for the prevalence of the outcome at the mean covariate values; it can also help us determine an appropriate R^2 value based on the linear regression of age on BMI and history of prior fracture!

```
# Function that helps us calculate fitted probabilities more easily
expit <- function(x){
  return (exp(x)/(1 + exp(x)))
}

# Reading in the dataset
glow <- read.csv("glow.csv")
summary(glow)

# Determining the mean covariate values in our dataset,
# as well as the standard deviation for age
mean.age <- mean(glow$AGE); sd.age <- sd(glow$AGE)
mean.bmi <- mean(glow$BMI); mean.priorfrac <- mean(glow$PRIORFRAC)

# Fitting both the unadjusted and adjusted regression models
mod.2 <- glm(FRACTURE ~ AGE + BMI + PRIORFRAC, family = "binomial", data = glow)

# Calculating the fitted probabilities for a one standard deviation increase in age
p1 <- (expit( coef(mod.2)[1] + mean.age*coef(mod.2)[2] +
               mean.bmi*coef(mod.2)[3] +
               mean.priorfrac* coef(mod.2)[4]))
p2 <- (expit( coef(mod.2)[1] + (mean.age + sd.age)*coef(mod.2)[2] +
               mean.bmi*coef(mod.2)[3] +
               mean.priorfrac* coef(mod.2)[4]))

# R^2 between age, BMI, and history of prior fracture
summary(lm(AGE ~ BMI + PRIORFRAC, data=glow))$r.squared
```

```
"p1       p2"
"0.234    0.315"
```

(a) Fitted fracture probabilities from GLOW.

```
"R^2"
"0.134"
```

(b) R^2 for the regression of age on BMI and prior fracture.

Using the above values, we can then determine the necessary number of patients to enroll in GLOOM if we want 90% power. Please see

`Lab14_EL_R`

for code representing the method presented in "A simple method of sample size calculation for linear and logistic regression" by F. Y. Hsieh et al. Here also is a means for doing it in Stata using the information above:

```
* Determining power and sample size for GLOOM
powerlog, p1(0.234) p2(0.315) rsq(0.134)
```

```
Logistic regression power analysis
One-tailed test: alpha=.05  p1=.234  p2=.315  rsq=.134  odds ratio=1.505334081
> 976417

power          n
0.60         166
0.65         189
0.70         215
0.75         245
0.80         281
0.85         325
0.90         386
```

So in order to achieve 90% power, we'll need to recruit 386 individuals into our study!

How would the above sample size results change if we had larger R^2 values (i.e., if the covariates BMI and prior history explained a larger proportion of the variability in age)? If we had smaller R^2 values? Why?

```
* Examining how the results change for different R^2 values
*         R^2 = 0
powerlog, p1(0.234) p2(0.315)
*         R^2 = 0.5
powerlog, p1(0.234) p2(0.315) rsq(0.5)
```

power	n
0.60	144
0.65	164
0.70	186
0.75	212
0.80	243
0.85	282
0.90	334

(a) Power and sample size calculations with $R^2 = 0$.

power	n
0.60	288
0.65	328
0.70	373
0.75	425
0.80	486
0.85	563
0.90	668

(b) Power and sample size calculations with $R^2 = 0.5$.

(Yet Another) Example: Power Calculations for Survival Models

Finally, let's consider the chemotherapy data that we've been working with in class. The data are from a small pilot study of 23 chemotherapy patients currently in remission, and are available on Canvas under the Lab section for Week 14, in the file `chemotherapy.dta`. In this study, 11 patients were randomized to receive maintenance chemotherapy while in an effort to prolong remission time, and the remaining 12 were randomized to no intervention. The outcome of interest was time until relapse (in weeks).

Since there are only 23 participants and 18 events, the log-rank test comparing the time-to-relapse between the two groups fails to reach significance, and our estimate for the association between maintenance chemotherapy and relapse is quite wide:

```
# Reading in the data
chemo <- read.dta("chemotherapy.dta")

# Hazard ratio and CI for the pilot study
mod.3 <- coxph(Surv(time, relapse)~group,data=chemo)
summary(mod.3)
```

```
Call:
coxph(formula = Surv(time, relapse) ~ group, data = chemo)

  n= 23, number of events= 18

         coef exp(coef) se(coef)      z Pr(>|z|)
group -0.9155    0.4003   0.5119 -1.788   0.0737 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
group    0.4003      2.498    0.1468     1.092

Concordance= 0.619  (se = 0.073 )
Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.06581
Wald test            = 3.2  on 1 df,   p=0.07371
Score (logrank) test = 3.42  on 1 df,   p=0.06454
```

We want to turn our pilot study into a full study, but aren't quite sure (1) what effect size we should power our study for, and (2) how many events we need in order to achieve a reasonable power. Let's think through this problem, and then try running some power and sample size calculations to help us actually design our study!

Based on everything we've seen and discussed so far, how do you think the power of the log-rank test will change for different hazard ratios between 0.15 and 1.1, assuming we keep the number of events fixed?

Now suppose that we've fixed the power at 80%. How will the necessary sample size (i.e., number of events) change under different values in our set of plausible hazard ratios?

We can use the following code to implement these calculations in R:

```
# Determining the power for a fixed sample size and an array of HRs
for (hr in c(0.15, 0.3, 0.4, 0.7, 0.9, 1.01, 1.1)){
  print(paste("HR:", hr, "N:", 9))
  print(paste("Power:",
              round(powerCT.default(nE = 9, nC = 9, pE = 1, pC = 1, RR = hr),5) ))
}
```

```
# Determining the sample size for a fixed power and an array of HRs
for (hr in c(0.15, 0.3, 0.4, 0.7, 0.9, 1.01, 1.1)){
  print(paste("HR:", hr, "Power:", 0.8))
  print(round(ssizeCT.default(power = 0.8, k = 1, pE = 1, pC = 1, RR = hr,
                              alpha = 0.05),5))
}
```

```
[1] "HR: 0.15 N: 9"
[1] "Power: 0.88018"
[1] "HR: 0.3 N: 9"
[1] "Power: 0.62723"
[1] "HR: 0.4 N: 9"
[1] "Power: 0.44366"
[1] "HR: 0.7 N: 9"
[1] "Power: 0.1129"
[1] "HR: 0.9 N: 9"
[1] "Power: 0.04122"
[1] "HR: 1.01 N: 9"
[1] "Power: 0.02626"
[1] "HR: 1.1 N: 9"
[1] "Power: 0.03938"
```

(a) Power calculations.

```
[1] "HR: 0.15 Power: 0.8"
[1] 8 8
[1] "HR: 0.3 Power: 0.8"
[1] 14 14
[1] "HR: 0.4 Power: 0.8"
[1] 22 22
[1] "HR: 0.7 Power: 0.8"
[1] 127 127
[1] "HR: 0.9 Power: 0.8"
[1] 1417 1417
[1] "HR: 1.01 Power: 0.8"
[1] 158552 158552
[1] "HR: 1.1 Power: 0.8"
[1] 1731 1731
```

(b) Sample size calculations.

Our sample size calculations are all in terms of the number of events that we need to observe during the course of our trial. But the catch is that—when designing a study—we need to account for the fact that some patients in our study may be lost to follow-up, or may not experience the event (relapse) before the end of the trial. The withdrawal/censoring rate in the pilot study was approximately 20%. *If our main study has the same 20% censoring rate as the pilot study, how will this affect our sample size calculations? (Hint: recall from lecture the fraction* $\frac{1}{1-\lambda}$*)*

Note: Power and sample size calculations for a Cox proportional hazards model with a single binary covariate are exactly the same as the calculations for the two-group log-rank that we did above. See lab code for example code for calculating power and sample size for Cox proportional hazards models with a continuous covariate.