Sample size justification

Steve Simon, https://github.com/pmean/papers-and-presentations/blob/master/sample-size-justification/2023-02-talk.pdf

Created 2022-02-02

Who am I?

Steve Simon

- PhD Statistics, 1982, U Iowa
- Teach in Biomedical and Health Informatics
  - Previous jobs at CMH, CDC
- Part-time independent statistical consultant (P.Mean Consulting)
- Married to a Pediatric Cardiologist (retired)
- Run 5K and 4 mile races

Let me introduce myself. I am Steve Simon. I got a PhD in Statistics almost 40 years ago from the University of Iowa. The world has changed a lot since then, but I have tried to keep up. Today, if you want to sound trendy, you are a "data scientist". I teach in the Department of Biomedical and Health Informatics at UMKC. I have had previous jobs at Children's Mercy Hospital, and the Centers for Disease Control and Prevention. I'm also a part-time statistical consultant. I have a sole proprietorship, P.Mean Consulting. That's short for Professor Mean. For people who don't get the joke, I point out that Professor Mean is not just your average Professor. Related to this talk, I should point out that I am obsessed with computers.

Outline of this talk

- Pop quiz
- Sample size justification is an economic justification
- Definitions
- Specifying a research hypothesis
- Identifying the variation in your outcome
- Determining the minimum clinically important difference
- Repeat of pop quiz

We'll start with a pop quiz. Don't worry, it's not graded.

Pop quiz, question #1

- A good sample size will produce
  (1) Large values for both alpha and beta.
  (2) A large value for alpha and a small value for beta.
  (3) A small value for alpha and a large value for beta.
  (4) Small values for both alpha and beta.
  (5) I'm awfully glad I'm a Beta, because I don't work so hard.
  (6) I don't know the answer.

Does anyone know the reference to the quote in #5. It is from Aldous Huxley's Brave New World, a dystopic book about a future where people are bred into classes of alphas, betas, and gammas. And if you were a beta you were brainwashed into believing that you should be glad to be a beta.

Pop quiz, question #2

- One of the three things you need to calculate an appropriate sample size is
  (1) A confidence interval for your outcome variable
  (2) A range for your outcome variable
  (3) A standard deviation for your outcome variable.
  (4) A standard error for your outcome variable.
  (5) Any of these is fine.
  (6) I don't know the answer.

Here's the second question.

Pop quiz, question #3

- The minimum clinically important difference is determined by
  (1) Finding a balance between the benefits and the harms of a new drug.
  (2) Finding a balance between the cost of sampling an additional patient and the incremental reduction in uncertainty.
  (3) Finding a balance between Type I and Type II error rates.
  (4) Finding a balance between your work and your family.
  (5) More than one answer above is correct.
  (6) I don't know the answer.

Here's the third question. If you don't know the answers to these questions now, that's okay. But if you don't know the answer after I finish my lecture, then I have not done a good job.

A sad tale of research

- A researcher is finishing up a six year, ten million dollar NIH grant and writes up in the final report
  - "This is a new and innovative surgical procedure and we are 95% confident that the cure rate is somewhere between 3% and 98%."

Here's a story I tell all the time when someone comes in and asks me what their sample size should be. It's one of those stories that isn't true but should be. When I tell this story to someone, it doesn't matter how much or how little they know about Statistics. They still get the punchline. Ten years and six million dollars and the best you could do is come up with a confidence interval that goes from 3% to 98%? Good grief! That's a huge waste.

The heart and soul of all sample size calculations is economic. You want to insure that your research dollars are well spent, that you are getting something of value for your investment.

A second sad tale of research

- Confidence interval for an odds ratio: 0.82 to 3.14.
  - Interpret this interval

Here's a more subtle example that I typically include in a talk about confidence intervals. Recall that an odds ratio of 1 implies that the control and treatment arm have the same risk.

Now most people I show this interval to look at with a blank expression. No! This is something that should make you gag and retch. This is a terrible interval.

It includes the value of 1 which means no change in risk. But it also includes the value of 2, which means a doubling of risk. So you are looking at a research study with such a pathetically small sample size that it cannot distinguish a difference between no change in risk and a doubling of risk.

No, it's worse than that. This study can't distinguish between no change and a tripling of risk.

Sample size justification is an economic justification

- Larger samples cost more money
  - but provide more precision
  - diminishing returns
- Balance cost versus precision

There's a tradeoff between cost and precision. You want to spend as little money as possible, but not so little that you get a confidence interval for a cure that goes from 3% to 98% or for an odds ratio that goes from 0.82 to 3.14. So keep adding money to pay for a larger sample size until you get to the point where any additional money spent provides such a small improvement in precision that it is not worth your effort anymore.

Real example of economic tradeoff

- Overbilling problem
  - Need to estimate excess to process a refund
- Audit of all records too expensive
  - 2,000 records, $100 per record
- Sample how many records?

Sometimes it is hard to put a dollar figure on precision, but here's a simple example where you can see the relationship.

Solution

- Estimate 95% confidence interval for excess
  - Pay the upper limit
- Balance size of upper limit versus cost

The solution I proposed was to pay not the estimated excess charges but the upper limit of the 95% confidence interval. It seems fair for us to bear the cost of the uncertainty associated with our failure to do a 100% audit.

So now, we have a good idea of when to stop sampling. Say that it costs 100 dollars to add a new record to the sample, but the interval shrinks by 200 dollars. That's more than enough justification to increase the sample size.

But these calculations have a diminishing return because the interval width is inversely proportional to the square root of the sample size. So you might get to the point where an additional record only buys you a 50 dollar reduction in the upper confidence limit. This means you've gone too far.

Solution displayed graphically

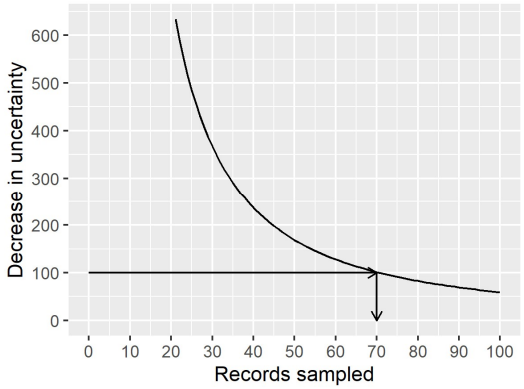

Figure 1. Decrease in uncertainty versus sample size

This graph shows the decrease in uncertainty associated with the 95% confidence interval. Anything less than a sample size of 70, and the gain in precision more than offsets the cost of additional sampling. But once the sample size increases beyond 70, diminishing returns take over and a further increase in sample size is not justified.

The curve in the previous graph shows the general concept of diminishing returns. With a small sample size, each additional record has a large payoff in reduction of uncertainty. But as the sample size increases, the incremental payoff grows smaller. Eventually, the incremental benefit of further reducing uncertainty is counterbalanced by the incremental increase in sampling costs.

In most examples, there is not a one-to-one relationship between the shrinkage of the confidence interval and the saving of money. But the principle still applies.

– Your current level of uncertainty is costing you money.

– Sampling can reduce that uncertainty, but the sample itself costs money.

– You have to balance the benefits in reduction of uncertainty against the costs of

sampling

Ethical concerns about an inadequate sample size

People volunteer for research studies for three reasons:

1. To earn some money,
2. To find out more about the research process, or
3. To help other people.

An inadequate sample size invalidates third reason

The motivation for people to volunteer for a clinical trial fall into three broad categories. First, they may be in it for the money. You shouldn't trivialize this. The get up early and sacrifice their opportunitiy to watch their favorite Saturday morning cartoons. Especiall that weird one: Sponge Something Square Bob? It is perfectly fine to compensate them for their time, as long as the amount is not excessive.

Other people volunteer out of curiosity. They want to learn more about how medical research works.

But perhaps the biggest reason is that people volunteer because they want to help people in the future who have the same disease. They want to make the world safer and healthier.

If the research study has such a small sample size that the results are uninterpretable then you have helped no one. You have broken an implicit promise to those volunteers. Almost no one would volunteer for a study if the only value in it is the ability for you to pad your own resume.

Ethical concerns about an excessive sample size

Features of some (not all) clinical trials

- Small/moderate amount of pain
- Endure a risky procedure, or
- Forgoing an appropriate medical treatment

In these cases, too large a sample is problematic.

Too large a sample size is also an ethical problem.

Too large a sample size creates needless suffering among research volunteers. Not every study requires suffering and the amount that we ask people to endure has to be small or moderate. But sometimes there is some pain involved (for example, through a blood draw) or you ask the volunteers to undergo a procedure with small risks (such as a lumbar puncture) or if you ask them to accept possible randomization in an arm of the study that uses a placebo, thus forgoing a possibly superior medical treatment. If this is the case, you have to minimize the number of people who have to endure these sacrifices.

Definition: population

The group you wish to generalize your research results to.
Usually defined in terms of one or more of the following

- Demography,
- Geography,
- Occupation,
- Time,
- Care requirements,
- Diagnosis,

Example of a population

All infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

Here's an example of a population that hits on most of these elements

"infants" is demography

"Missouri" is geography

"1995" is time

"visits to the Emergency Room" is care requirements.

Definitions: sample

Subset of a population.

- Random sample: every item in the population has the same probability of being in the sample.
- Biased sample: some items in the population have a decreased probability of being in the sample.

A population is usually so large that it is impossible from a logistical or financial perspective to measure everyone in the population. So you take a sample, a subset of the population. Ideally, you have a sample that is representative

Definitions: parameter and statistic

- Parameter: number computed from a population
  - μ, σ, π, ρ
- Statistic: number computed from a sample
  - $\bar{X}, S, p$ or $\hat{p}, r$.

A parameter is a number computed from a population. It is often designated by a Greek letter. The Greek letters mu and sigma represent the population mean and population standard deviation, pi represents a population proportion, and rho represents a population correlation.

A statistic is a number computed from a sample. It is usually represented by a letter in the standard (Roman) alphabet, though sometimes with a symbol above it. The sample mean and sample standard deviation are represented by X-bar and S, the sample proportion by p or p-hat, and the sample correlaton by r.

Putting it all togther

Statistics is the use of one or more **statistics** computed from a **sample** to make an inference about one or more **parameters** from a **population**.

It's a bit simplistic, perhaps, but here's a definition of statistics.

Definition: null hypothesis

- The null hypothesis is, by tradition, a hypothesis that represents
  - no change,
  - no difference, or
  - the status quo
- The null hypothesis is expressed using parameters (Greek)
  - $H_0\ \mu = 100$
  - $H_0\ \pi_1 = \pi_2$
  - $H_0\ \rho = 0$

The null hypothesis is often described as the “negative” hypothesis, but this is judgemental. It represents no change, no difference, or the status quo. It is written in terms of parameters and usually involves an equality.

Definition: alternative hypothesis

- The alternative hypothesis is complementary to the null hypothesis
  - $H_1\ \mu \neq 100$
  - $H_1\ \pi_1 \neq \pi_2$
  - $H_1\ \rho \neq 0$
- Note: no discussion of
  - one-sided hypotheses,
  - equivalence hypotheses,
  - non-inferiority hypotheses

The alternative hypothesis is complementary. I need to mention one-sided hypotheses. These are controversial but I don't wnat to get into all the details about when you use them. There are also equivalence hypotheses and non-inferiority hypotheses.

Definitions: Type I error

- A Type I error is
  - rejecting the null hypothesis when the null hypothesis is true.
  - It is also known as a false positive.
- Example involving drug approval:
  - a Type I error is allowing an ineffective drug onto the market.

One you have the null and alternative hypotheses, you collect statistics to help you accept or reject the null hypothesis. Since a sample is only part of the population, there is sampling error. This means that your statistic is only an estimate of a parameter. If you make a choice to accept or reject the null hypothesis, you could make an error. There are two types of errors, labeled Type I and Type II errors.

A type I error is rejecting the null hypothesis when the null hypothesis is true. Let's consider what this means from the perspective of drug approval.

The FDA requests that a new drug be shown to be superior, either to the current standard of care or sometimes to a placebo. Let's assume that the test is versus a placebo. The null hypothesis says that the new drug is equal to a placebo. The alternative hypothesis is that the new drug is better than placebo.

If you reject the null hypothesis, you are claiming that the data supports the superiority of your new drug over placebo. But in reality, you made a mistake. In the population, the drug is no better than a placebo, but your sample tells you a different story. This is caused by sampling error–errors associated with the sampling process.

If you make this type of error, you are letting an ineffective drug, one that is no better than a placebo, onto the market. Maybe this is good news for the drug company, but from every other perspective, those of the patients, their doctors, and society in general this is a disaster. Patients end up taking a drug that is no better than a sugar pill, so it is a huge waste of money. If there are other drugs out there on the market that are truly effective, this new drug might tempt some patients away from something that does work to something that doesn't work.

Actually, even from the drug company's perspective, a Type I error is bad. Sure it gets you more money, but when the medical community finally figures out that this drug is a bust, it is a publicity black eye for the company.

So think of a Type I error as allowing an ineffective drug onto the market.

Errors associated with the sampling process lead you to a false positive conclusion. This is not good because you are now able to sell a drug for a substantial amount of money, but it is actually no better

Definitions: Type II error

- A Type II error is
  - accepting the null hypothesis when the null hypothesis is false.
  - It is also known as a false negative
- Example involving drug approval:
  - a Type II error is keeping an effective drug off of the market.

This is not the only way you can make a mistake. A Type II error is accepting the null hypothesis when the null hypothesis is false.

Again recall that the null hypothesis is the hypothesis of the status quo, no change, no difference. If you accept the null hypothesis, it is because in your sample, there was little or no difference between the sample mean of the new drug and the sample mean of the placebo. But in the population, there is indeed a difference.

So this is bad news for the drug company. They had a drug that could actually be sold and make money for them. But because they falsely accepted the null hypothesis, they were not able to convince FDA to put the drug on the market. They lose out on a fair amount of profit. It's also bad from everyone else's perspective. Patients lose the opportunity to take a drug that is effective. If there are other drugs on the market that work, then it just means fewer drugs to pick from, less competition, and higher prices. But let's suppose that nothing currently on the market works. People suffer when there is a new drug barred from the market that could help them.

Now balancing the risks between Type I errors and Type II errors is a difficult task. Everyone suffers when an error is made, but the drug companies suffer more from a

Type II error than a Type I error. So they might be tempted to tip the scales in a way that decreases the Type II error rate but at the expense of an increased Type I error rate. Thankfully, FDA controls the drug approval process and assures that both risks are balanced.

Actually, FDA (and the research community in general) used to be too much concerned with Type I errors and not enough with Type II errors. This changed largely because of the AIDS crisis. AIDS (Acquired Immune Deficiency Syndrome) was a disease that, in the 1970s and 80s was rapidly fatal and with no known treatment. Then a class of drugs, anti-retrovirals came along that showed some promise in managing this disease. The burden of AIDS was such that FDA rethought its perspective on how to test new drugs in a disease that had close to a 100% mortality rate and no approved treatment. There were patient advocacy groups like ACT UP that lobbied for these changes. And thanks to their work and the changing perspective of FDA, we now are taking a more balanced approach to drug approval, one that examines the costs to patients and to society in general when an ineffective drug is allowed on the market versus the costs when an effective drug is kept off the market.

Definition: Power function

- Probablity of rejecting the null hypothesis for various values in the alternative hypothesis
- Power is a function of
  - Parameter(s)
  - Alpha
  - Sample size
- Power = 1 - β

Many sample size justifications mention the power function. Power is the probability of rejecting the null hypothesis.

Example of a power function

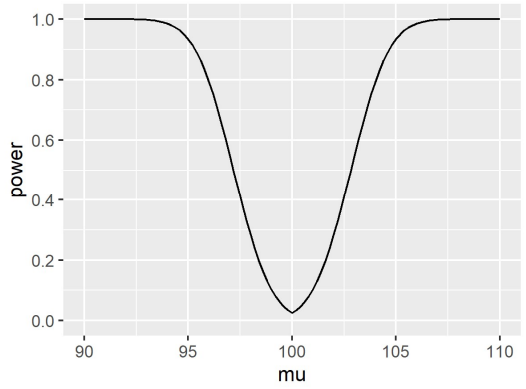

Figure 2. Relationship between power and hypothesized mean

Here's an example of a power function for a one sample t-test of the hypothesis H0 mu=100 for a sample of size 50 and a standard deviation of 10. Notice that the power is lowest when you are close to the hypothesized value of 100. When you move away from 100 in either direction, the power increases to almost 1.0.

Power as a function of sample size

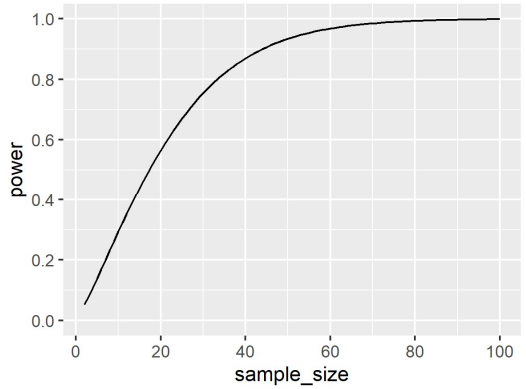

Figure 3. Relationship between power and sample size

This graph shows the power for detecting a 5 unit shift in the mean for a group that has a standard deviation of 10 (an effect size of 0.5). This power is calculated for sample sizes of 2 all the way to 100. Notice a sharp rise in power early, but the graph also shows the diminishing returns. An increase of 10 subjects makes a big difference in power when your sample size is 30, but an increase in sample size of 10 when the sample size is 60 shows much less of an impact of power.

Sample size justification

In a nutshell, choose a sample size large enough

- to insure that the chances of making a Type I or a Type II error are both small. or
- (equivalently) to insure that alpha is small and power is large. or
- to produce a reasonably narrow confidence interval.

There are two basic ways that you justify a sample size. You can choose a sample size large enoug so that the chances of making either type of error (Type I or Type II) is small. Since power is inversely related to beta, the probability of a Type II error, you can say equivalane

Three things that you need to justify your sample size.

- Research hypothesis
- Standard deviation of your outcome measure
- Minimum clinically important difference

When a client asks me for a sample size, I ask for three things. The first is a research hypothesis. The client may not have a hypothesis, so my job is to develop one with them. Of course, not all research has to have a hypotheis and I want to talk a bit more about that case later.

The second thing I ask for is a bit harder, the standard deviation of the outcome measure.

Hardest yet is the minimum clinically important difference.

Finding the standard deviation

- Run a pilot study
- From previous research
  - Maybe already sitting in your draft bibliography
  - Reasonably comaparable
    - Same outcome measure
    - Similar study population

You might consider getting an estimate of the standard deviation from the data you collectin in a pilot study. There are lots of good reasons to run a pilot study anyway, so this should always be a consideration.

You've already done a literature review haven't you? If so, search through the papers in your review that used the same outcome measure that you are proposing in your study and a reasonably similar patient population. This is not always easy, and you will sometimes be forced to use a study where the patients are quite different from your patients. Don't fret too much about this, but make a good faith effort to find the most representative population that you can.

Some clients will raise an objection here and say that their research is unique, so it is impossible to find a comparable paper. It is true that most research is unique (otherwise it wouldn't be research). But what these people are worried about is that their intervention is unique. It's easier to find a match on the other details of the proposed research: the patients being studied and the outcome being measured. These are the more important details. If you find a study where the patient population and the outcome match, the you are probably going to get a good estimate of variation.

Inferring a standard deviation from other measures of variation

- If the paper does not have a standard deviation, look for
  - a standard error,
  - a coefficient of variation,
  - a confidence intervals, or
  - a range ($S \approx RANGE/4$)
- For a binary outcome
  - No standard deviation needed
  - Find the baseline or control proportion instead

If you find a good research paper that is a reasonable match to your proposed research, you might not see a standard deviation in any of the tables. If instead you find a standard error, that's good enough. The standard error is just the standard deviation divided by the square root of the sample size, so it is easy to calculate one from the other.

Likewise, if you know the coefficient of variation (defined as the mean divided by the standard deviation), you can usually figure out what the standard deviation is.

If you can find a confidence interval for the outcome measure, that's a bit more work, but you can usually invert the confidence intervals formula to get the standard deviation.

There is only an approximate relationship between the range and the standard deviation, but if that is the only measure of variation in the paper, the range divided by 4 is usually a pretty good estimate of the standard deviation.

If your outcome is binary (two possible outcomes) then you don't need a standard deviation to calculate power and justify your sample size. Instead, find the proportion

at baseline or in your control group.

Determining the minimum clinically important difference

- Requires clinical judgement
  - Some differences are clinically trivial
  - Large enough to change professional opinion
  - Large enough so that patients notice
- Example: Pain score
  - Measured on visual analog scale (0-100)
  - Patients only feel better with a shift of 15 points

The minimum clinically significant difference is the boundary between a difference so small that no one would adopt the new intervention on the basis of such a meager changer and a difference large enough to make a difference (that is, to convince health care professionals to change their behavior and adopt the new therapy). Or ask yourself what size difference is going to be noticed by the patients.

An example of establishing the minimum clinically important difference is with pain as an outcome measure. In many studies, pain is measured on a visual analog scale. The patient is shown a line of exactly 100 millimeters with one end marked as “no pain” and the other end marked as “worst” or “ubearable” pain. The patient makes a mark anywhere in that line and you measure the distance.

Studies have been done that showed that patients who report a reduction in pain after receiving analgesics move the mark about 15 millimeters to the right.

Cost-benefit tradeoff

- For a binary outcome
  - X = the cost of the treatment
  - kX = benefit of the cure
  - MCD = 1/k
- Examples
  - Aromatherapy
  - Gerson therapy (coffee enemas)

For binary outcomes, the choice is not too difficult in theory. Suppose that an intervention "costs" X dollars in the sense that it produces that much pain, discomfort, and inconvenience, in addition to any direct monetary costs. Suppose the value of a cure is kX where k is a number greater than 1. A number less than 1, of course, means that even if you could cure everyone, the costs outweigh the benefits of the cure.

For k>1, the minimum clinically significant difference in proportions is 1/k. So if the cure is 10 times more valuable than the costs, then you need to show at least a 10% better cure rate (in absolute terms) than no treatment or the current standard of treatment. Otherwise, the cure is worse than the disease.

It helps to visualize this with certain types of alternative medicine. If your treatment is aromatherapy, there is almost no cost involved, so even a very slight probability of improvement might be worth it. But Gerson therapy, which involves, among other things, coffee enemas, is a different story. An enema is reasonably safe, but is not totally risk free. And it involves a substantially greater level of inconvenience than aromatherapy. So you'd only adopt Gerson therapy if it helped a substantial fraction of patients. Exactly how many depends on the dollar value that you place on having

to endure a coffee enema, a task that I will leave for someone else to quantify.

Practical example of sample size justification

- Comparison of two skin barriers for burn patients
  - Pain
  - Healing time
  - Cost

In a study of two different skin barriers for burn patients, we are interested in three outcome measures: pain, healing time, and cost. We will randomly assign half of the patients to one skin barrier and half to the other.

Sample size justification using pain, the formula

- Oucher scale
  - S = 1.5
  - MCD = 1

$$n = (\sigma_1^2 + \sigma_2^2)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 / MCD^2$$

For pediatric patients we usually measure pain with the Oucher, a five point scale that has been validated for children. A review of previous studies using the Oucher have shown that it has a standard deviation of about 1.5 units. We would be interested in seeing how large a sample size is needed to show a change of 1 unit, the smallest individual change attainable on the Oucher. We want to have a power of .80, or equivalently, the probability of a Type II error of .20.

The formula for the sample size in each group is a bit messy, but nothing that you can't handle. Plug in the standard deviation of 1.5 in both spots, the various percentiles for the standard normal distribution, and the minimum clinically important difference. The sample size required to produce 80% power for a two-sided test at an alpha level of 0.05 is 35.3. Round this up to n=36 per group.

Sample size justification using pain, the calculations

$$n = (\sigma_1^2 + \sigma_2^2)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 / MCD^2$$

$$n = (\sigma_1^2 + \sigma_2^2)(z_{0.975} + z_{0.80})^2 / MCD^2$$

$$n = (1.5^2 + 1.5^2)(1.96 + 0.84)^2 / 1^2$$

n=35.3, round up to n=36 per group.

The formula for the sample size in each group is a bit messy, but nothing that you can't handle. Plug in the standard deviation of 1.5 in both spots, the various percentiles for the standard normal distribution, and the minimum clinically important difference. The sample size required to produce 80% power for a two-sided test at an alpha level of 0.05 is 35.3. Round this up to n=36 per group.

Sample size calculation using R, program code

```
> power.t.test(
    delta=1,
    sd=1.5,
    sig.level=0.05,
    power=0.80,
    n=NULL,
    type="two.sample")
```

Here's how you would do this in R. I love R and would be thrilled to give you an introduction to it, but I don't have enough time. Suffice it to say that You won't be quizzed on this later.

For those of you who do know R, this is how it works. Specify four out of the following five parameters:

delta, the minimum clinically important difference

sd, the standard deviation

sig.level or alpha, the probability of a Type I error. This is traditionally set at 0.05, but there may be times that you want it to be 0.10 or 0.01.

power or 1-beta. This is traditionally set at 0.8 or 0.9.

n, the sample size per group.

The parameter you want to compute is left as NULL. So if you want a sample size that

produces 80% power, leave n as NULL. If you want to estimate power when the sample size is 50 per group, set power to NULL.

Sample size calculation using R, output

```
     Two-sample t test power calculation

              n = 36.3058
          delta = 1
             sd = 1.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Here's the output. If you want 80% power, you need a sample size of (rounding up) 37 per group.

Power calculation using PiFace, where to find it

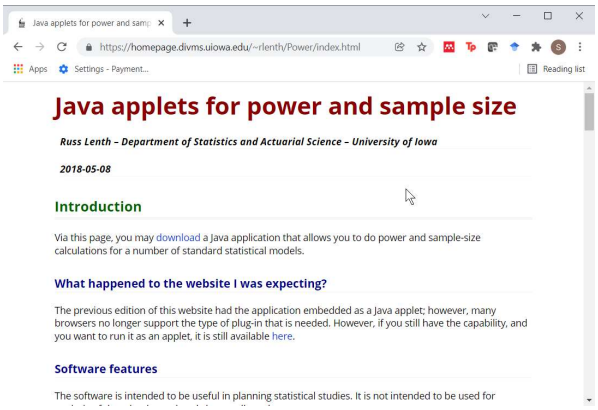

Figure 4. Excerpt from PiFace website

Let me show you a different way to calculate power. It uses a program called PiFace, that was developed by Russ Lenth at the University of Iowa. Russ was my dissertation advisor, and I would have loved to work on something like PiFace, but he developed it several years after I graduated. Anyway, it is a great program and is freely available. Do a web search on “PiFace” and “Power” and this page should pop up. PiFace is written in Java, and you need to download an install the Java Runtime Environment (JRE) prior to using this program. Both PiFace and JRE are free downloads.

Power calculation using PiFace,

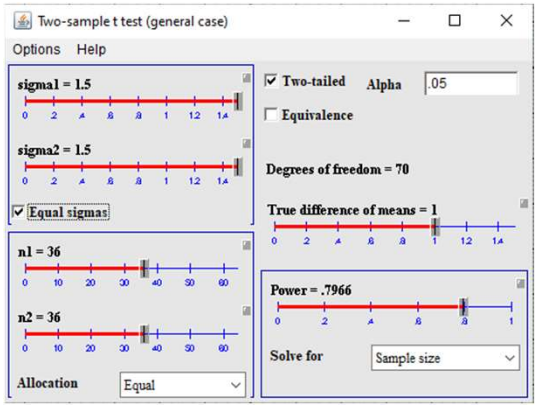

Figure 5. Excerpt from PiFace website

I'm going to try to demonstrate this program live, but if that doesn't work, here is what the results are. A sample size of 36 per group gives you a power of 0.7966.

There are lots of other programs out there that do power and sample size calculations.

Pop quiz, question #1

- A good sample size will produce
  (1) Large values for both alpha and beta.
  (2) A large value for alpha and a small value for beta.
  (3) A small value for alpha and a large value for beta.
  (4) Small values for both alpha and beta.
  (5) I'm awfully glad I'm a Beta, because I don't work so hard.
  (6) I don't know the answer.

Ideally, (4) is the best answer. Larger sample sizes can insure fewer chances for Type I and Type II errors. But if you want to quibble and say that we always hold alpha constant at 0.05 and only beta would decrease, that's a fine answer also.

Pop quiz, question #2

- One of the three things you need to calculate an appropriate sample size is
  (1) A confidence interval for your outcome variable
  (2) A range for your outcome variable
  (3) A standard deviation for your outcome variable.
  (4) A standard error for your outcome variable.
  (5) Any of these is fine.
  (6) I don't know the answer.

Actually any of these is fine. The simplest thing is to have a standard devation, but if you know any of the others, then it is pretty easy to calculate an exact value for the standard deviation except in the case of the range, where you can get an approximate standard deviation.

Pop quiz, question #3

- The minimum clinically important difference is determined by
  (1) Finding a balance between the benefits and the harms of a new drug.
  (2) Finding a balance between the cost of sampling an additional patient and the incremental reduction in uncertainty.
  (3) Finding a balance between Type I and Type II error rates.
  (4) Finding a balance between your work and your family.
  (5) More than one answer above is correct.
  (6) I don't know the answer.

It's tricky to calculate the minimum clinically important difference, but I did show one example where you compare the costs (pain, discompfort, inconvenience) versus the benefits of cure to establish what proportion of cures is needed to outweigh the costs.