8. Detectable Difference, Power, and Sample Size

Readings: Rosner-7.5-7.7
R power and sample size functions (Appendix)

Homework: Homework 3 due by noon on September 24
Homework 4 due by noon on October 1

Overview

A) Important probabilities and definitions
B) Evaluating the performance of a statistical test
C) Derivation of power for two-sided, one-sample Z-test (variance known)
D) How power, detectable difference, and sample size relate to each other
E) Using R for power and sample size (with one-sample Z-test (known variance) and one-sample t-test (unknown variance))

Often in planning studies we are asked “How many subjects do I need for the study?” Or, “If I have 30 subjects, do I have enough power to see a significant result?” These questions are not as readily answered as they might seem. The answers depend on several factors, which we will now examine. Finding rejection regions and p-values requires calculating probabilities assuming H_0: $\mu = \mu_0$ is true. To find power we need to consider the flip side: H_0 is not true, and some specific alternative hypothesis H_1 is true.

e.g. Generic example: A collaborating investigator says, “I want to do a study where I compare my new method to the standard method. How many subjects do I need? I think I can get 50, is that enough?” What questions would you want to ask your colleague, and how would you answer?

Recall: A hypothesis is a claim or statement about a population parameter or parameters, and a hypothesis test is a statistical method of quantifying evidence (using sample information) to reach a decision about a hypothesis.

First question to ask: “What is the (null) hypothesis?”

A) Important probabilities and definitions

Recall: Based on the data, we make a decision to reject H_0 or to not reject H_0, and we quantify the evidence against the H_0 in the form of a p-value. (Remember, we do not "accept H_0" since all we can do is say if we have enough evidence to reject H_0 or not!)

Reality ⇒ What we decide ⇓	**H_0 True**	**H_0 False/H_1 True**
Fail to reject H_0	*Correct* Probability of correct decision = $1-\alpha$ = Level of confidence	*Type II Error* P(Type II Error) = β
Reject H_0	*Type I error* P(Type I error) = α Level of significance	*Correct* Probability of correct decision = $1-\beta$ = Power

	H_0 True	**H_0 False**
Fail to reject H_0	$1-\alpha$ Level of confidence	β P(Type II Error)
Reject H_0	α Level of significance P(Type I Error)	$1-\beta$ Power: Probability of finding difference if it exists

P(Type I Error) = α = Probability of rejecting the null hypothesis when it is true;
P(Reject $H_0 \mid H_0$ is true); Level of significance

P(Type II Error) = β = Probability of failing to reject the null hypothesis when it is false;
P(Fail to Reject $H_0 \mid H_0$ is false)

Approach: We want both α and β to be small.
Since β increases as α decreases, this is not a well-defined problem.
We know how to fix α (usually at 0.05). How do we make sure β is not too big?

Power:
P(Reject $H_0 \mid H_0$ is false) = 1 – P(we do not reject $H_0 \mid H_0$ is false) = P(Reject $H_0 \mid H_1$ true) = $1 - \beta$

B) Evaluating the performance of a hypothesis test

There are 4 important quantities that we must consider:

1) Level of significance of a test = α
2) Power of a test = $1 - \beta$
3) Sample size = n
4) Detectable difference in means: $|\mu_0 - \mu_1|$

We usually set α (typically at 0.05). Then, by fixing two of the other quantities, we can solve for the last remaining one:

A) Find the *power* of a test with *n* subjects to detect a *difference in means* of $|\mu_0 - \mu_1|$
For a two-sided test with significance level α where H_0: $\mu = \mu_0$ vs. H_1: $\mu = \mu_1$, using a one sample Z-test (i.e. assuming σ is known), our power can be calculated as:

$$Z_{1-\beta} = \frac{|\mu_0 - \mu_1|}{se(\bar{X})} - Z_{1-\alpha/2} = \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} - Z_{1-\alpha/2}$$

$$1 - \beta = \Phi\left[\frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} - Z_{1-\alpha/2}\right]$$

B) Find the *difference in means* that can be detected with *n* subjects and *power = 1-β*

Detectable difference: $|\mu_0 - \mu_1| = (Z_{1-\beta} + Z_{1-\alpha/2})\, \sigma/\sqrt{n}$

C) Find the *number of subjects* needed to detect a *difference* of $|\mu_0 - \mu_1|$ with *power = 1-β*

Required sample size: $n = \frac{\sigma^2 \left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2}{(\mu_0 - \mu_1)^2}$

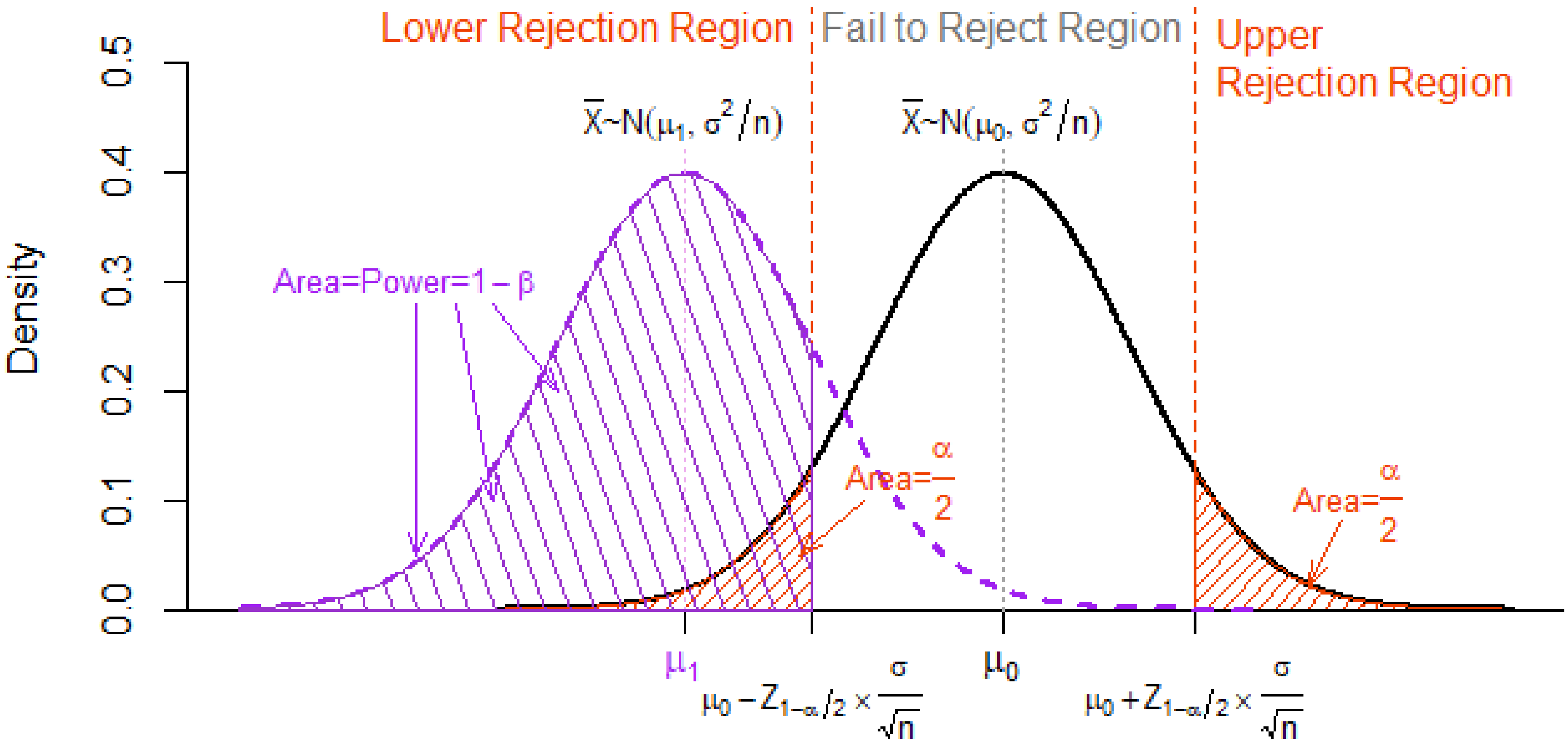
Note: If we change α, the values of all of the above would also change.

Note: Lots of versions of these equations exist and can be confusing, potentially with different assumptions.

Note: For all these scenarios, we also need one additional piece of information: σ, the population standard deviation. In practice, its true value is hard to come by. Previous work reported in the literature is a good source, but sometimes a pilot study needs to be done to get a reasonable (in the ballpark) estimate. If the estimate is a good one, assuming the Z-distribution is fine. If we want to reflect the uncertainty in the value of σ then we would need to use the t-distribution. Since this depends on df=n-1, sample size estimation would be an *iterative* process in this case.

C) Derivation of power for a two-sided, one-sample Z-test

Assume we have normally distributed data, and we want to test H_0: $\mu = \mu_0$ vs. H_1: $\mu = \mu_1$. Consider the distribution of $\overline{\boldsymbol{X}}$ if $\mu = \mu_0$ and if $\mu = \mu_1$:



The power to detect the difference $|\boldsymbol{\mu_0} - \boldsymbol{\mu_1}|$ in our figure: $P(\text{Reject H}_0|\text{H}_0 \text{ false}) = 1 - \beta$

$$
\begin{aligned}
P(\text{Reject H}_0|\text{H}_0 \text{ false}) &= P(\text{Reject H}_0|\mu = \mu_1) \\
&= P(\bar{X} < \mu_0 - Z_{1-\frac{\alpha}{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) \text{ OR } \bar{X} > \mu_0 + Z_{1-\frac{\alpha}{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) |\mu = \mu_1) \\
&= P\left(\bar{X} < \mu_0 - Z_{1-\frac{\alpha}{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) \middle| \mu = \mu_1\right) + P\left(\bar{X} > \mu_0 + Z_{1-\frac{\alpha}{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) |\mu = \mu_1\right) \\
&= P\left(\bar{X} < \mu_0 - Z_{1-\frac{\alpha}{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) \middle| \mu = \mu_1\right) + \approx 0 \\
&= P\left(\frac{\bar{X} - \mu_1}{{}^{\sigma}/_{\sqrt{n}}} < \frac{\mu_0 - Z_{1-{}^{\alpha}/_{2}}\left({}^{\sigma}/_{\sqrt{n}}\right) - \mu_1}{{}^{\sigma}/_{\sqrt{n}}} |\mu = \mu_1\right) \\
&= P\left(Z < \frac{\mu_0 - \mu_1}{{}^{\sigma}/_{\sqrt{n}}} - Z_{1-{}^{\alpha}/_{2}}\right) \\
&= \Phi(Z < Z_{1-\beta}) \\
&= 1 - \beta
\end{aligned}
$$

The formulas for sample size and detectable difference derive from this formula.

D) How power, detectable difference, and sample size relate to each other

$$|\mu_0 - \mu_1| = \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)\frac{\sigma}{\sqrt{n}};\ 1-\beta = \Phi\left[\frac{|\mu_0-\mu_1|}{\sigma/\sqrt{n}} - Z_{1-\frac{\alpha}{2}}\right];\ n = \frac{\sigma^2\left(Z_{1-\beta}+Z_{1-\alpha/2}\right)^2}{(\mu_0-\mu_1)^2}$$

As sample size n increases:

Power:

Detectable Difference $|\mu_0 - \mu_1|$:

As the difference to be detected, $|\boldsymbol{\mu_0} - \boldsymbol{\mu_1}|$, increases:

Power:

Required Sample Size:

As desired power increases:

Required Sample Size:

Detectable Difference $|\mu_0 - \mu_1|$:

$$|\mu_0 - \mu_1| = \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)\frac{\sigma}{\sqrt{n}};\ 1-\beta = \Phi\left[\frac{|\mu_0-\mu_1|}{\sigma/\sqrt{n}} - Z_{1-\frac{\alpha}{2}}\right];\ n = \frac{\sigma^2\left(Z_{1-\beta}+Z_{1-\alpha/2}\right)^2}{(\mu_0-\mu_1)^2}$$

As σ, the population s.d., increases:

Power:

Detectable Difference $|\mu_0 - \mu_1|$:

Required Sample Size:

As α, the significance level of the test, increases:

Power:

Detectable Difference $|\mu_0 - \mu_1|$:

Required Sample Size:

e.g. A new calcium channel blocker is to be tested for treatment of unstable angina. We measure the change in heart rate after 48 hours.
"How many subjects do I need?" "If I use 20 subjects, what will I be able to see?"

What's the hypothesis?

To carry out the test, we need to specify the α-level, e.g. 0.05.

To estimate the required sample size, we will still need:

1.

2.

3.

Say we are told σ = 10 beats/min for changes using another drug and we want to detect a difference of 5 beats/min.

To test the drug for decreasing heart rate (HR), each subject is tested in two conditions:
1. No drug and 2. With the drug

We compute $D = HR_{\text{no drug}} - HR_{\text{drug}}$ = data

Null: H_0: $\mu = \mu_0 = 0$ (mean difference = 0, drug has no effect)
Alternative: H_1: $\mu = \mu_1 \neq 0$ (expect $\mu_1 > 0$; but will do a two-sided test)

Situation 1: Suppose drug truly decreases HR by 2 bpm. (μ_1 = 2 bpm)

Situation 2: Suppose drug truly decreases HR by 10 bpm. (μ_1 =10 bpm)

Q. In repeated experiments, in which situation will we be more likely to detect the effect of the drug, i.e. to reject H_0, if all other parameters are held constant between studies? What if we only know μ_1?

A. The proportion of experiments detecting the difference is approximately 1- β (i.e., power). With all parameters are constant Situation 2 will have higher power. If we only know μ_1, then either Situation could have more power depending on the other parameters. Recall, power depends on:

1. n = sample size
2. size of detectable difference $|\mu_0 - \mu_1|$
3. σ for the difference or change in HR
4. α-level

Different approaches to power analysis depending on investigator's question:

<u>"How many subjects are needed?"</u> : Investigator might want to specify the power to detect a specific change, e.g. a change of 5 bpm vs. 0 bpm. Using α = 0.05, power = 0.80, and s.d. for change of 10 bpm:

$$n = \frac{\sigma^2 \left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2}{(\mu_0 - \mu_1)^2} = \frac{10^2 (Z_{0.8} + Z_{0.975})^2}{5^2} = \frac{10^2 (0.84 + 1.96)^2}{25} = 31.36$$

So the investigator will need about 32 subjects. (Note, in order to achieve *at least* the desired power we must round up, even if the estimate was n=31.01 subjects needed. Otherwise we would be slightly *underpowered*.)

"If 20 subjects are used, what power is there to detect a difference?": Suppose the investigator estimates they can only get a certain number of subjects in the time period of the study and wants to know how much power that will give to detect a specified difference. Using α = 0.05, n=20, and s.d. for change of 10 bpm:

$$Z_{1-\beta} = \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} - Z_{1-\frac{\alpha}{2}} = \frac{5}{10/\sqrt{20}} - 1.96 = 2.236 - 1.96 = 0.276$$

$$\Phi(0.276) = 0.608726 \approx 0.61 = 1 - \beta = \text{Power}$$

A sample size of 20 provides a power of 61% to detect a difference of at least 5 bpm, given that it exists.

"If 20 subjects are used, what is the detectable difference?": Another interpretation of this question could be: how large a difference is detectable with 20 subjects and a power of 80%?

$$|\mu_0 - \mu_1| = \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)\frac{\sigma}{\sqrt{n}} = (Z_{0.8} + Z_{0.975})\frac{10}{\sqrt{20}} = (0.84 + 1.96)(2.236) = 6.26\text{ bpm}$$

With a sample size of 20, there is 80% power to detect a difference of at least 6.26 bpm, given that it exists.

One-Sided Power Analysis:

For $\mu_1 > \mu_0$: $Z_{1-\beta} = \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}} + Z_\alpha = \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}} - Z_{1-\alpha}$; $1-\beta = \Phi\left[\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}} - Z_{1-\alpha}\right]$

For $\mu_1 < \mu_0$: $Z_{1-\beta} = \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + Z_\alpha = \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - Z_{1-\alpha}$; $1-\beta = \Phi\left[\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - Z_{1-\alpha}\right]$

Example: What is the power?

H_0: μ = 90 mmHg DBP vs. H_1: μ > 90 mmHg DBP

Let's say that specific alternative of interest is H_1: μ_1 = 96 mmHg. What is the power if σ^2 = 64 $(\text{mmHg})^2$, n = 16, and α = 0.05?

We reject H_0: μ = 90 mmHg DBP if: $\bar{X} > \mu_0 + Z_{1-\alpha}\left(\sigma/\sqrt{n}\right) = 90 + 1.645\left(\frac{8}{4}\right) = 93.3\text{ mmHg}$

Power : $Z_{1-\beta} = \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}} - Z_{1-\alpha} = \frac{96-90}{8/\sqrt{16}} - 1.645 = 1.355 \rightarrow 1-\beta = \Phi[1.355] = 0.9115$

We have over 90% power to detect a difference of at least 6 mmHg greater than 90 mmHg.

What happens to power if μ_1 = 92 mmHg?

Power : $Z_{1-\beta} = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} - Z_{1-\alpha} = \frac{92-90}{8/\sqrt{16}} - 1.645 = -0.645 \rightarrow 1 - \beta = \Phi[-0.645] = 0.2595$

We have less than 26% power to detect a difference of at least 2 mmHg greater than 90 mmHg.

Failure to find a difference:

- "Publish or perish" pressure in academia makes it highly desirable to find a difference between groups, treatments, etc. because it's "more interesting, worthy of journal space, etc." This creates an incomplete distribution of true results since only the significant results make it to press. Consequence is called **publication bias**, or the "file drawer problem", in the language of meta-analysis.

- Regression to the Mean is one of several phenomena that lead to the observed lack of reproducibility of scientific results. For a thorough and engrossing essay on the topic of reproducibility, be sure to read the *New Yorker* article: http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer.

- *Post-hoc* power analyses can be done but they are not too informative. Power should be based on a pre-specified detectable difference, not the one seen in the study itself. See the paper mentioned on the next slide by Hoenig and Heisey (in the Canvas Paper Repository).

- Power should be set very high for studies aiming at a definitive result. This is done to avoid the "slippery slope of power", i.e. where power is sensitive to assumptions about σ and detectable difference $|\mu_0 - \mu_1|$. This is illustrated with regard to detectable difference in scenario (b) below.

Useful references on Power and Sample Size Calculations:

Lachin, JM. (1981). Introduction to Sample Size Determination and Power Analysis of Clinical Trials. *Controlled Clinical Trials*, 2: 93-113. (On the Canvas course site in the Paper Repository).

Hoenig, JM and Heisey, DM. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, 55:19-24. (On the Canvas course site in the Paper Repository).

E) Using R for power and sample size

Let’s consider some scenarios for testing a single mean (“one-sample Z-test” or “one-sample t-test”) based on a single sample and see how we can approach them using R.

a) σ = 10 (known); N=15, 20, 25; Detectable Difference between null and alternative means = -15 to 15; α = 0.01, 0.05, 0.10 (two-sided). Find Power.

b) s = 10 (σ unknown); N=15, 20, 25; Detectable Difference between null and alternative means = -15 to 15; α = 0.01, 0.05, 0.10 (two-sided). Find Power.

c) s = 10 (σ unknown); Detectable Difference between null and alternative means = 5 to 10; α = 0.01, 0.05, 0.10 (two-sided); Power = 0.80, 0.90, 0.95. Find N.

d) s = 10 (σ unknown); N = 15, 20, 25; α = 0.01, 0.05, 0.10 (two-sided); Power = 0.80, 0.9, 0.95. Find detectable difference between null and alternative means.

e) Confidence interval s = 10 (σ unknown); α = 0.01, 0.05, 0.10 (two-sided); halfwidth = 3 6 9 12; prob (halfwidth) = 0.80, 0.9, 0.95. Find N.

a) σ = 10 (known); N=15, 20, 25; Detectable Difference between null and alternative means = -15 to 15; α = 0.01, 0.05, 0.10 (two-sided). Find Power.

In R it is easiest to program our own function from the equations provided in this lecture to calculate the power for our two-sided Z:

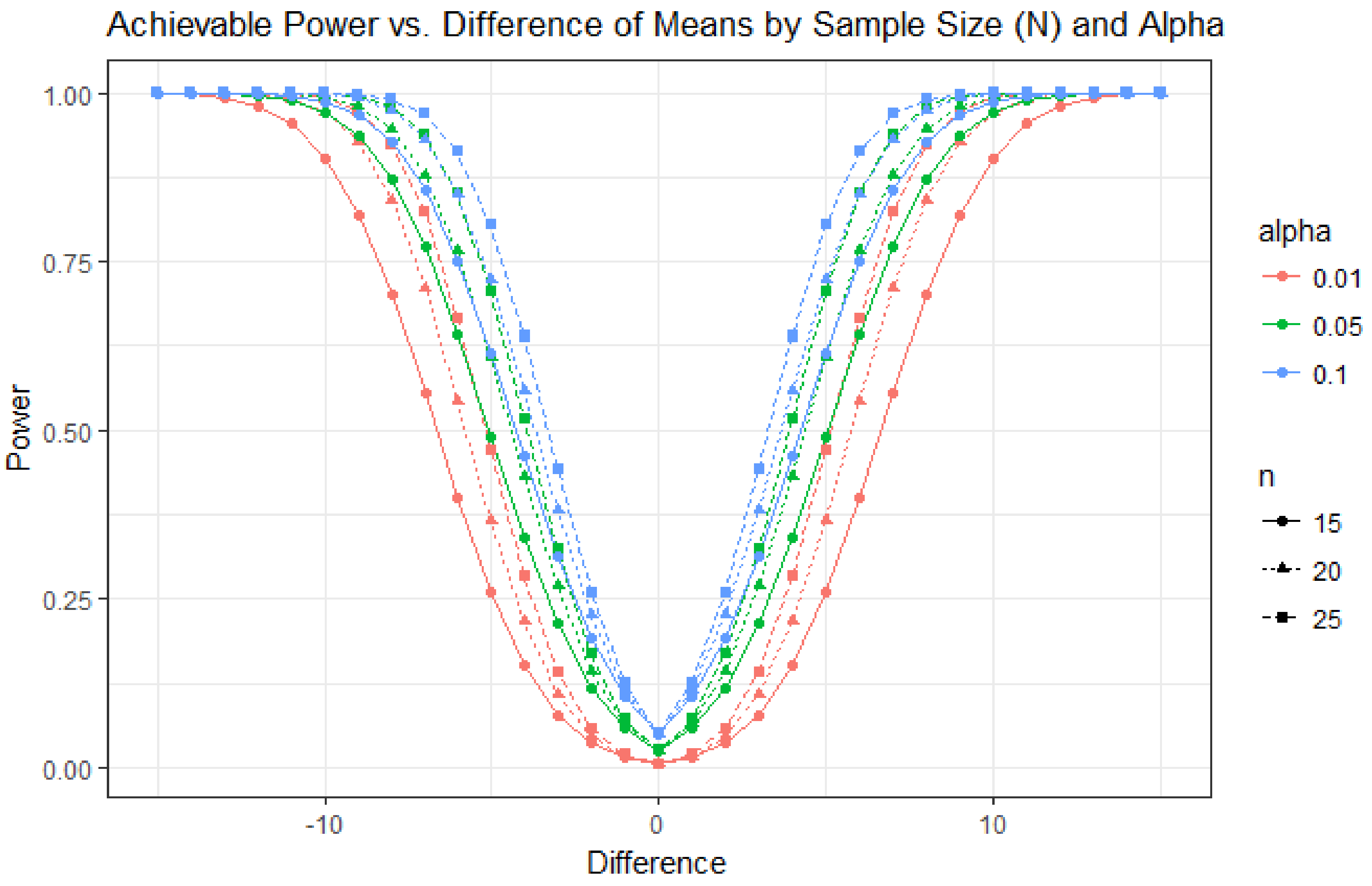
```
findPowerZ <- function(diff = 5, sd = 1, n = 10, alpha = 0.05){
    z.alpha <- qnorm(1 - (alpha/2))
    power <- pnorm(diff/(sd/sqrt(n)) - z.alpha)
    return(power)
}
```

Note: We didn't put absolute values around our diff function, so we will need to account for that for our two-sided test:

```
# R code to obtain power for two-sided Z-test
findPowerZ(diff=abs(5), sd=10, n=20, alpha=0.05)
[1] 0.6087659

findPowerZ(diff=abs(-5), sd=10, n=20, alpha=0.05)
[1] 0.6087659
```

a) continued:


Achievable Power vs. Difference of Means by Sample Size (N) and Alpha
Power
1.00
0.75
0.50
0.25
0.00
-10
0
10
Difference
alpha
0.01
0.05
0.1
n
15
20
25

b) s = 10 (σ unknown); N=15, 20, 25; Detectable Difference between null and alternative means = -15 to 15; α = 0.01, 0.05, 0.10 (two-sided). Find Power.

R provides a default function we can use to calculate power, sample size, etc. when the variance is unknown: power.t.test

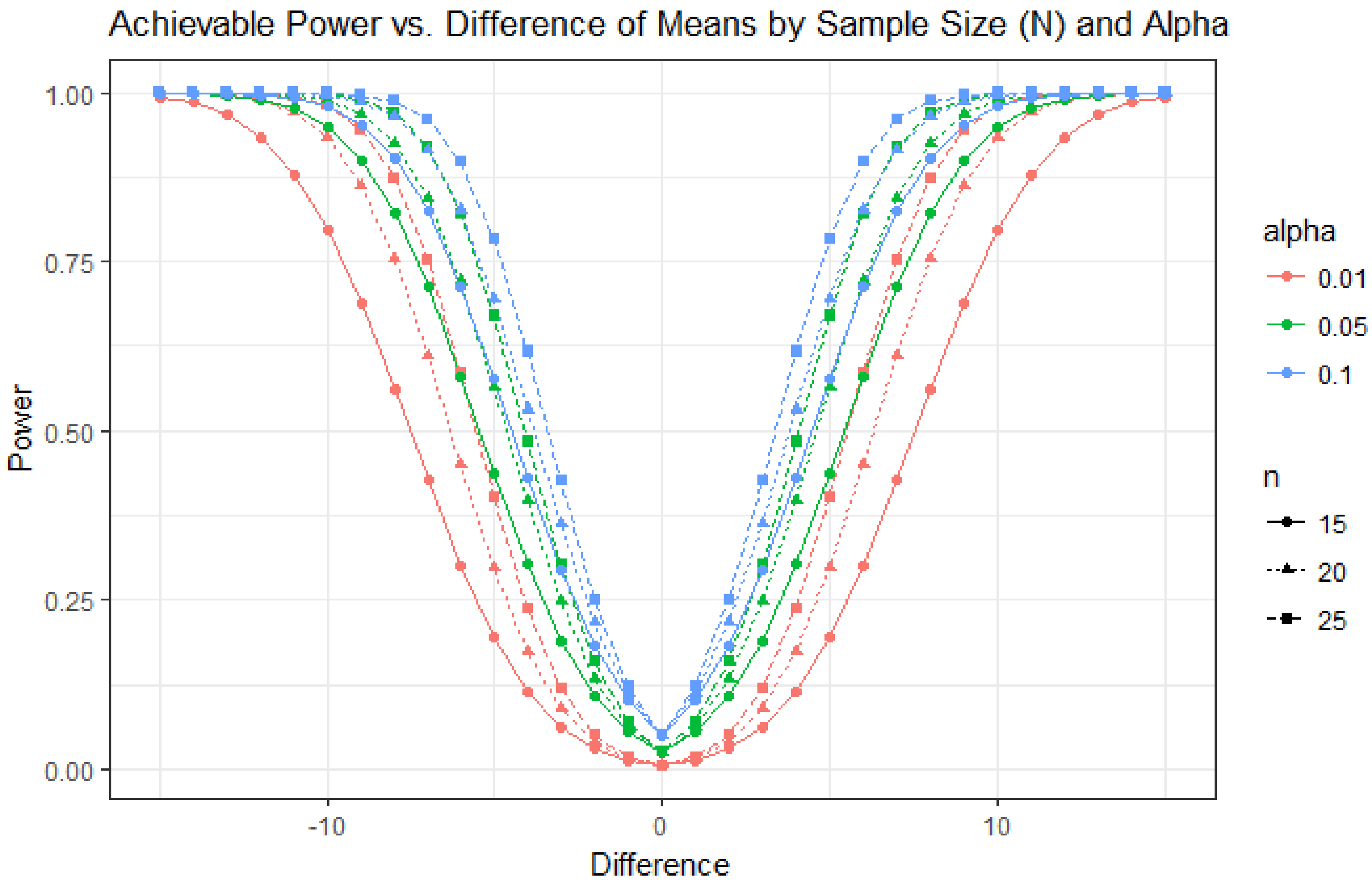
```
# R code to obtain power for t-test
pwrt_b <- power.t.test(n = 15, sd = 10, sig.level = 0.01,
delta=-15, type = "one.sample", alternative = "two.sided")

pwrt_b
     One-sample t test power calculation

              n = 15
          delta = 15
             sd = 10
      sig.level = 0.01
          power = 0.9937996
    alternative = two.sided

pwrt_b$power
[1] 0.9937996
```

b) continued:


Achievable Power vs. Difference of Means by Sample Size (N) and Alpha
Power
1.00
0.75
0.50
0.25
0.00
Difference
-10
0
10
alpha
0.01
0.05
0.1
n
15
20
25

c) $s = 10$ (σ unknown); Detectable Difference between null and alternative means = 5 to 10; $\alpha = 0.01, 0.05, 0.10$ (two-sided); Power = 0.80, 0.90, 0.95. Find N.

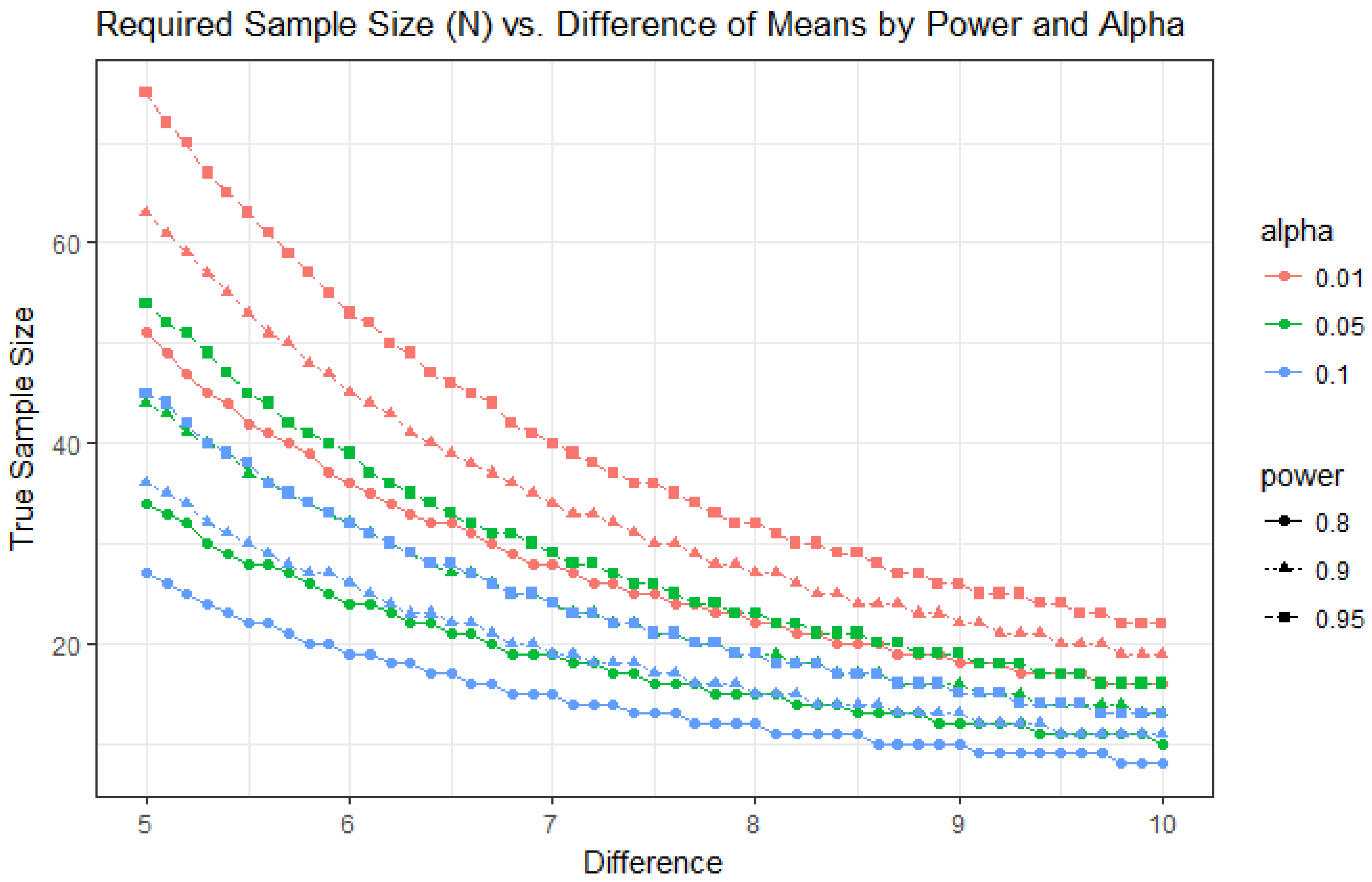
```
# R code to obtain N for t-test
pwrt_c <- power.t.test(power=0.8, sd = 10, sig.level = 0.01, 
delta=5, type = "one.sample", alternative = "two.sided")

pwrt_c
     One-sample t test power calculation 

              n = 50.0647
          delta = 5
             sd = 10
      sig.level = 0.01
          power = 0.8
    alternative = two.sided

ceiling(pwrt_c$n) #use ceiling to round up
[1] 51
```

c) continued:


Required Sample Size (N) vs. Difference of Means by Power and Alpha
True Sample Size
60
40
20
Difference
5
6
7
8
9
10
alpha
0.01
0.05
0.1
power
0.8
0.9
0.95

d) s = 10 (σ unknown); N = 15, 20, 25; α = 0.01, 0.05, 0.10 (two-sided); Power = 0.80, 0.9, 0.95. Find detectable difference between null and alternative means.

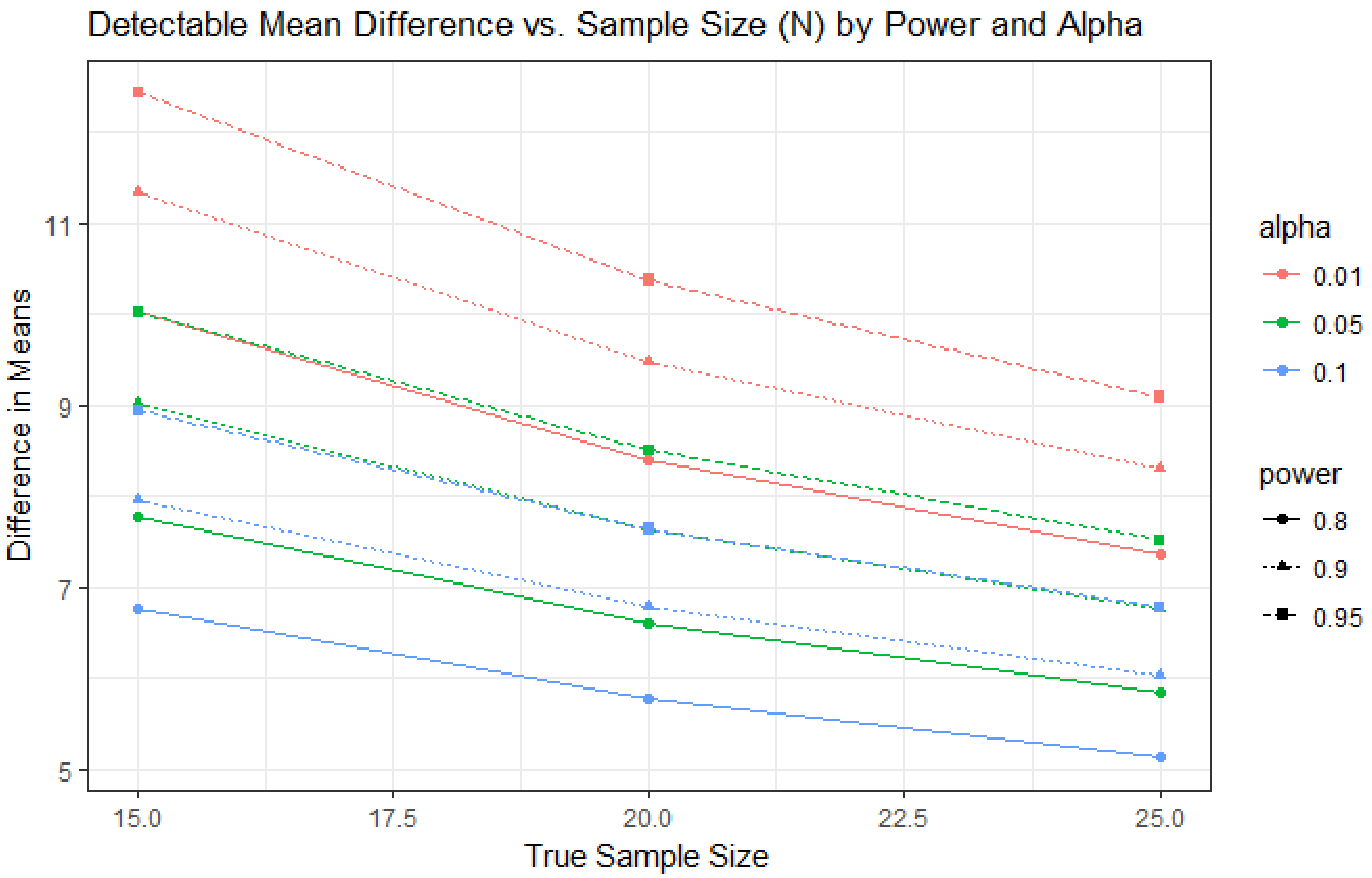
```
# R code to obtain difference for t-test
pwrt_d <- power.t.test(power=0.8, sd = 10, sig.level = 0.01,
n = 15, type = "one.sample", alternative = "two.sided")

pwrt_d
     One-sample t test power calculation

              n = 15
          delta = 10.03483
             sd = 10
      sig.level = 0.01
          power = 0.8
    alternative = two.sided

pwrt_d$delta
[1] 10.03483
```

d) continued:


Detectable Mean Difference vs. Sample Size (N) by Power and Alpha
Difference in Means
11
9
7
5
15.0
17.5
20.0
22.5
25.0
True Sample Size
alpha
0.01
0.05
0.1
power
0.8
0.9
0.95

e) Confidence interval s = 10 (σ unknown); α = 0.01, 0.05, 0.10 (two-sided); halfwidth = 3 6 9 12; prob (halfwidth) = 0.80, 0.9, 0.95. Find N.

Here we need to code up our own custom function to calculate the sample size needed for a given halfwidth of the confidence interval. Recall, the formula for a confidence interval where we don't know σ can be represent by

$$\bar{X} \pm Z_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

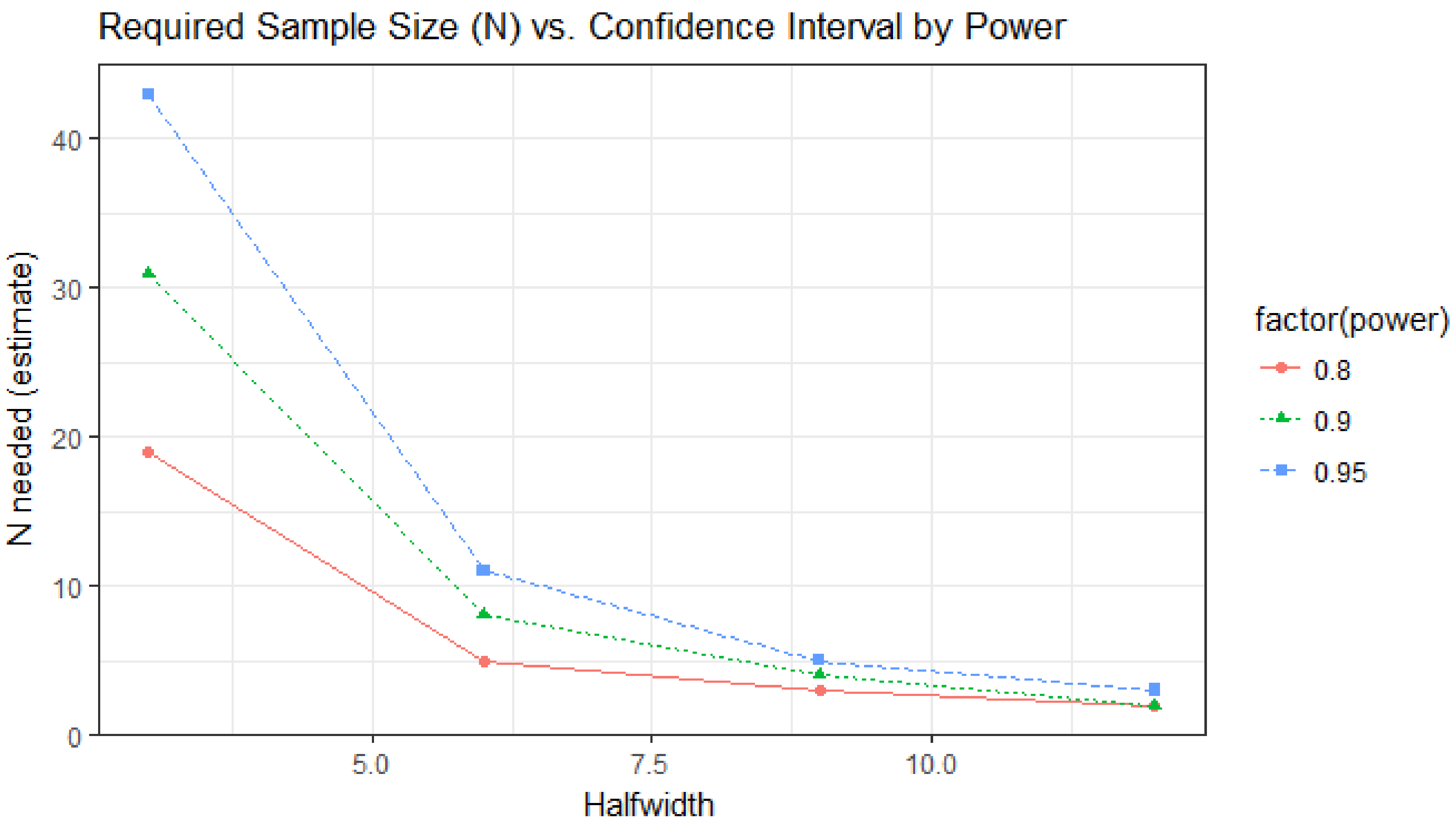
We can take the green second-half of the equation for our halfwidth value and solve for n:

$$w = \frac{Z_{1-\alpha/2} \times s}{\sqrt{n}} \rightarrow n = \left(\frac{Z_{1-\alpha/2} \times s}{w}\right)^2$$

```
findNfromCI <- function(h = 3, pw = 0.95, s = 10) {
     alpha <- 1 - pw
     z <- qnorm(1 - (alpha/2))
     n <- (z * s/h)^2
     out <- ceiling(n)
     return(out)
}

findNfromCI(3)
[1] 43
```

e) continued:


Required Sample Size (N) vs. Confidence Interval by Power
N needed (estimate)
0
10
20
30
40
5.0
7.5
10.0
Halfwidth
factor(power)
0.8
0.9
0.95