STA305/1004 - Class 6

September 24, 2019

Today's Class

- Introduction to Phase III Clinical Trials
- Introduction to power
- Power of the one-sample z-test
- Power of the one-sample t-test

What are clinical trials?

Clinical trials are prospective intervention studies with human subjects to investigate experimental drugs, new treatments, medical devices, or clinical procedures (Yin, 2012).

Phases of clinical trials

Developing a new drug for cancer.

- **Preclinical studies:** In vitro (e.g. slides, test tubes) and in vivo (living organism such as rodents) studies on wide range of doses of experimental agents. This stage of study provides preliminary toxicity and efficacy data including pharmacokinetics (PK) and pharmacodynamics (PD) information.
- **Phase I:** Usually first study in humans to investigate the toxicity and side effects of the new agent. Identify MTD.
- **Phase II:** Assess if drug has sufficient efficacy. The drug is usually administered around the MTD. If drug does not show efficacy or is too toxic then further testing is discontinued.

Phases of clinical trials

- **Phase III:** If drug passes phase II testing then it is compared to the current standard of care or placebo. These are long-term, large scale randomized studies that may involve hundreds or thousands of patients.
- If the drug is proven to be effective (e.g. two positive phase III trials required for FDA approval) the company will file an application with regulatory agencies to sell the drug. If approved then the drug will be available to the general population in the country where it was approved.
- **Phase IV:** After approval a study might follow a large number of patients over a longer period of time to monitor side effects and drug interactions. For example, findings from these studies might add a warning label to the drug.

Phases of clinical trials

- The four phases are usually conducted sequentially and separately.
- Each trial requires an independent study design and a study protocol.
- Every aspect of trial design, monitoring, and data analysis call upon statistical methods.
- In randomized clinical trials a treatment group is often referred to as an **arm**.

Phases of clinical trials

- Experimental design plays a very important role in the design of clinical trials.
- Two arm clinical trials use all of theory of randomization that we learned about last week. Randomization is used to design phase III clinical trials since causation can usually be assessed using a randomized design.

How can causation be assessed using a randomized design?

- Suppose that patients are randomized in a two arm clinical trial where one of the arms is the standard treatment and the other arm is an experimental treatment
- A statistically significant difference in the outcome between the two arms is observed showing the experimental treatment is more efficacious.
- The interpretation is that the experimental treatment *caused* patients to have a better outcome since the only difference between the two arms is the treatment. Randomization is supposed to ensure that the groups will be similar with respect to all the factors measured in the study and all the factors that are not measured.

How many patients should be enrolled in a Phase III clinical trial?

- In a phase III trial sample size is the most critical component of the study design. The sample size has implications for how many subjects will be exposed to a drug that has no proven efficacy.
- The investigator needs to specify type I, II error rates, and the effect sizes.
- Standard practice is to compute the smallest sample size required to detect a clinically important/significant treatment difference with sufficient.

How many patients should be enrolled in a Phase III clinical trial?

- If the sample size is too small then the trial might fail to discover a truly effective drug because the statistical test cannot reach the significance level (5%) due to a lack of power.
- If the sample size is overestimated then resources wasted and drug development delayed since patient enrollment is often the main factor in time to complete a trial.

Statistical hypotheses

Suppose that subjects are randomized to treatments A or B with equal probability. Let μ_A be the mean response in the group receiving drug A and μ_B be the mean response in the group receiving drug B. The null hypothesis is that there is no difference between A and B, the alternative claims there is a clinically meaningful difference between them.

$$H_0 : \mu_A = \mu_B \text{ versus } H_0 : \mu_A \neq \mu_B$$

Statistical hypotheses

The type I error rate is defined as:

$$\begin{aligned}\alpha &= P\left(\text{type I error}\right) \\ &= P\left(\text{Reject } H_0 | H_0 \text{ is true}\right).\end{aligned}$$

Statistical hypotheses

The type II error rate is defined as:

$$\begin{aligned}\beta &= P\left(\text{type II error}\right) \\ &= P\left(\text{Accept } H_0 | H_1 \text{ is true}\right).\end{aligned}$$

Statistical hypotheses

Power is define as:

$$\begin{aligned}\text{power} &= 1 - \beta \\ &= 1 - P\left(\text{Accept } H_0 | H_1 \text{ is true}\right) \\ &= P\left(\text{Reject } H_0 | H_1 \text{ is true}\right).\end{aligned}$$

Power

The probability that a fixed level α test will reject H_0 when a particular alternative value of the parameter is true is called power of the test to detect that alternative.

Power

Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? Based on results of a previous study $\sigma = 2$ for the percent change in TBBMC over the 6-month period. A change in TBBMC of 1% would be considered important. Is 25 subjects a large enough sample size for this project?

Power of the one sample z-test

Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. A test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_0 : \mu \neq \mu_0$$

will reject at level α if and only if

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2},$$

or

$$\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)^{th}$ percentile of the $N(0, 1)$.

Power of the one sample z-test

The power of the test at $\mu = \mu_1$ is

$$\begin{aligned} 1 - \beta &= 1 - P(\text{type II error}) \\ &= P(\text{Reject } H_0 | H_1 \text{ is true}) \\ &= P(\text{Reject } H_0 | \mu = \mu_1) \\ &= P\left(\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2} | \mu = \mu_1 \right) \end{aligned}$$

Subtract the mean μ_1 and divide by $\sigma/\sqrt{n}$ to obtain (why?):

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right) \right) + \Phi\left(-z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right) \right),$$

where $\Phi(\cdot)$ is the $N(0,1)$ CDF.

Power of the one sample z-test

The power function of the one-sample z-test is:

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right) + \Phi\left(-z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right).$$

What is the limit of the power function as:

- $n \to \infty$
- $\mu_1 \to \mu_0$

Why is Power and Sample Size Important in Phase III Clinical Trials?

- If a new treatment is to be used in patients then it should be compared to the standard treatment.
- Evidence is required that the new treatment is effective and safe.
- The form of the evidence is a hypothesis test.
- Will the hypothesis test reject if a difference between the treatments really exists?
- High power will ensure that if a difference exists then the hypothesis test will have a high probability of rejecting.
- The most practical way to ensure the test is powerful is to enrol enough patients in each arm of the trial.

Sample Size and Power in Phase III Clinical Trials

- The sample size is calculated under the alternative hypothesis based on the type I error rate α and power $1 - \beta$.
- Specify a clinically meaningful difference that is to be detected at the conclusion of the trial.
- Intuitively, if a small difference (effect size) is expected between the two treatments in comparison, a large sample size would be required, and vice versa. Why?
- Sample size also depends on the variance.
- The larger the variance, the harder it is to detect the difference and thus a larger sample size is needed.

Power of the one sample z-test

Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. A test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_0 : \mu \neq \mu_0$$

will reject at level α if and only if

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2},$$

or

$$\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)^{th}$ percentile of the $N(0, 1)$.

Power of the one sample z-test

The power of the test at $\mu = \mu_1$ is

$$\begin{aligned} 1 - \beta &= 1 - P(\text{type II error}) \\ &= P(\text{Reject } H_0 | H_1 \text{ is true}) \\ &= P(\text{Reject } H_0 | \mu = \mu_1) \\ &= P\left(\left| \bar{X} - \mu_0 \right| \geq \frac{\sigma}{\sqrt{n}} z_{\alpha/2} | \mu = \mu_1 \right) \end{aligned}$$

Subtract the mean μ_1 and divide by $\sigma/\sqrt{n}$ to obtain (why?):

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right) \right) + \Phi\left(-z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right) \right),$$

where $\Phi(\cdot)$ is the $N(0,1)$ CDF.

Power of the one sample z-test

The power function of the one-sample z-test is:

$$1 - \beta = 1 - \Phi\left(z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right) + \Phi\left(-z_{\alpha/2} - \left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right).$$

What is the limit of the power function as:

- $n \to \infty$
- $\mu_1 \to \mu_0$
- $\sigma \to 0$

Power of the one-sample z-test

The power function for a one-sample z-test can be calculated using R.

$$1-\beta = 1-\Phi\left(z_{\alpha/2}-\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)\right)+\Phi\left(-z_{\alpha/2}-\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)\right).$$

```
pow.z.test <- function(alpha,mu1,mu0,sigma,n){
  arg1 <- qnorm(1-alpha/2)-(mu1-mu0)/(sigma/sqrt(n))
  arg2 <- -1*qnorm(1-alpha/2)-(mu1-mu0)/(sigma/sqrt(n))
  1-pnorm(arg1)+pnorm(arg2)
}
```

Power of the one-sample z-test

For example the power of the test

$$H_0 : \mu = 0 \text{ versus } H_0 : \mu = 0.2$$

with $n = 30, \sigma = 0.2, \alpha = 0.05$ can be calculated by calling the above function.

```
pow.z.test(.05,.15,0,.2,30)
```

```
[1] 0.9841413
```

What does this mean?

Power of the one-sample t-test

Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. A test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_0 : \mu \neq \mu_0$$

will reject at level α if and only if

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \geq t_{n-1,\alpha/2},$$

where $t_{n-1,\alpha/2}$ is the $100(1 - \alpha/2)^{th}$ percentile of the t_{n-1}.

Power of the one-sample t-test

It can be shown that

$$\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] = \frac{Z+\gamma}{\sqrt{V/(n-1)}},$$

where,

$$Z = \frac{\sqrt{n}(\bar{X}-\mu_1)}{\sigma}$$
$$\gamma = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$$
$$V = \frac{(n-1)}{\sigma^2}S^2.$$

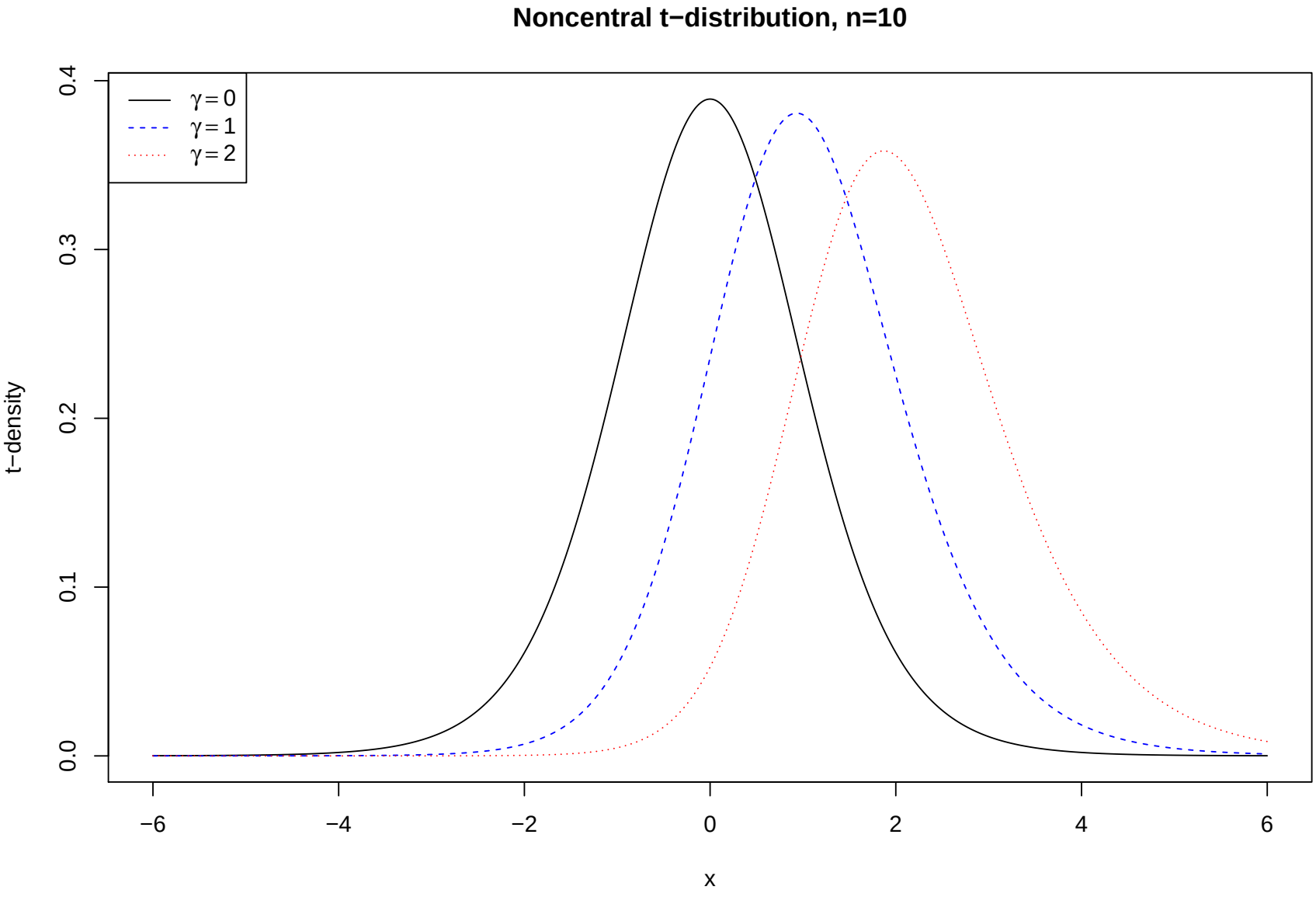
$Z \sim N(0,1)$ and $V \sim \chi^2_{n-1}$ and Z is independent of V.

Power of the one-sample t-test

- If $\gamma = 0$ then then $\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] \sim t_{n-1}$. This is sometimes called the **central t-distribution**.
- If $\gamma \neq 0$ then $\sqrt{n}\left[\frac{\bar{X}-\mu_0}{S}\right] \sim t_{n-1,\gamma}$, where $t_{n-1,\gamma}$ is the **non-central t-distribution** with non-centrality parameter γ.

Power of the one-sample t-test

A plot of the central ($\gamma = 0$) and non-central t ($\gamma = 1, 2$) are shown in the plot below.



Power of the one-sample t-test

The power of the test at $\mu = \mu_1$ is

$$
\begin{aligned}
1 - \beta &= 1 - P\,(\text{type II error}) \\
&= P\,(\text{Reject } H_0 | H_1 \text{ is true}) \\
&= P\,(\text{Reject } H_0 | \mu = \mu_1) \\
&= P\left(\left| \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right| \geq t_{n-1,\alpha/2} | \mu = \mu_1 \right) \\
&= P(t_{n-1,\gamma} \geq t_{n-1,\alpha/2}) + P(t_{n-1,\gamma} < -t_{n-1,\alpha/2})
\end{aligned}
$$

Power of the one-sample t-test

$$P(t_{n-1,\gamma} \geq t_{n-1,\alpha/2}) + P(t_{n-1,\gamma} < -t_{n-1,\alpha/2})$$

The following function calculates the power function for the one-sample t-test in R:

```
onesampttestpow <- function(alpha,n, mu0, mu1,sigma)
{delta <- mu1-mu0
t.crit <-qt(1-alpha/2,n-1)
t.gamma <- sqrt(n)*(delta/sigma)
t.power <- 1-pt(t.crit,n-1,ncp=t.gamma)
            +pt(-t.crit,n-1,ncp=t.gamma)
return(t.power)
}
```

Power of the one-sample t-test

The power of the t-test for testing

$$H_0 : \mu = 0 \text{ versus } H_0 : \mu = 0.15$$

with $n = 10, \sigma = 0.2, \alpha = 0.05$ can be calculated by calling the above function is

```
onesampttestpow(.05,10,0,.15,0.2)
```

```
[1] 0.5619339
```

Power of the one-sample t-test

Use the built-in function in R to calculate the power of t-test `power.t.test()`.

```
power.t.test(n = 10,delta = 0.15,sd = 0.2,
             sig.level = 0.05,type = "one.sample" )
```

```
     One-sample t test power calculation

              n = 10
          delta = 0.15
             sd = 0.2
      sig.level = 0.05
          power = 0.5619339
    alternative = two.sided
```

Power of the two-sample t-test

- Consider a two-sample comparison with continuous outcomes. Let Y_{ik} be the observed outcome for the i^{th} subject in the kth treatment group, for $i = 1, ..., n_k$, and $k = 1, 2$. The outcomes in the two groups are assumed to be independent and normally distributed with different means but an equal variance σ^2,

$$Y_{ik} \sim N(\mu_k, \sigma^2).$$

- Let $\theta = \mu_1 - \mu_2$, the difference in the mean between treatment 1 (the new therapy) and treatment 2 (the standard of care).
- To test whether the effects of the two treatments are the same, we formulate the null and alternative hypotheses as

$$H_0 : \theta = 0 \text{ versus } H_0 : \theta \neq 0.$$

Power of the two-sample t-test

- Consider a two-sample comparison with continuous outcomes. Let Y_{ik} be the observed outcome for the i^{th} subject in the kth treatment group, for $i = 1, ..., n_k$, and $k = 1, 2$. The outcomes in the two groups are assumed to be independent and normally distributed with different means but an equal variance σ^2,

$$Y_{ik} \sim N(\mu_k, \sigma^2).$$

- Let $\theta = \mu_1 - \mu_2$, the difference in the mean between treatment 1 (the new therapy) and treatment 2 (the standard of care).
- To test whether the effects of the two treatments are the same, we formulate the null and alternative hypotheses as

$$H_0 : \theta = 0 \text{ versus } H_0 : \theta \neq 0.$$

Power of the two-sample t-test

The two-sample t statistic is given by

$$T_n = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2}.$$

- $T_n \sim t_{n_1+n_2-2}$ under H_0
- $T_n \sim t_{n_1+n_2-2,\gamma}$ with noncentrality parameter

$$\gamma = \frac{\mu_1 - \mu_2}{\sigma\sqrt{1/n_1 + 1/n_2}},$$

under H_1.

Power of the two-sample t-test

H_0 is rejected if

$$|T_n| \geq t_{n_1+n_2-2,\alpha/2},$$

where $t_{df,\alpha/2}$ is the $100(1-\alpha/2)th$ percentile of the central t-distribution with df degrees of freedom. (Yin, pg. 164-165)

- use `t.test()` to do the calculations.

Power of the two-sample t-test

The power of the test is

$$1 - \beta = 1 - P(t_{n_1+n_2-2,\gamma} \geq t_{n_1+n_2-2,\alpha/2}) + P(t_{n_1+n_2-2,\gamma} < -t_{n_1+n_2-2,\alpha/2})$$

The sample size can be solved from this equation which does not have a closed form.

The sample size can be determined by specifying:

- type I and type II error rates,
- the standard deviation,
- the difference in treatment means that the clinical trial aims to detect.

Power of the two-sample t-test

$$1-\beta = 1 - P(t_{n_1+n_2-2,\gamma} \geq t_{n_1+n_2-2,\alpha/2}) + P(t_{n_1+n_2-2,\gamma} < -t_{n_1+n_2-2,\alpha/2})$$

```
twosampttestpow <- function(alpha,n1,n2, mu1, mu2,sigma){
delta <- mu1-mu2
t.crit <-qt(1-alpha/2,n1+n2-2)
t.gamma <- delta/(sigma*sqrt(1/n1+1/n2))
t.power <- 1-pt(t.crit,n1+n2-2,ncp=t.gamma)+
              pt(-t.crit,n1+n2-2,ncp=t.gamma)
return(t.power)
}
```

Power of the two-sample t-test

A clinical trial to test a new treatment against the standard treatment for colon cancer is being designed. The investigators feel that the smallest meaningful difference in tumour growth is 1cm. The standard deviation of tumour growth is 3cm. The investigators feel that they can enrol 50 subjects per arm. Will this clinical trial have adequate power to detect a difference between the treatments?

- What are the parameters of interest?
- What are the null and alternative hypotheses?
- How can the power of the study be calculated?

Power of the two-sample t-test

```
twosampttestpow(.05,50,50,1,2,3)
```

```
[1] 0.3785749
```

Power of the two-sample t-test

- `power.t.test()` can calculate the number of subjects required to achieve a certain power.
- Suppose the investigators want to know how many subjects per group would have to be enrolled in each group to achieve 80% power under the same conditions?

```
power.t.test(power = 0.8,delta = 1,sd = 3,sig.level = 0.05)
```

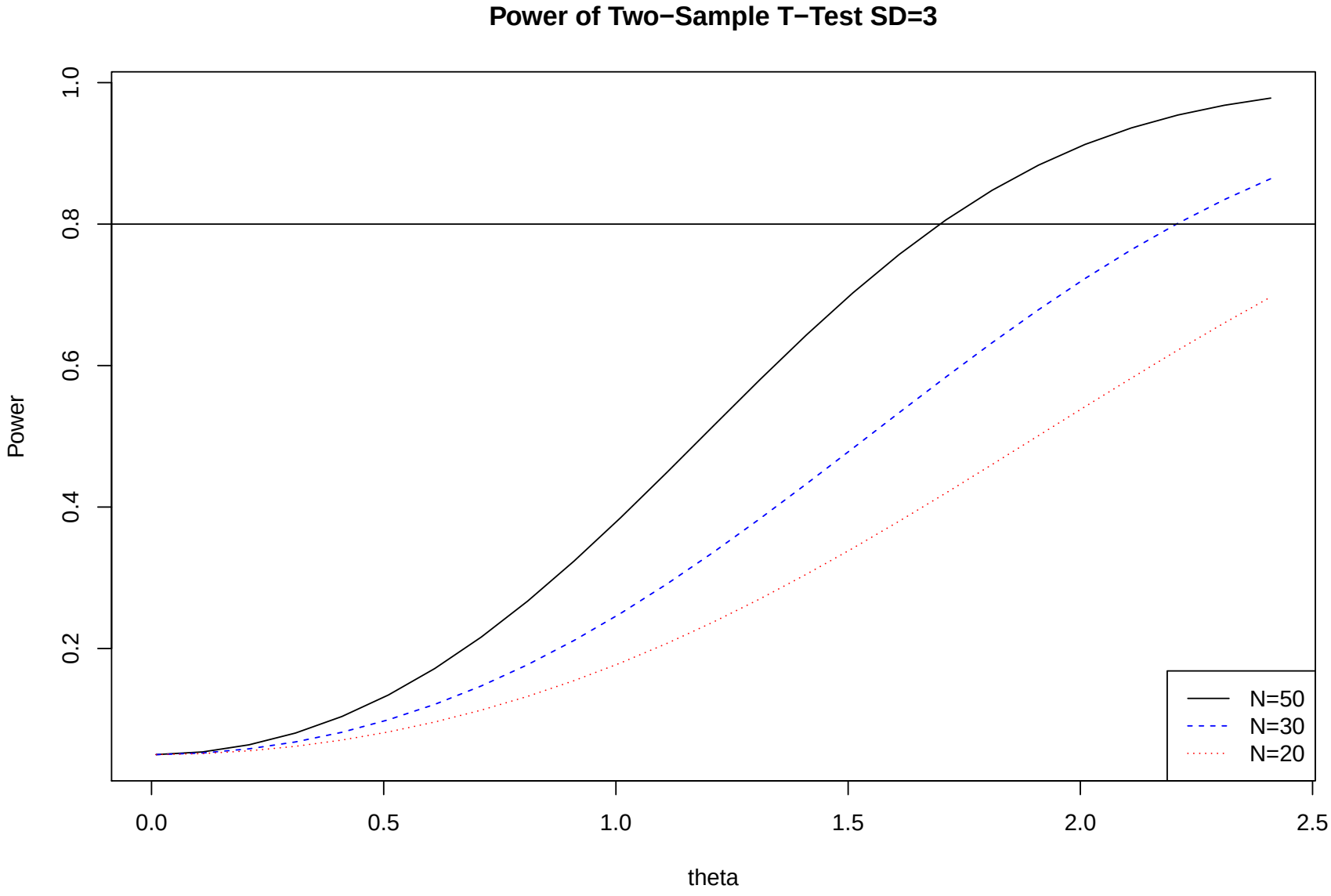
```
     Two-sample t test power calculation

              n = 142.2466
          delta = 1
             sd = 3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

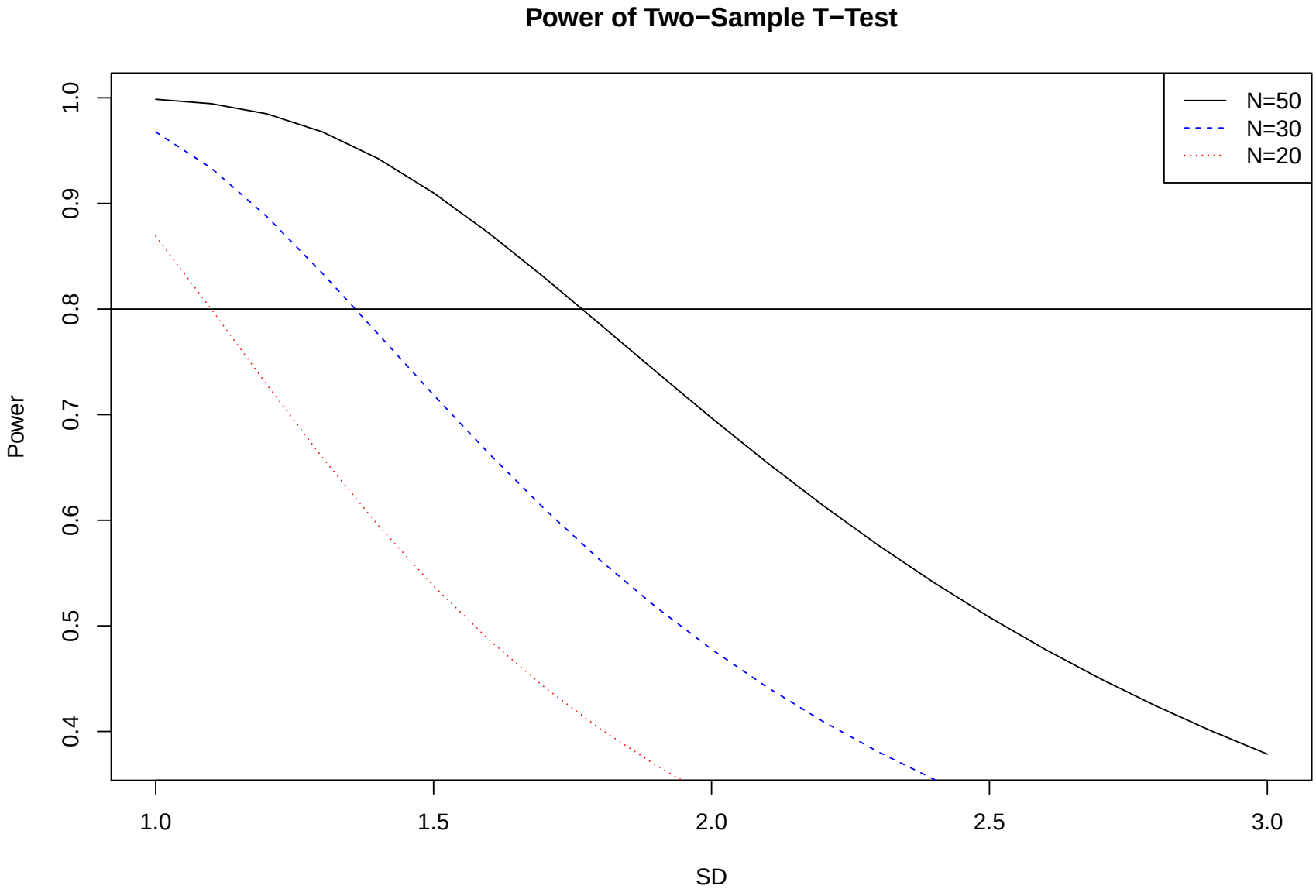
Power of the two-sample t-test

The following plot shows power of the two-sample t-test as a function of the difference $\theta = \mu_1 - \mu_2$ to be detected and equal sample size per group.



Power of the two-sample t-test

This plot shows power as a function of σ and sample size per group.



Power of the two-sample t-test

In some studies instead of specifying the difference in treatment means and standard deviation separately the ratio

$$ES = \frac{\mu_1 - \mu_2}{\sigma}$$

can be specified.

- ES is called the scaled effect size.
- Cohen (1992) suggests that effect sizes of 0.2, 0.5, 0.8 correspond to small, medium , and large effects respectively.

Power of the two-sample t-test

Power as a function of effect size can be investigated.

The plot shows that for $n_1 = n_2 = 10$ the two-sample t-test has at least 80% power for detecting effect sizes that are at least 1.3.