Power Analysis

11/04/2022 (Week 11)

Jingjing Yang, PhD

Assistant Professor of Human Genetics

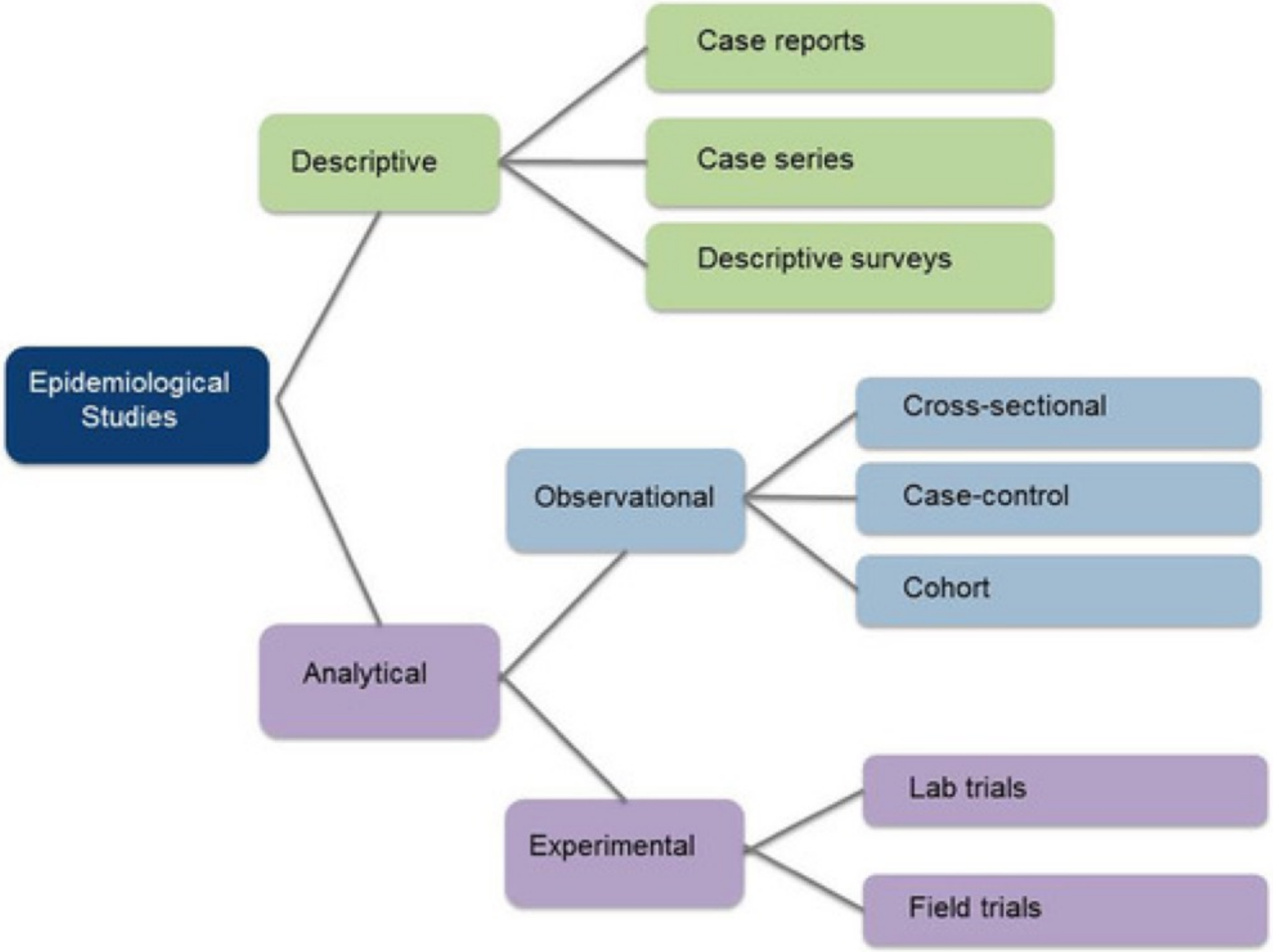
Jingjing.yang@emory.edu

Outline

- Study Design
- Review Type I and Type II Error
- Power
- Power Analysis

Study Design

https://s4be.cochrane.org/blog/2021/04/06/an-introduction-to-different-types-of-study-design/



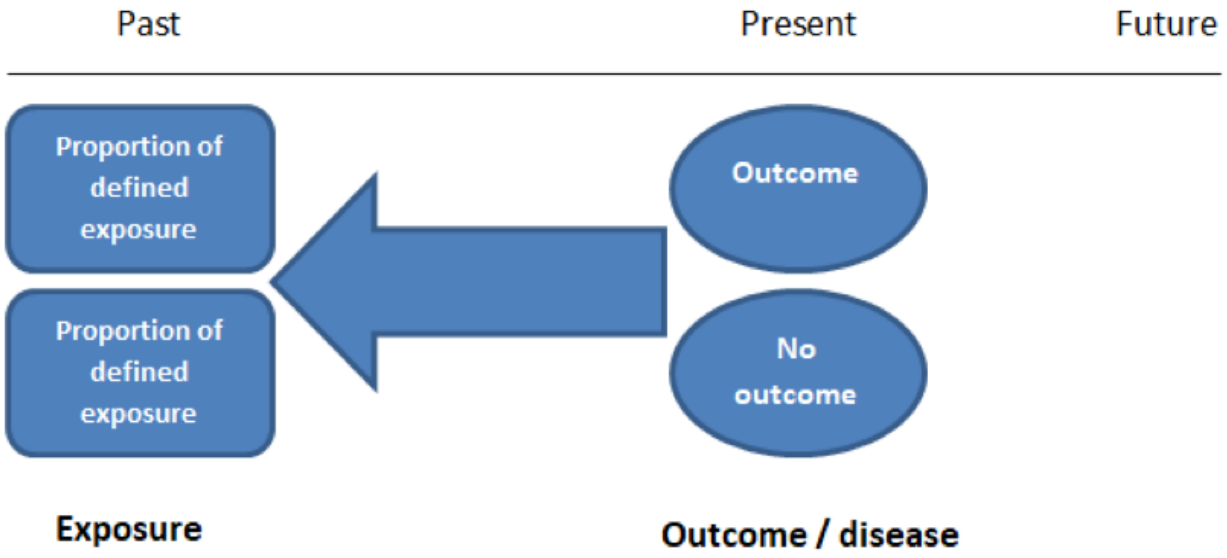
Case-Control Study

- We conduct this study by comparing 2 groups: one group with the disease (**cases**) and another group without the disease (**controls**)
- This design is always **Retrospective**
- We aim to find out the odds of having a risk factor or an exposure if an individual has a specific disease **(Odds ratio)**
- Relatively easy to conduct

Case-Control Study

For example

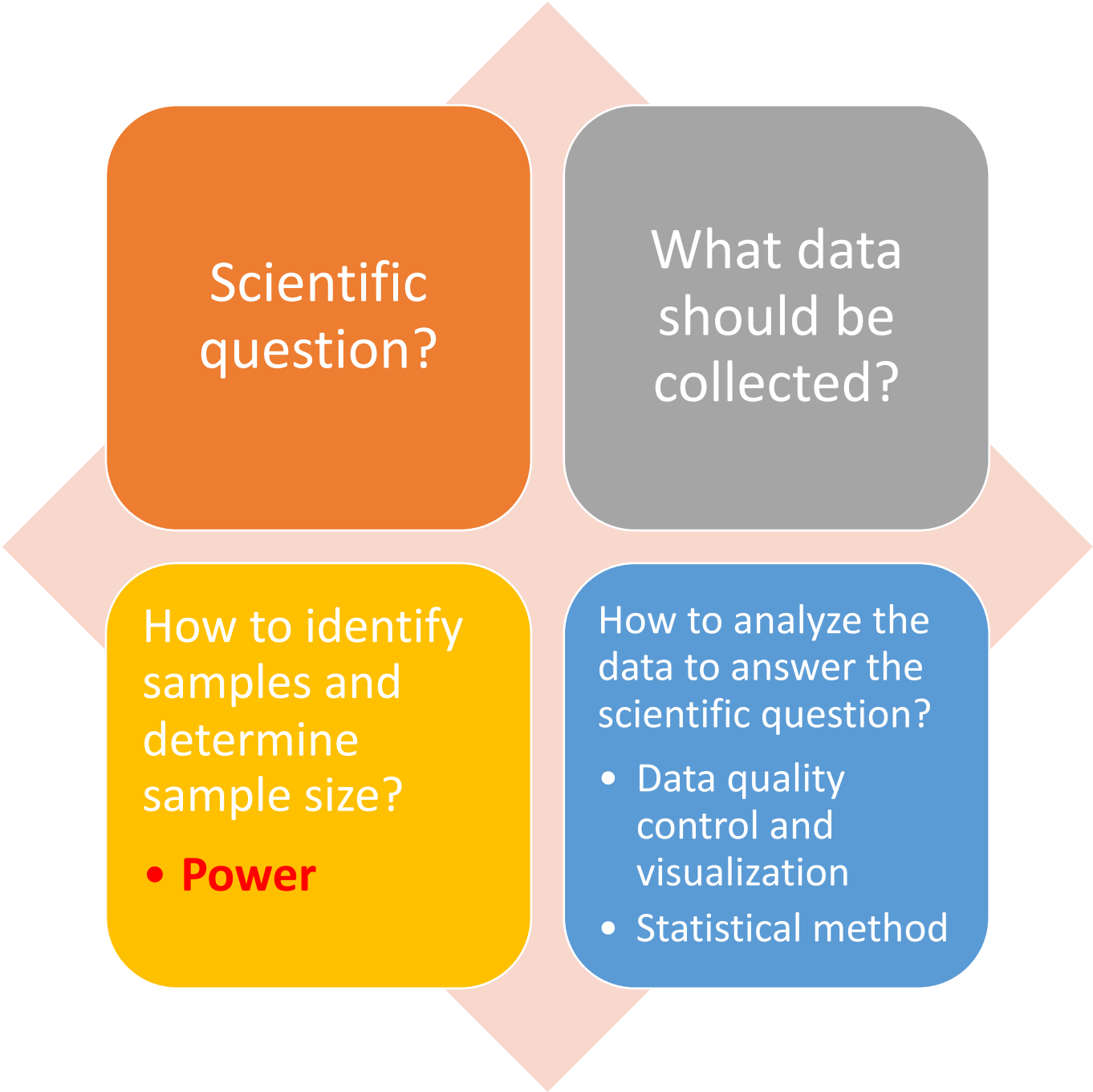
- we want to study the odds of being a smoker among hypertensive patients compared to normotensive ones.
- To do so, we choose a group of patients diagnosed with hypertension and another group that serves as the control (normal blood pressure).
- Then we study their smoking history to find out if there is a correlation.



Experimental Studies

- Also known as interventional studies
- Can involve animals and humans
- Pre-clinical trials involve animals
- Clinical trials are experimental studies involving humans
- In randomized controlled trials, one group of participants receives the control, while the other receives the tested drug/intervention. Those studies are the best way to evaluate the efficacy of a treatment.

Questions need to address?



How important is statistical power?

- Important for funding
  - Needed for a successful grant application
  - NIH applications often rejected due to:
    - Low power (sample size too small)
    - Lack of power calculations
    - Inappropriate power calculations
  - Why do NIH reviewers care?
- Important for success of study
  - Underpowered studies are less likely to yield significant findings
  - When they do yield significant results, much more likely to be false positives
- So power is a primary concern in every analysis we do.
  - Second most important thing, after validity

Hypothesis Testing

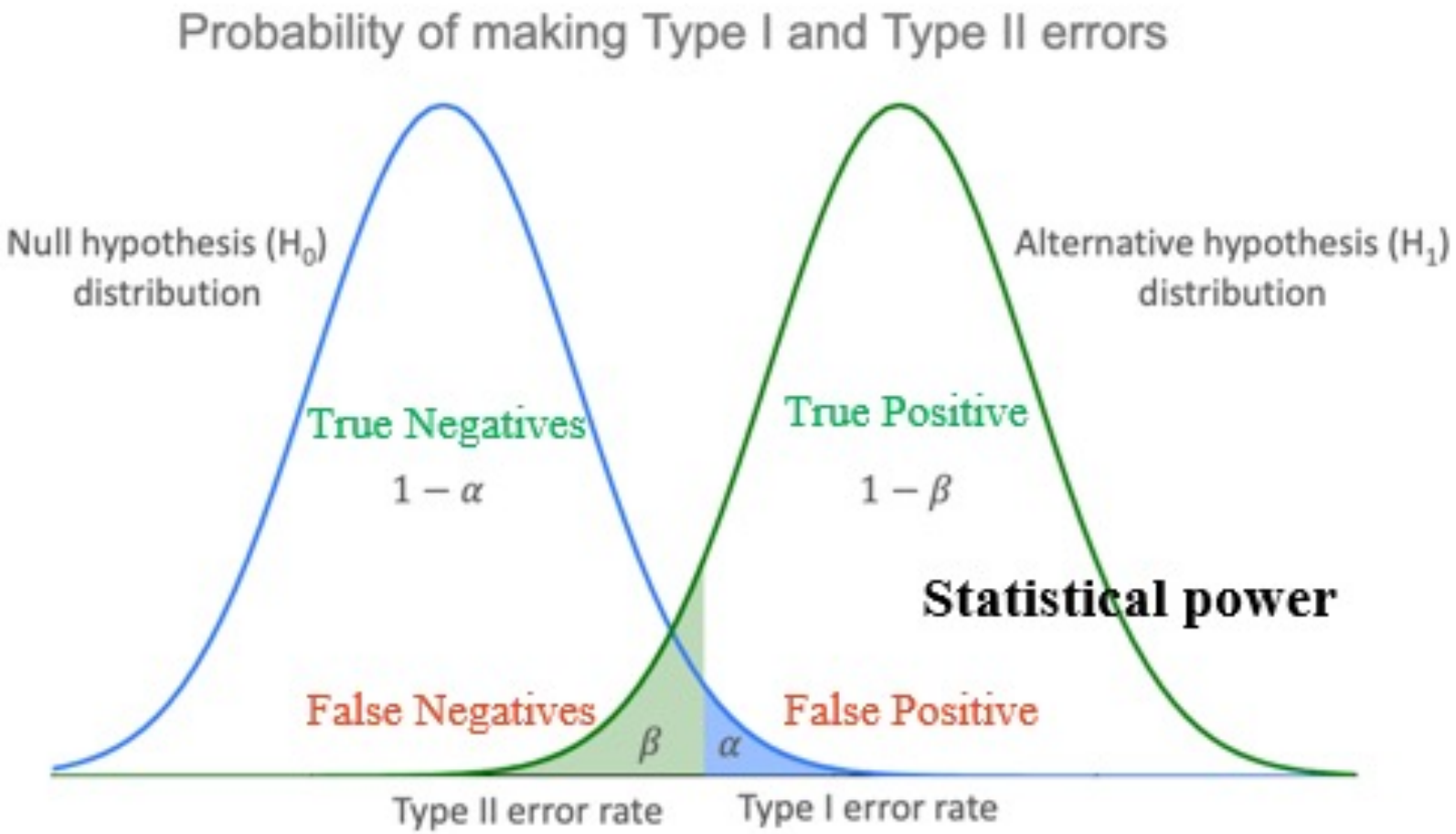
- Null Hypothesis: H_0
- Alternative Hypothesis: H_a
- Test Statistic: T

1. Determine the distribution (probability density function) of T under H_0
2. Calculate the value of test statistic T_{obs} with observed data
3. Calculate p-value under H_0:
   $Prob(T\ is\ as\ or\ more\ exterem\ than\ T_{obs}|H_0)$
4. Reject H_0 if p-value < α.

Type I and Type II Error

Null hypothesis is...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = 1−β
Not rejected	**Correct decision** True negative Probability = 1−α	**Type II error** False negative Probability = β

Hypothesis Testing



Which Test Statistic/Method to Use?

- Dose lower Type I Error mean lower Type II Error as well?
- Control for Type I Error?
- Best power (lowest Type II Error)?
- Should be appropriate for your data type!

Power

Power

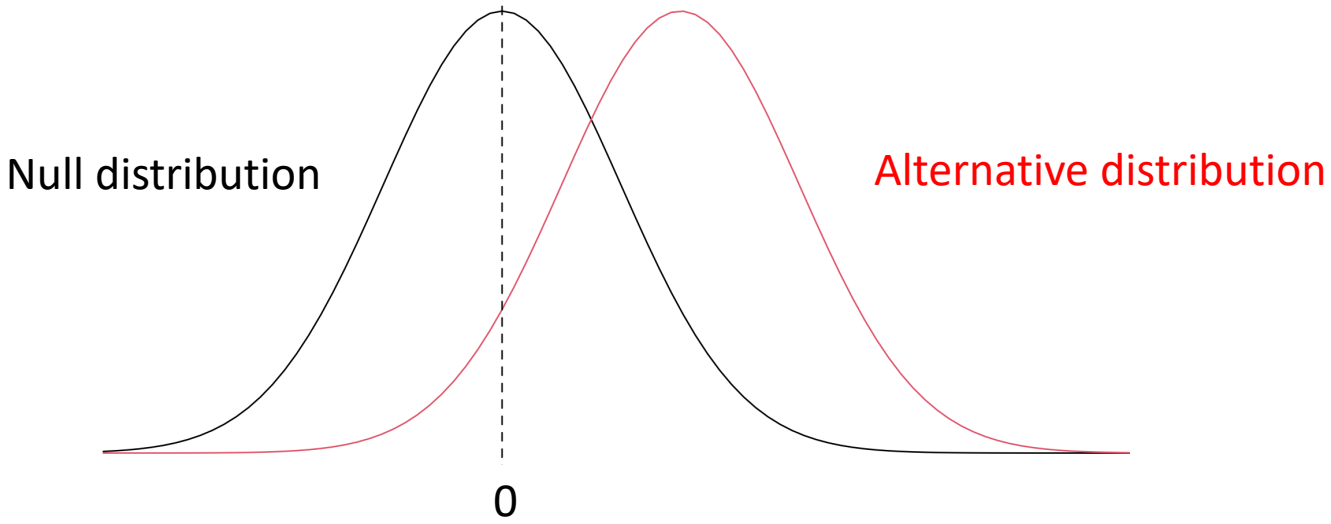
The Probability to Reject Null Hypothesis when the Null Hypothesis is Truly FALSE (True Positive Probability)

$$\text{Power} = 1 - \text{Type II Error } (\beta)$$

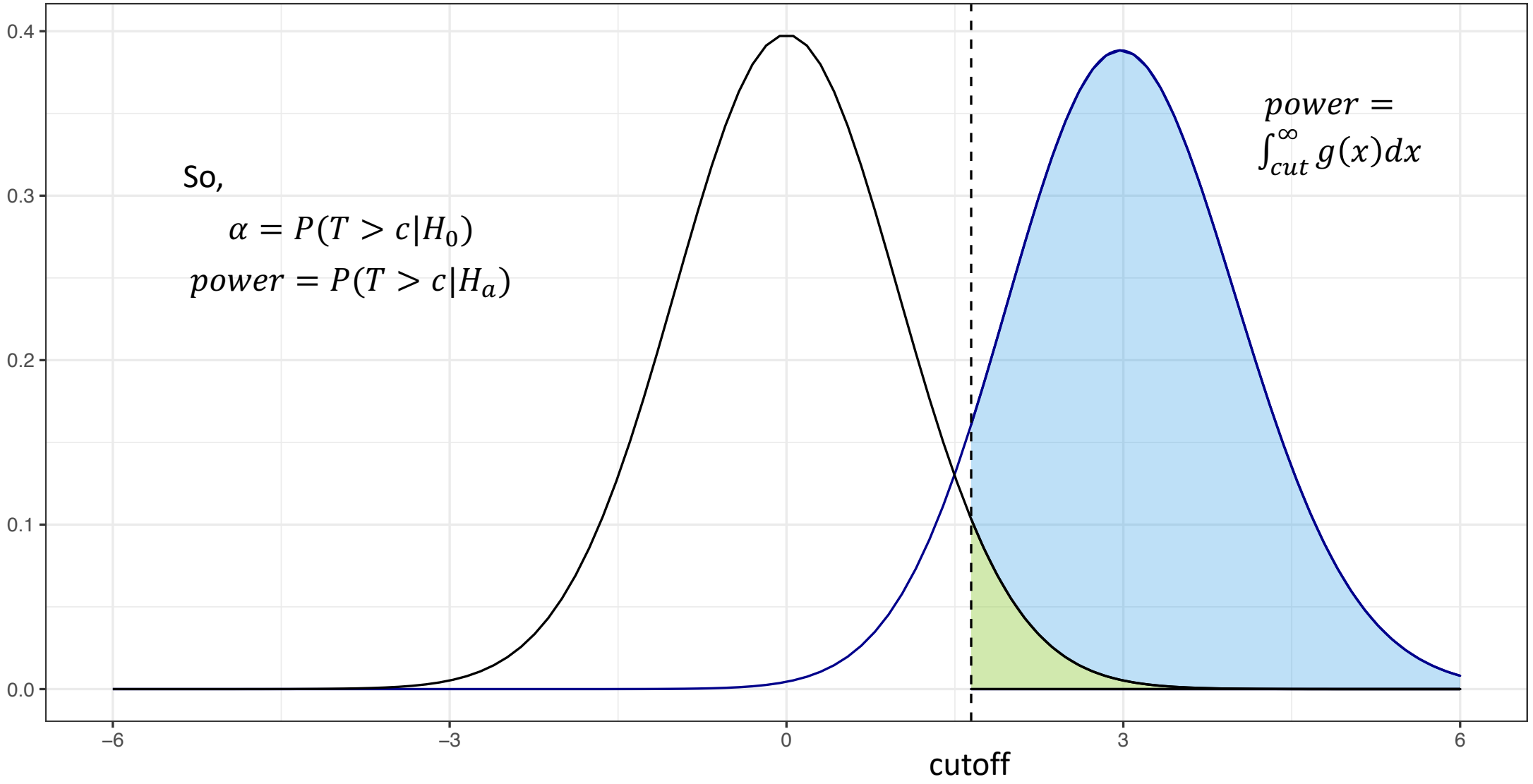
- Chance of detecting a real/true effect
- Analytical power function would be known if we know the probability density distribution of test statistic under H_a

Alternate distribution

- T-test example: $T = \frac{\bar{X}_1 - \bar{X}_2}{s/\sqrt{n}}$
  - s is standard deviation of outcome data
  - n is sample size
- How is *T* distributed under the alternative hypothesis?
- Alternative hypothesis is that $\bar{X}_1 - \bar{X}_2 > 0$. But this is very non-specific!
- If we **assume a particular effect size**, it is possible to derive the distribution of *T assuming that effect size is correct*.

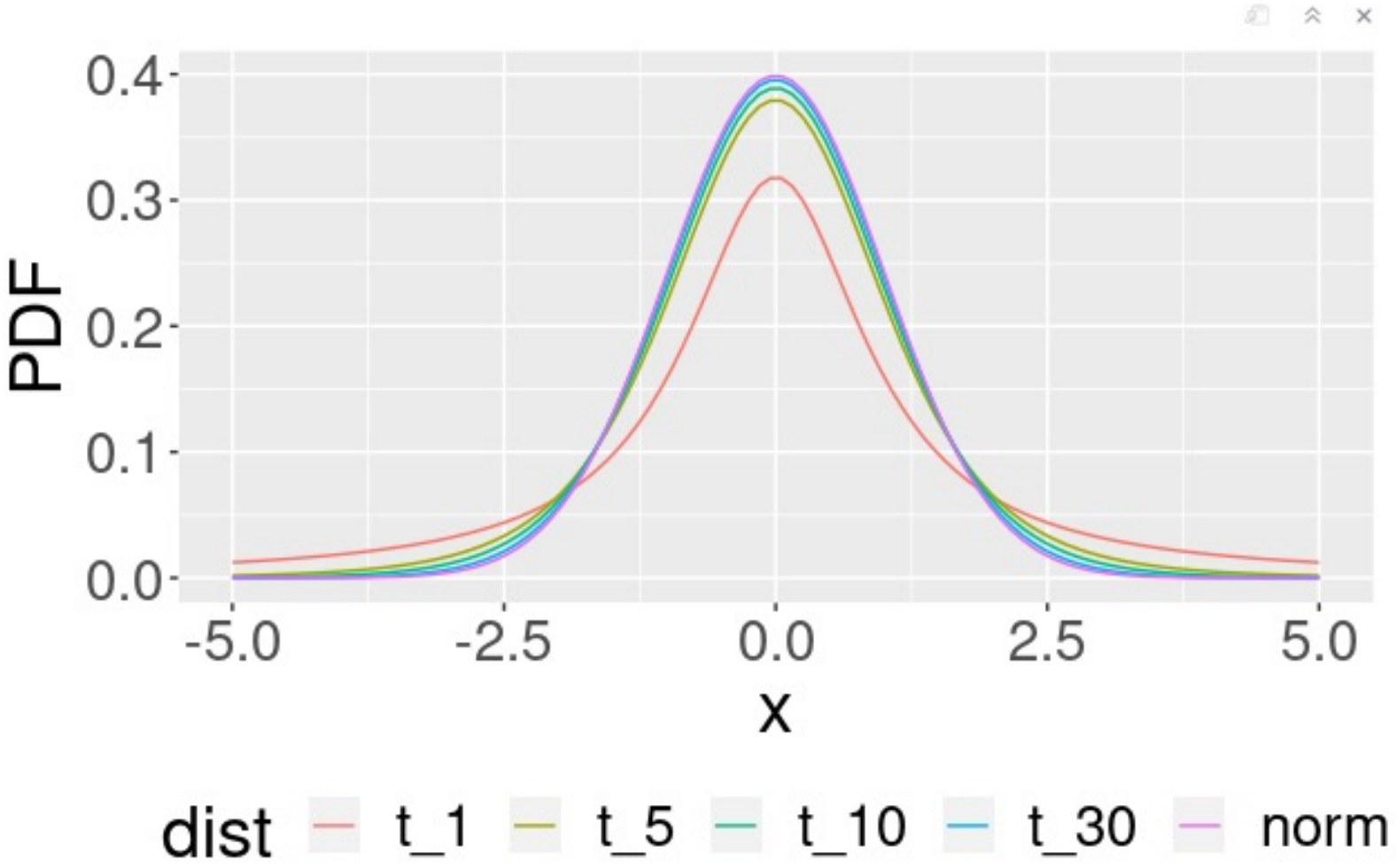


Power can be computed as AUC from alternative distribution



T-distribution and the standard normal distribution

T-distribution would converge to a standard normal distribution as the degrees of freedom increase!



What factors influence the power of a study?

1. Sample size
2. True effect size (size of difference or association in the population)
   - Mean difference
   - Slope
   - Correlation (r or r^2)
3. Significance criterion – what α-level are we using?

- **The alternate distribution depends on true effect size and sample size**
- **The significance criterion determines how easy/hard it is to reject**

A General Formula for Power

$$Power = 1 - \beta \propto \frac{ES\ \alpha\sqrt{n}}{\sigma}$$

- ES : Effect Size
- α : Type I Error
- n : Sample Size
- σ : Out Come Standard Deviation

$$n \propto \left(\frac{\sigma\ (1\ - \beta)}{\alpha\ ES}\right)^2$$

$$ES\ \propto \frac{(1-\beta)\sigma}{\alpha\sqrt{n}}$$

Effect Size (ES)

Type of Test	Effect Size (ES)
One-sample t-test	Sample mean
Two-sample t-test	Differences between means
Linear regression	Slope/coefficient
One-way ANOVA	Differences between means

Review of effect sizes

Effect size (useful for interpretation)	Test statistic (useful for inference)
Estimate of the underlying "true value"	Variance-standardized version of this estimate
Scale may vary, may depend on units of measurement	Standardization solves this problem
Example: group difference in means $\bar{X}_1 - \bar{X}_2$	Example: T-statistic $\frac{\bar{X}_1 - \bar{X}_2}{s/\sqrt{n}}$
Other examples: correlation coefficient, regression slope, odds ratio, Cohen's d (Cohen's d = $(\bar{X}_1 - \bar{X}_2)/s$)	Other examples: F-statistic, chi-squared statistic, p-value
Does not depend on N under null or alternative	If alternative is true, becomes more extreme as N increases (why?)

When Power Function is Analytically Known

R library “pwr”

- pwr.t.test() for t-test

For t-tests, the effect size is assessed as

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

where μ_1 = mean of group 1
μ_2 = mean of group 2
σ^2 = common error variance

Cohen suggests that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

Input/Output Arguments for pwr.t.test():

- n : Sample Size
- d : Cohen’s Effect Size (standardized), difference between the means divided by the pooled standard deviation
- power : Power
- sig.level: significance level, e.g., 0.05 by default
- type : type of test, e.g., "two.sample", "one.sample", "paired"
- alternative : “two.sided”, “less”, “greater”

Details

Exactly one of the parameters 'd','n','power' and 'sig.level' must be passed as NULL, and that parameter is determined from the others. Notice that the last one has non-NULL default so NULL must be explicitly passed if you want to compute it.

Power Analysis

Example 1: One-sample t-test

Example 1: Determine the power
For given sample size n and effect size d

```
### Find the power for given sample size and effect size
```{r}
pwr.t.test(n = 100, d = 0.29, sig.level = 0.05,
                          type = "one.sample", alternative = "two.sided")
```
```

```
     One-sample t test power calculation

              n = 100
              d = 0.29
      sig.level = 0.05
          power = 0.8190623
    alternative = two.sided
```

Example 1: One simulation with sample size 100 and mean (effect size) 0.29

```{r}
n = 100
# Simulate x variable values
x = rnorm(n, mean = 0.29, sd = 1)
t.test(x)
```

```
        One Sample t-test

data:  x
t = 3.3162, df = 99, p-value = 0.001276
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1297010 0.5161124
sample estimates:
mean of x
0.3229067
```

```{r}
# sample mean
x_mean = mean(x)

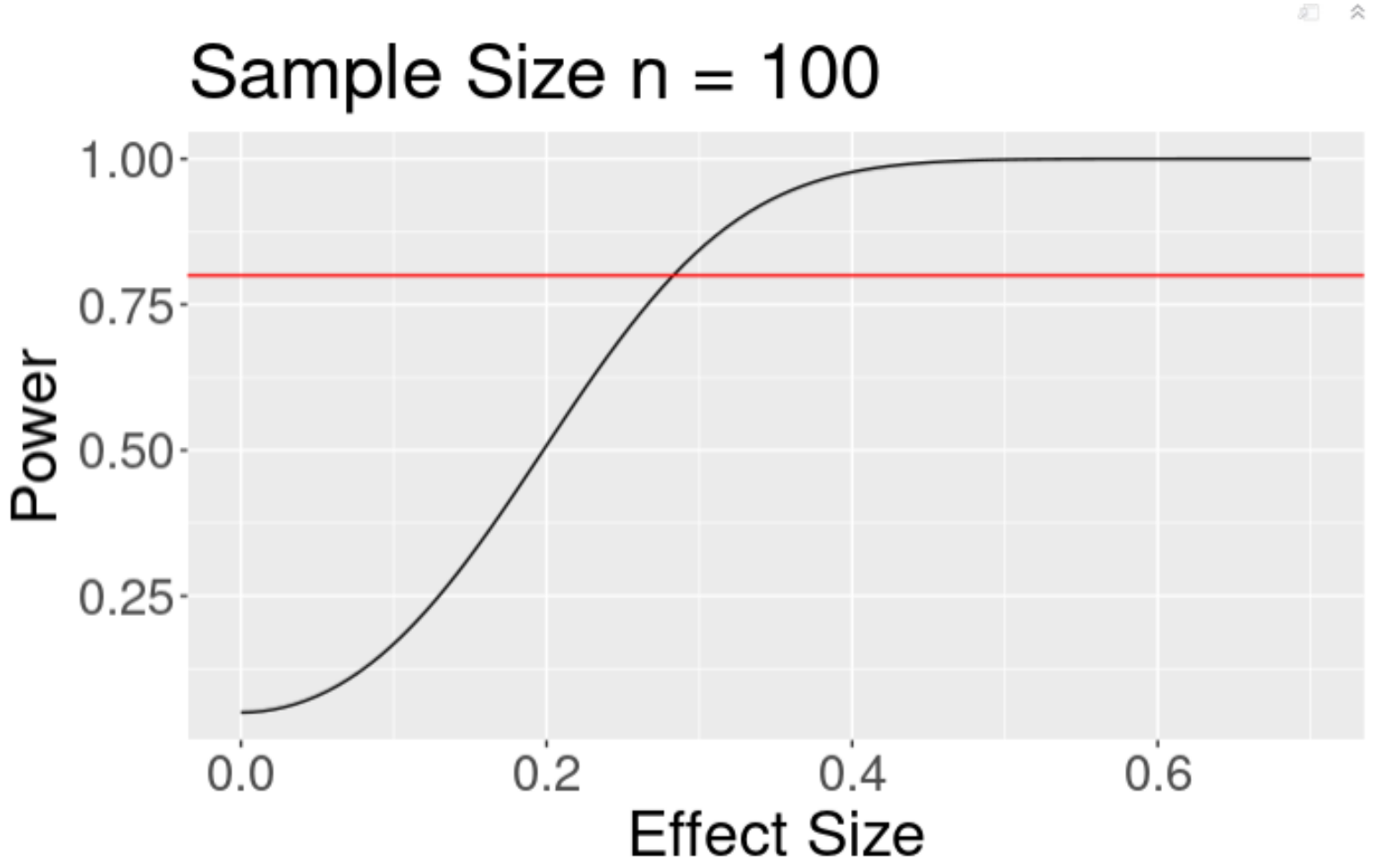
# sample deviation
x_sd = sd(x)

# t-test statistic value
t_stat = x_mean / (x_sd / sqrt(n))

print(c(x_mean, x_sd, t_stat))
```

```
[1] 0.3229067 0.9737125 3.3162425
```

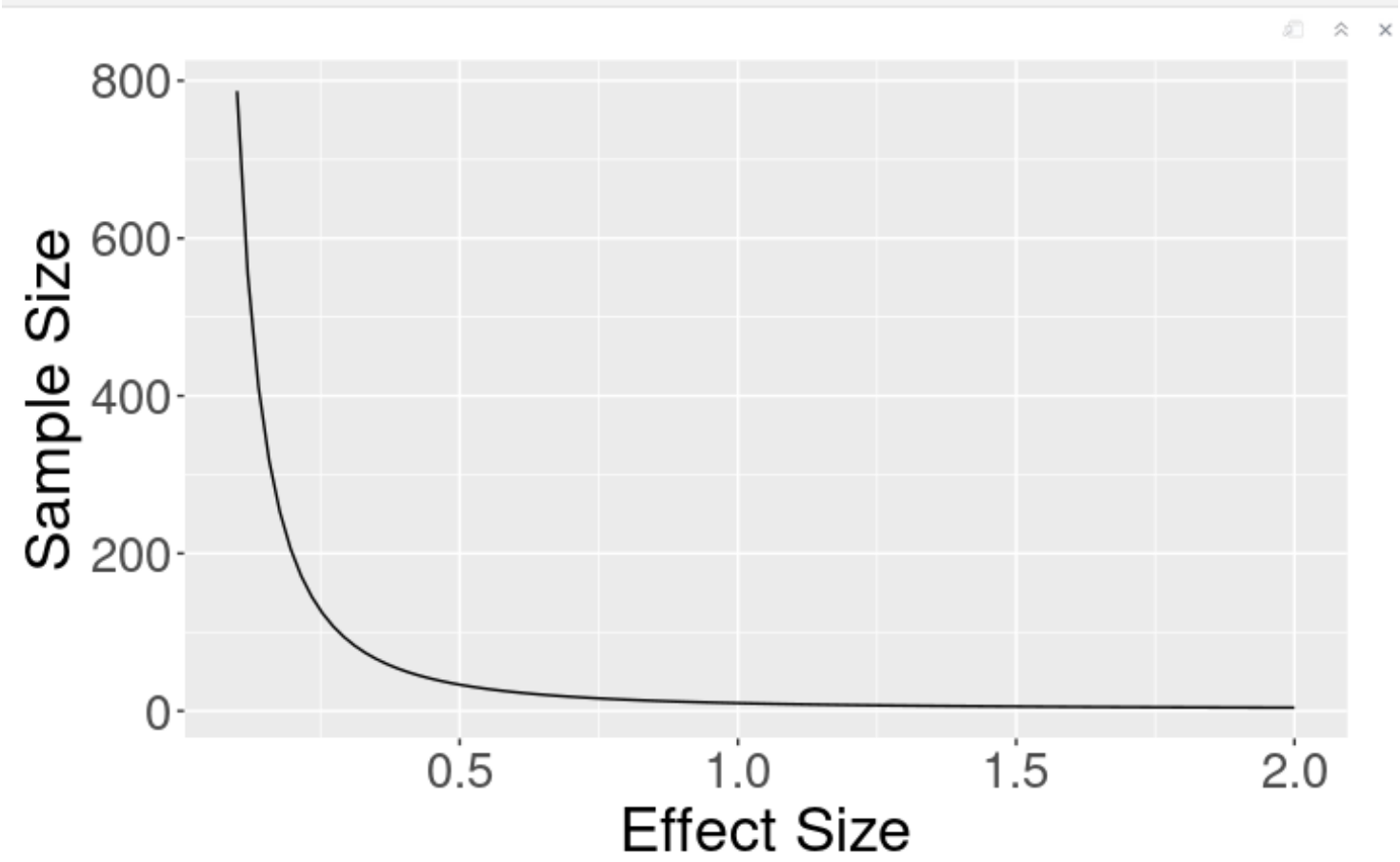
Example 1: Evaluate power along effect sizes with fixed sample size

```
#### Evaluate Power
```{r}
n_vec = 100
ES_vec = seq(0, 0.7, length.out = 100)
power_vec = rep(NA, 100)

# d is the Effect Size
for(i in 1:100){
  # Find desired sample size
  power_vec[i] = pwr.t.test(n = 100, d = ES_vec[i],
                            sig.level = 0.05,
                          type = "one.sample", alternative =
"two.sided")$power
}

qplot(ES_vec, power_vec, geom = "line") +
  labs(x = "Effect Size", y = "Power", title = "Sample Size n = 100") +
  geom_hline(yintercept = 0.8, colour = "red")
```
```

Example 1: Evaluate power along effect sizes with fixed sample size



Example 1: Determine sample size for targeted 80% power

```
### Find desired sample size
```{r}
ES_vec = seq(0.1, 2, length.out = 100)
n_vec = rep(NA, 10)

# d is the Effect Size
for(i in 1:100){
  # Find desired sample size
  n_vec[i] = pwr.t.test(d = ES_vec[i], sig.level = 0.05, power = 0.8,
                        type = "one.sample", alternative = "two.sided")$n
}

qplot(ES_vec, n_vec, geom = "line") + labs(x = "Effect Size", y = "Sample
Size")
```
```

Example 1: Determine sample size for targeted 80% power

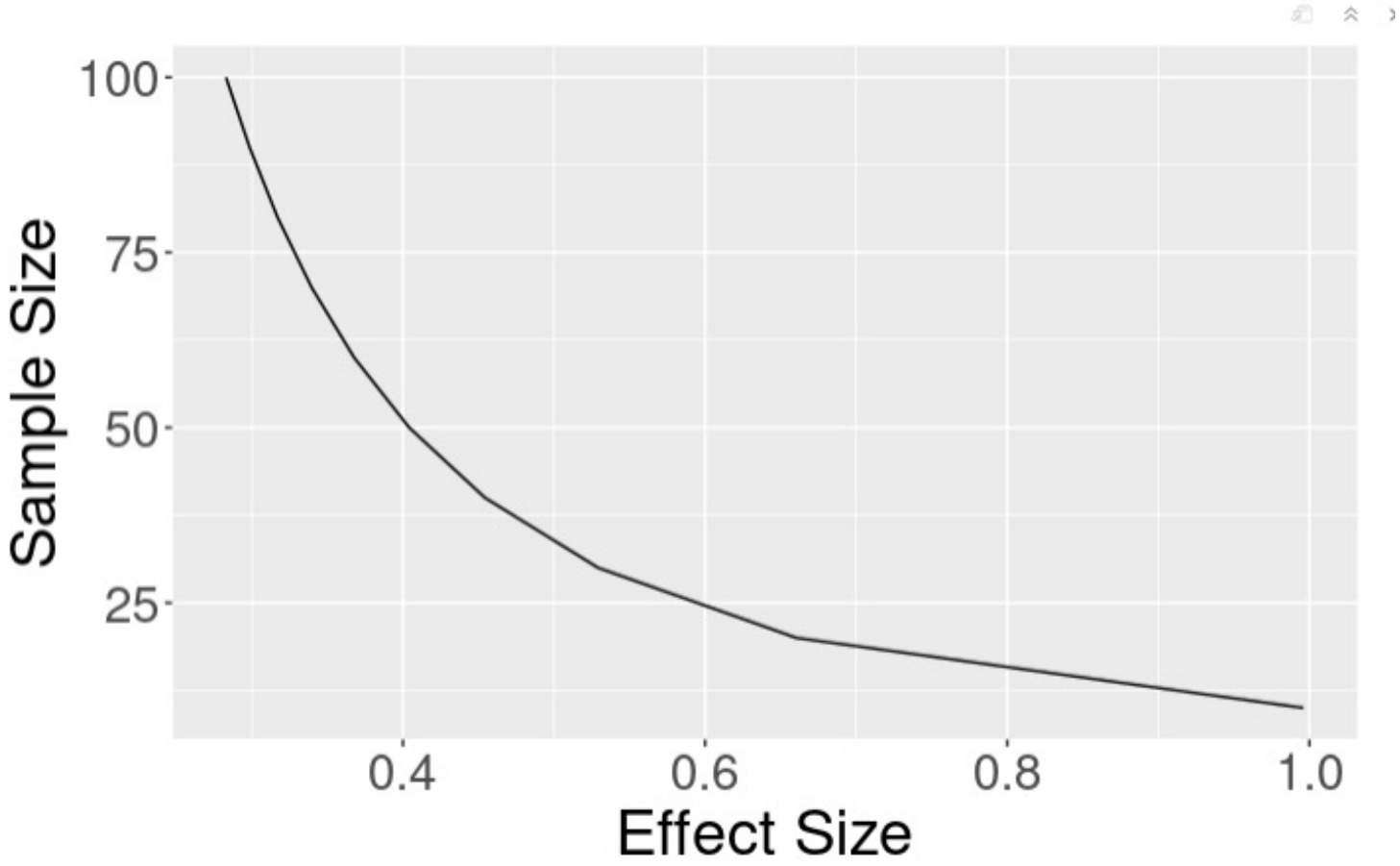


Example 1: Determine effect size for targeted 80% power

Find desired effect size

```{r}
n_vec = seq(10, 100, by = 10)
ES_vec = rep(NA, 10)

# d is the Effect Size
for(i in 1:10){
  # Find desired sample size
  ES_vec[i] = pwr.t.test(n = n_vec[i], sig.level = 0.05, power = 0.8,
                         type = "one.sample", alternative = "two.sided")$d
}

qplot(ES_vec, n_vec, geom = "line") + labs(x = "Effect Size", y = "Sample
Size")
```

Example 1: Determine effect size for targeted 80% power



Example 2: Balanced one-way ANOVA

`pwr.anova.test` *Power calculations for balanced one-way analysis of variance tests*

Description

Compute power of test or determine parameters to obtain target power (same as power.anova.test).

Usage

```
pwr.anova.test(k = NULL, n = NULL, f = NULL, sig.level = 0.05, power = NULL)
```

Arguments

`k`	Number of groups
`n`	Number of observations (per group)
`f`	Effect size
`sig.level`	Significance level (Type I error probability)
`power`	Power of test (1 minus Type II error probability)

For a one-way ANOVA effect size is measured by f where

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i * (\mu_i - \mu)^2}{\sigma^2}}$$

where $p_i = n_i / N$,
n_i = number of observations in group i
N = total number of observations
μ_i = mean of group i
μ = grand mean
σ^2 = error variance within groups

Cohen suggests that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes respectively.

Example 2: Balanced one-way ANOVA

- Additional reading about how to determine sample size for functions in the R library ”pwr”:
  - https://www.statmethods.net/stats/power.html

Details

Exactly one of the parameters 'k','n','f','power' and 'sig.level' must be passed as NULL, and that parameter is determined from the others. Notice that the last one has non-NULL default so NULL must be explicitly passed if you want to compute it.

Value

Object of class '"power.htest"', a list of the arguments (including the computed one) augmented with 'method' and 'note' elements.

Example 2: Determine power with effect size 0.28, 4 groups, 20 samples per group

```r
# Determine Power for given effect size f=0.28, 4 groups, 20 samples per group
pwr.anova.test(f=0.28, k=4, n=20, sig.level=0.05)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 20
              f = 0.28
      sig.level = 0.05
          power = 0.5149793

NOTE: n is number in each group
```

Example 2: Determine sample size with effect size 0.28, 4 groups, targeted power 80%

```{r}
# Determine required sample size for given effect size f=0.28, 4 groups, power = 0.8
pwr.anova.test(f=0.28, k=4, power=0.80, sig.level=0.05)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 35.75789
              f = 0.28
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Experiment Design (Power Analysis)

- Choose appropriate test statistic/method
- Choose significance threshold: type I error tolerance rate
- Decide expected effect sizes based on preliminary data or previous results or pilot studies
- Plot power function vs. sample size, grouped by different effect sizes
- Goal
  - $\alpha \leq 0.05$ and sample size to achieve 80% power (β=20%) in biomedical research
  - $\alpha \leq 0.05$ and sample size to achieve 90% power (β=10%) in clinical trials
  - Lower significance threshold is needed for α when multiple testing exists

In-Class Exercise 1

Power Analysis for t-test by pwr.t.test() function

When Power Function is not Analytically Known

Monte Carlo Simulations

Simulation can also be used to evaluate type I error

Evaluate Power by Simulation

Simulate

Simulate outcome data under the Alternative hypothesis

- Calculate test statistic value
- Calculate p-value, and decide whether to reject the current simulation

Repeat

Repeat such simulation for a large number of iterations, e.g., 1000

Calculate

Calculate the proportion of simulations (POWER) that you indeed reject the null hypothesis

What would be your desired Power?

Evaluate Power by Simulation

With sample size 100, effect size 0.29

````
### Simulation to evaluate power for effect size = 0.29
```{r}
m = 1000 # number of simulation
n = 100 # sample size
sig_m = 0
alpha = 0.05
ES = 0.29
set.seed(123)
# Simulate x variable values for m times, conduct t.test per simulation
for (i in 1:m){
  x = rnorm(n, mean = ES, sd = 1) # generate x under Ha
  x_mean = mean(x) # sample mean
  x_sd = sd(x) # sample deviation
  t_stat = x_mean / (x_sd / sqrt(n)) # t-test statistic value
  # Calculate p-value
  pvalue_t_test = pt(abs(t_stat), df = n - 1, lower.tail = FALSE) * 2
  if(pvalue_t_test < alpha){
    sig_m = sig_m + 1 # if significant
  }
}
# Power
print(sig_m / m)
```
````

```
[1] 0.844
```

Evaluate Type I Error by Simulation

- Simulate outcome data under the NULL hypothesis
  - Calculate test statistic value
  - Calculate p-value
  - Decide whether to reject the current simulation
- Repeat such simulation for a large number of times
  - The total number of simulations will need to be determined based on the significance level you use
  - e.g., 1000 times with alpha = 0.05, 10000 times with alpha = 0.005
- Calculate the proportion of simulations (Type I Error) that you indeed reject the null hypothesis

What would be your desired Type I Error?

Evaluate Type I Error by Simulation

With sample size 100, effect size 0

Simulation to evaluate type I error

```{r}
m = 1000 # number of simulation
n = 100 # sample size
sig_m = 0
alpha = 0.05

set.seed(123)
# Simulate x variable values for m times, conduct t.test per simulation
for (i in 1:m){
  x = rnorm(n, mean = 0, sd = 1) # generate x under NULL
  x_mean = mean(x) # sample mean
  x_sd = sd(x) # sample deviation
  t_stat = x_mean / (x_sd / sqrt(n)) # t-test statistic value
  # Calculate p-value
  pvalue_t_test = pt(abs(t_stat), df = n - 1, lower.tail = FALSE) * 2
  if(pvalue_t_test < alpha){
    sig_m = sig_m + 1 # if significant
  }
}

# Type I Error
print(sig_m / m)
```

```
[1] 0.032
```

In-Class Exercise 2

Next Lectures

- Week 12 (11/11): Adjust for multiple tests (Karen)
  - Homework 7 due 11/11, revision due 11/17
- Week 13 (11/18): Permutation test (Jingjing)
  - Homework 8 distributed and due 12/01
- Week 14 (12/02): Machine learning (Jingjing)
  - Homework 9 distributed and due 12/08
- Grading for homework 8 & 9
  - Submitted and showed your work for all tasks: 10
  - Missed: 0
  - Feedbacks will still be provided