# Designing a clinical study: power considerations

## Fundamentals of Translational Oncology Workshop

Marcio Augusto Diniz, Ph.D.
Biostatistics and Bioinformatics Research Center
Cedars Sinai Medical Center

June 30, 2021

# Summary

## Study Design

- Feasibility/Acceptability studies;
- Phase I dose finding studies;
- Single-arm or two-arm phase II studies;
- Cluster randomized trials;
- Adaptive designs;
- Biomarker studies (Development and Validation);
- Predictive modeling;
- Surveys;
- and so on.

# Introduction

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Figure: Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR. Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews neuroscience. 2013 May;14(5):365-76.

## Reproducibility

- Low statistical power undermines the purpose of scientific research; it reduces the chance of detecting a true effect;
- Consequences of such low statistical power, which include overestimates of effect size and low reproducibility of results;
- There are ethical dimensions to the problem of low power; unreliable research is inefficient and wasteful.

## Reproducibility

Weber, F., Do, J. P. H., Chung, S., Beier, K. T., Bikov, M., Doost, M. S., Dan, Y. (2018).Regulation of REM and non-REM sleep by periaqueductal GABAergic neurons. Nature communications, 9(1), 354.

- Sample size: For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions.
- Is this experiment reproducible?

## What are power considerations?

- It is a set of statistical calculations aiming to evaluate whether a clinical study is able to test the hypothesis of interest given some assumptions and resources constraints.

# Introduction

## What are power considerations?

- It is a set of statistical calculations aiming to evaluate whether a clinical study is able to test the hypothesis of interest given some assumptions and resources constraints.

## Who does it matter for?

- Funding agencies: NIH, DOD, PCORI;
- Review boards: IRB, IACUC, PRMC;
- Regulatory agencies: FDA;
- Investigator who does not want to waste effort with a study that will fail with a large chance.

# Introduction

## Example

- Investigators are interested in designing a study to evaluate whether an intervention on eating habits can benefit cancer survivors with overweight/obesity.
- They proposed to enroll patients in a 6-month intervention and follow them up to 18 months. Patients will be measured at baseline, 6, 12 and 18 months.

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.
- How do we measure whether this intervention is efficacious?

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.
- How do we measure whether this intervention is efficacious?
  - BMI;
  - Waist circumference;
  - Diet Intake;
  - PROMIS;

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.
- How do we measure whether this intervention is efficacious?
  - BMI;
  - Waist circumference;
  - Diet Intake;
  - PROMIS;
- What is the primary endpoint?

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.
- How do we measure whether this intervention is efficacious?
  - BMI;
  - Waist circumference;
  - Diet Intake;
  - PROMIS;
- What is the primary endpoint?
  - BMI

# How can they design a study to evaluate this intervention?

## Biological hypotheses

- The intervention is efficacious.
- How do we measure whether this intervention is efficacious?
  - BMI;
  - Waist circumference;
  - Diet Intake;
  - PROMIS;
- What is the primary endpoint?
  - BMI
- The intervention decreases BMI over time.

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?
  - 18 months

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?
  - 18 months
- Let's say that after 18 months, the average decrease in BMI is 1 BMI unit.

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?
  - 18 months
- Let's say that after 18 months, the average decrease in BMI is 1 BMI unit.
  - Is that decrease an expected variation over time or generated by the intervention?

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?
  - 18 months
- Let's say that after 18 months, the average decrease in BMI is 1 BMI unit.
  - Is that decrease an expected variation over time or generated by the intervention?
  - It will be hard to established a causal effect without a control weight loss program (such as Weight Watchers).

# How can they design a study to evaluate this intervention?

## Biological hypothesis

- The intervention decreases BMI over time.
- Which time point is the most important one?
  - 18 months
- Let's say that after 18 months, the average decrease in BMI is 1 BMI unit.
  - Is that decrease an expected variation over time or generated by the intervention?
  - It will be hard to established a causal effect without a control weight loss program (such as Weight Watchers).
- The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.

# Parallel 2-arm longitudinal study design

## Biological hypothesis

- The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.

## Study design

- Patients will be randomized either to receive intervention or control arms during 6 months, and they will be followed up to 18 months.

# Parallel 2-arm longitudinal study design

## Biological hypothesis

- The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.

## Study design

- Patients will be randomized either to receive intervention or control arms during 6 months, and they will be followed up to 18 months.
- How many patients are needed?

# Parallel 2-arm longitudinal study design

## Biological hypothesis

- The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.

## Study design

- Patients will be randomized either to receive intervention or control arms during 6 months, and they will be followed up to 18 months.
- How many patients are needed?
- Power considerations should be performed.

# Parallel 2-arm longitudinal study design

## Power considerations

- It is a set of statistical calculations aiming to evaluate whether a clinical study is able to test the hypothesis of interest given some assumptions and resources constraints.

# Parallel 2-arm longitudinal study design

## Power considerations

- It is a set of statistical calculations aiming to evaluate whether a clinical study is able to test the hypothesis of interest given some assumptions and resources constraints.
- Biological hypothesis: The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.

# Parallel 2-arm longitudinal study design

## Power considerations

- It is a set of statistical calculations aiming to evaluate whether a clinical study is able to test the hypothesis of interest given some assumptions and resources constraints.
- Biological hypothesis: The intervention will show a larger decrease in BMI at 18 months when compared with a control arm.
- The biological hypothesis needs to be translated to statistical hypotheses.

# Summary

# Test of hypotheses

## Statistical Hypotheses

- **Null hypothesis**: The hypothesis that we want to disprove.
- **Alternative hypothesis**: The hypothesis that we want to prove.

## Translating Biological hypothesis to Statistical Hypotheses

- $H_0$: No differences will be observed in average BMI at 18 months between intervention and control;
- $H_1$: Intervention arm will show lower average BMI at 18 months than control arm.

# Test of hypotheses

## Translating Biological hypothesis to Statistical Hypotheses

- $H_0$: No differences will be observed in BMI at 18 months between intervention and control;
- $H_1$: Intervention arm will show lower BMI at 18 months than control arm.

## Assumption

- Randomization will ensure that average BMI at baseline will be the same at baseline for both arms. Therefore, it is enough to compare the BMI at 18 months.

## Uncertainty and decision making

- After data collection, we will make a decision based on a statistical test: **reject** or **not reject** the null hypothesis.
- However, wrong decisions can be done given uncertainty.
- Statistics allow us to measure uncertainty:

| | | Truth | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| 2*Decision | Fail to Reject $H_0$ | No error | Type II |
| | Reject $H_0$ | Type I | No error |

# Test of hypotheses

## Type I error

- **Rejecting $H_0$ when $H_0$ is true**;
- In our study, type I error implies into stating that the intervention arm is more efficacious than control arm in decreasing BMI at 18 months when the intervention arm is actually equal or worse than the control arm.
- It is denoted as $\alpha$;
- For historical reasons, the maximum type I error (significance level) is often defined at 1%, 5% and 10%.

# Test of hypotheses

## Type II error

- **Failing to reject $H_0$ when $H_0$ is false**;
- In our study, type II error implies into stating that the intervention arm is equal or less efficacious than control arm in decreasing BMI at 18 months when the intervention arm is actually better than the control arm.
- It is denoted as $\beta$.

# Test of hypotheses

## Power

- It is defined as 1 - Prob(type II error) $= 1 - \beta$;
- Power is the probability of **rejecting** the null hypothesis when the null is false, i.e., finding evidence favoring the hypothesis of interest when the hypothesis of interest is true;
- Clinical studies are often required to reach at least 80% of power.

# Test of hypotheses

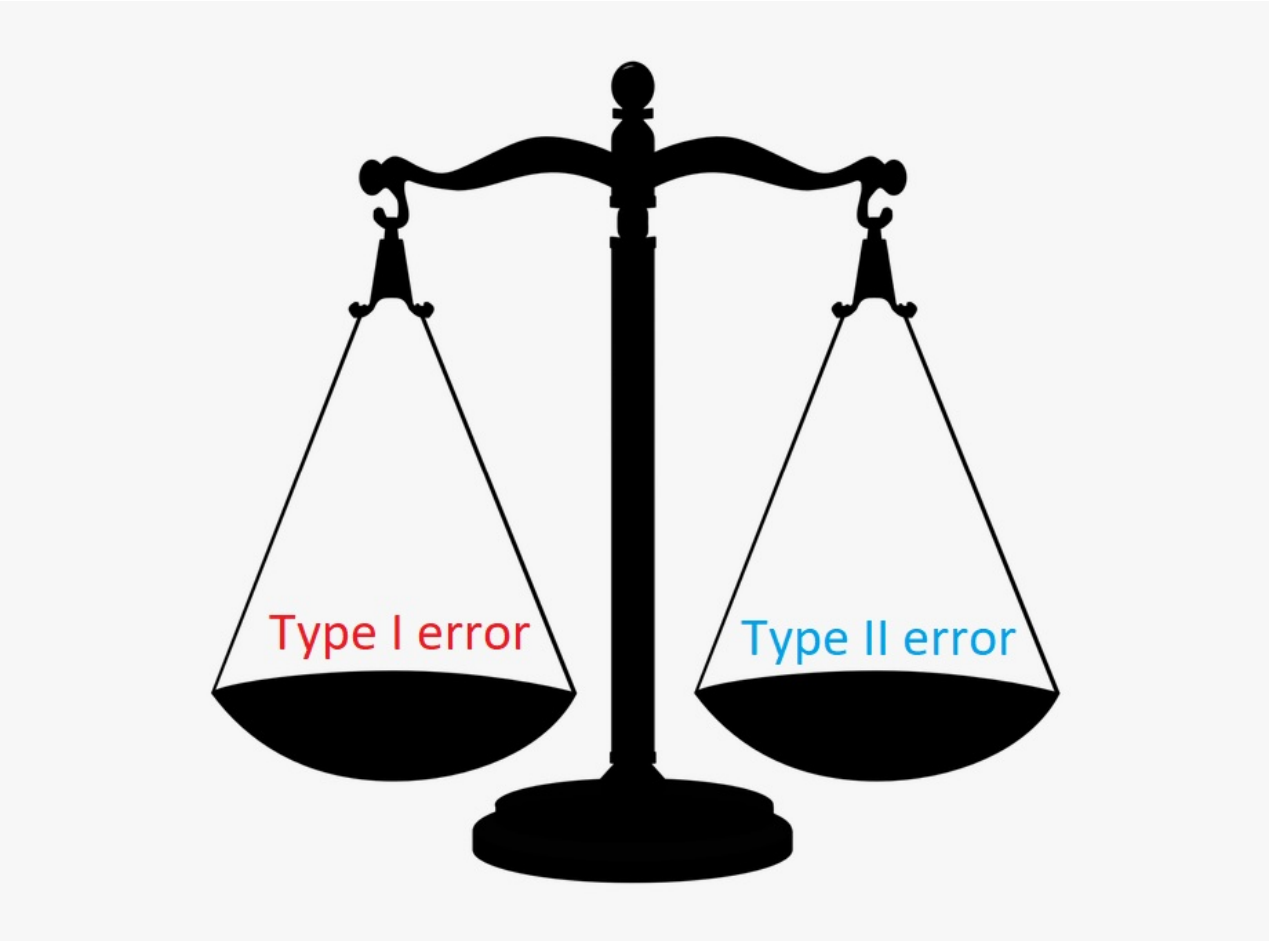



Figure: How could we balance significance level $\leq 5\%$ and Power $\geq 80\%$?

# Summary

# Which test will be used?

## Primary endpoint

- If BMI will be considered a numerical variable ($kg/m^2$), then we are interested in the average BMI by arm, and we will use a t-student test.
- If BMI will be considered as a categorical variable (Overweight or Obese: Yes/No)? Then, we are interested in the proportion of Overweight/Obese patients by arm, and we will use a proportion test.

# BMI as continuous variable

## Specific assumptions

- What is the standard deviation of BMI in the control arm?
- Will the standard deviation for the intervention arm be the same as the control arm?
- What is the expected (clinical meaningful) difference to be observed between arms?

## Standard assumptions

- Significance level?
- Power?

# BMI as continuous variable

## Specific assumptions

- Assuming a standard deviation of 2.5 from NHANES;
- Same standard deviation for both groups;
- At least 1 BMI unit.

## Standard assumptions

- Significance level of 5%;
- Power of 80%.

# Power Considerations

## Blurb

- The main hypothesis to be tested is whether there is difference in average BMI between intervention and control arms. We assume a standard deviation of 2.5 based on NHANES and 1 BMI unit as a minimum clinical difference. A sample size of 100 patients per group will reach 80% of power using a two-independent sample Student t-test at 5% significance level.

# Power Considerations

## Sample size will be based on...

- Hypothesis to be tested;
- Standard deviation (specific for t-test);
- Minimum clinical difference;
- Power;
- Significance level.

# Power Considerations - Assumptions

## Effect size

- Effect size is defined based on the statistical hypotheses that will be used to calculate sample size. For the difference between two groups,

$$\delta = \frac{mean_{intervention} - m_{control}}{sd} = \frac{1}{2.5} = 0.4$$

# Power Considerations - Assumptions

## Effect size

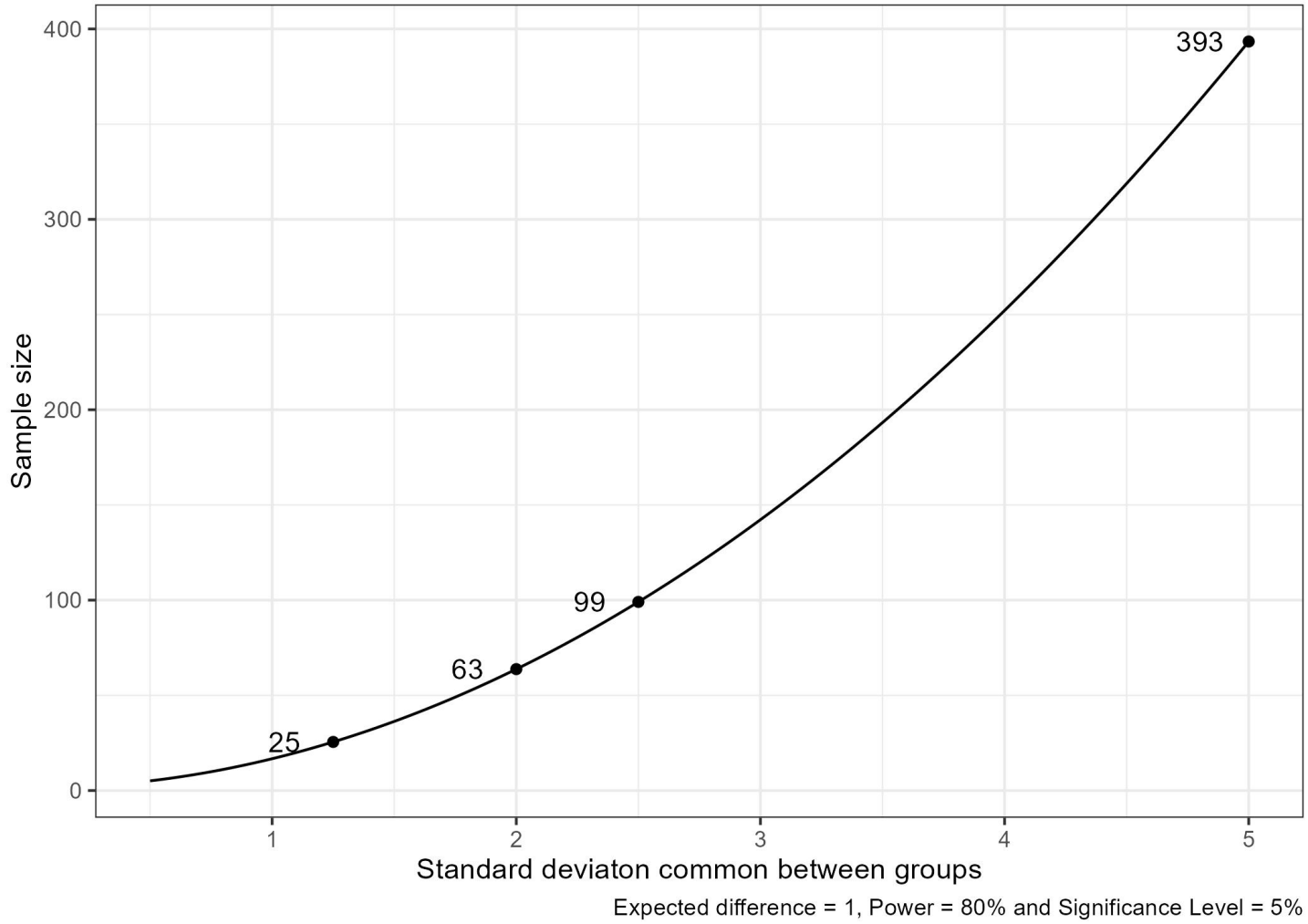
- Effect size is defined based on the statistical hypotheses that will be used to calculate sample size. For the difference between two groups,

$$\delta = \frac{mean_{intervention} - m_{control}}{sd} = \frac{1}{2.5} = 0.4$$

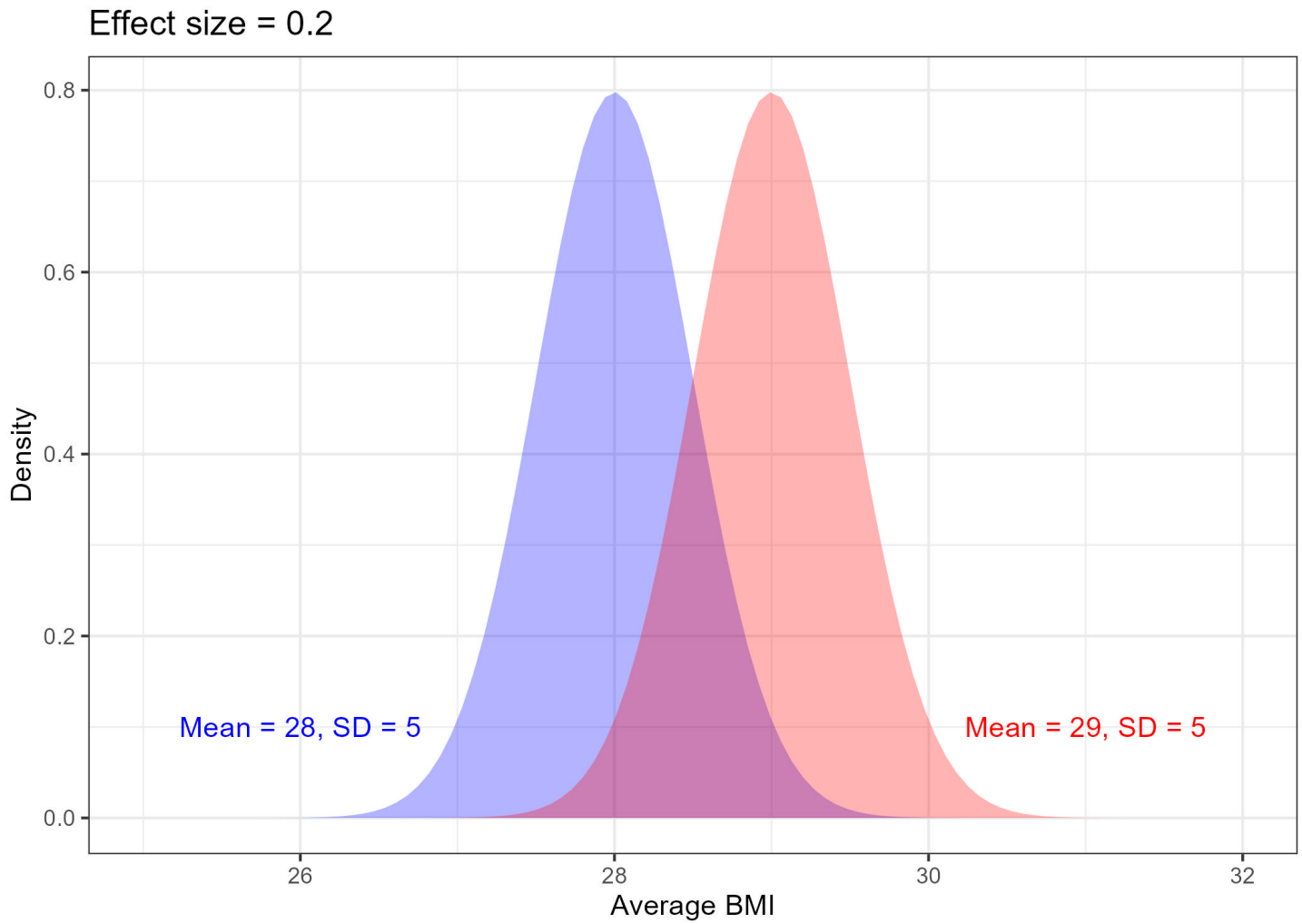
## Cohen (1988)

- Small: $0.2 \rightarrow mean_{intervention} - m_{control} = 0.5$ BMI units;
- Medium: $0.5 \rightarrow mean_{intervention} - m_{control} = 1.25$ BMI units;
- Large: $0.8 \rightarrow mean_{intervention} - m_{control} = 2$ BMI units;

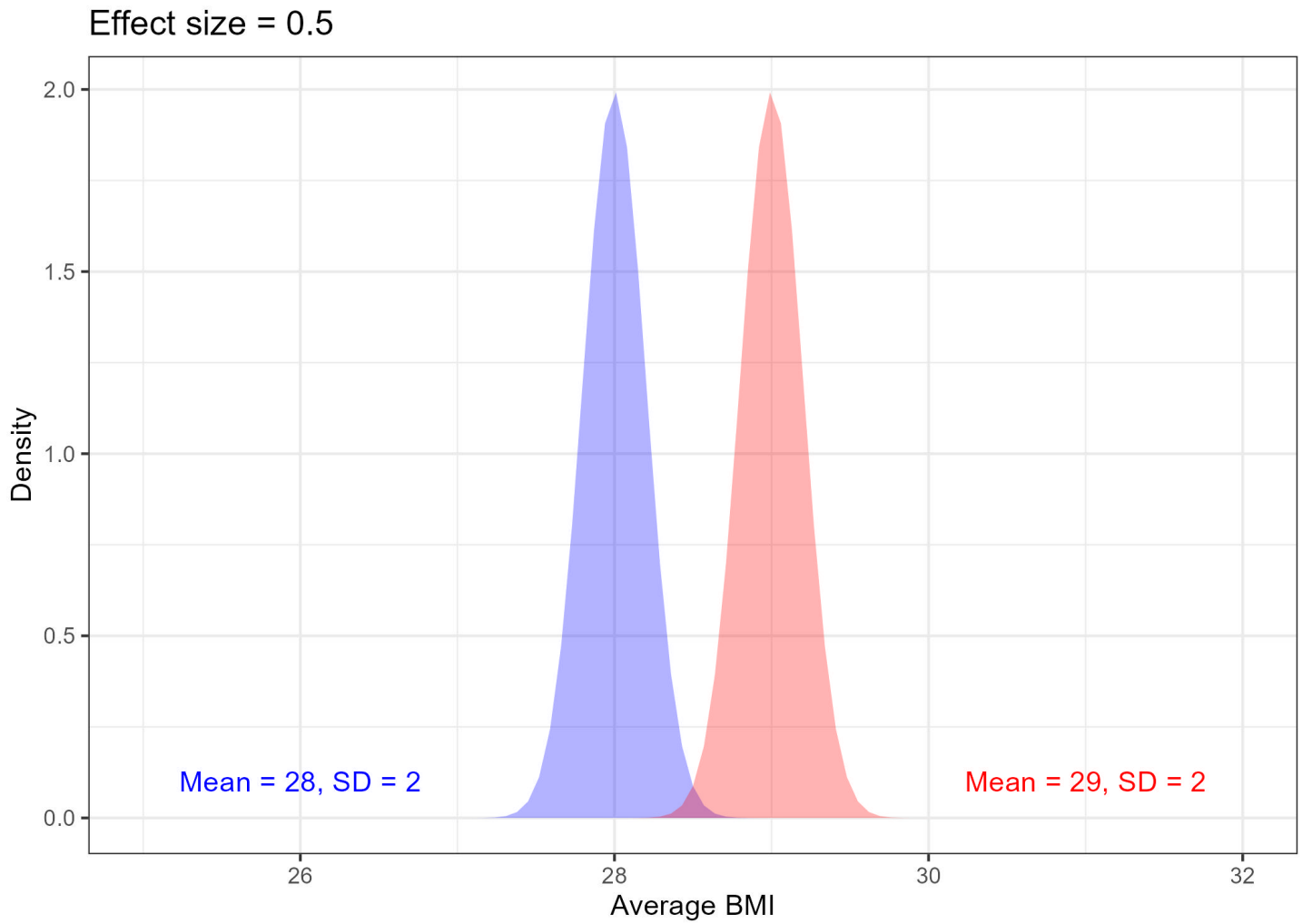
# How important is the assumption of standard deviation?



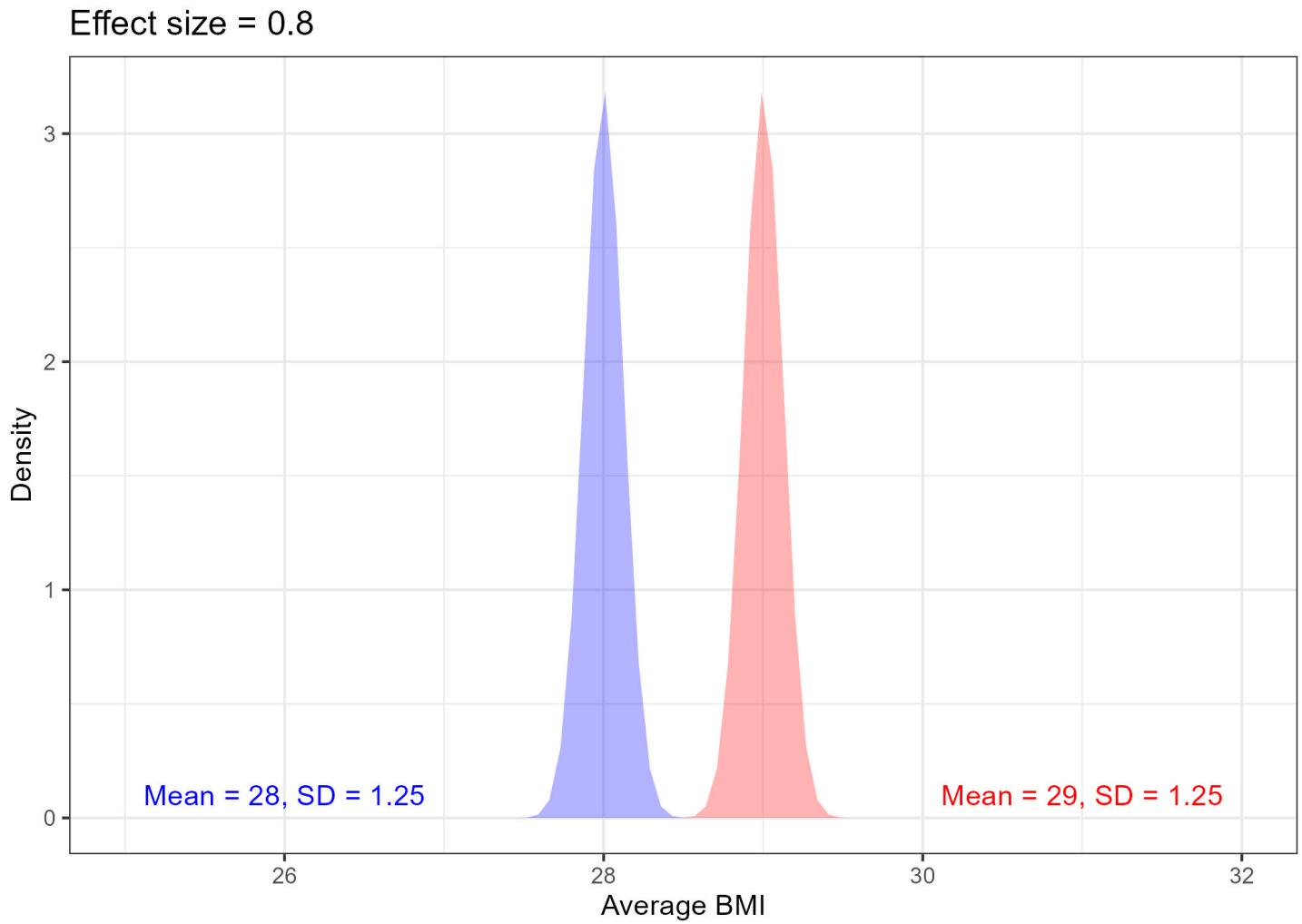
# How important is the assumption of standard deviation?



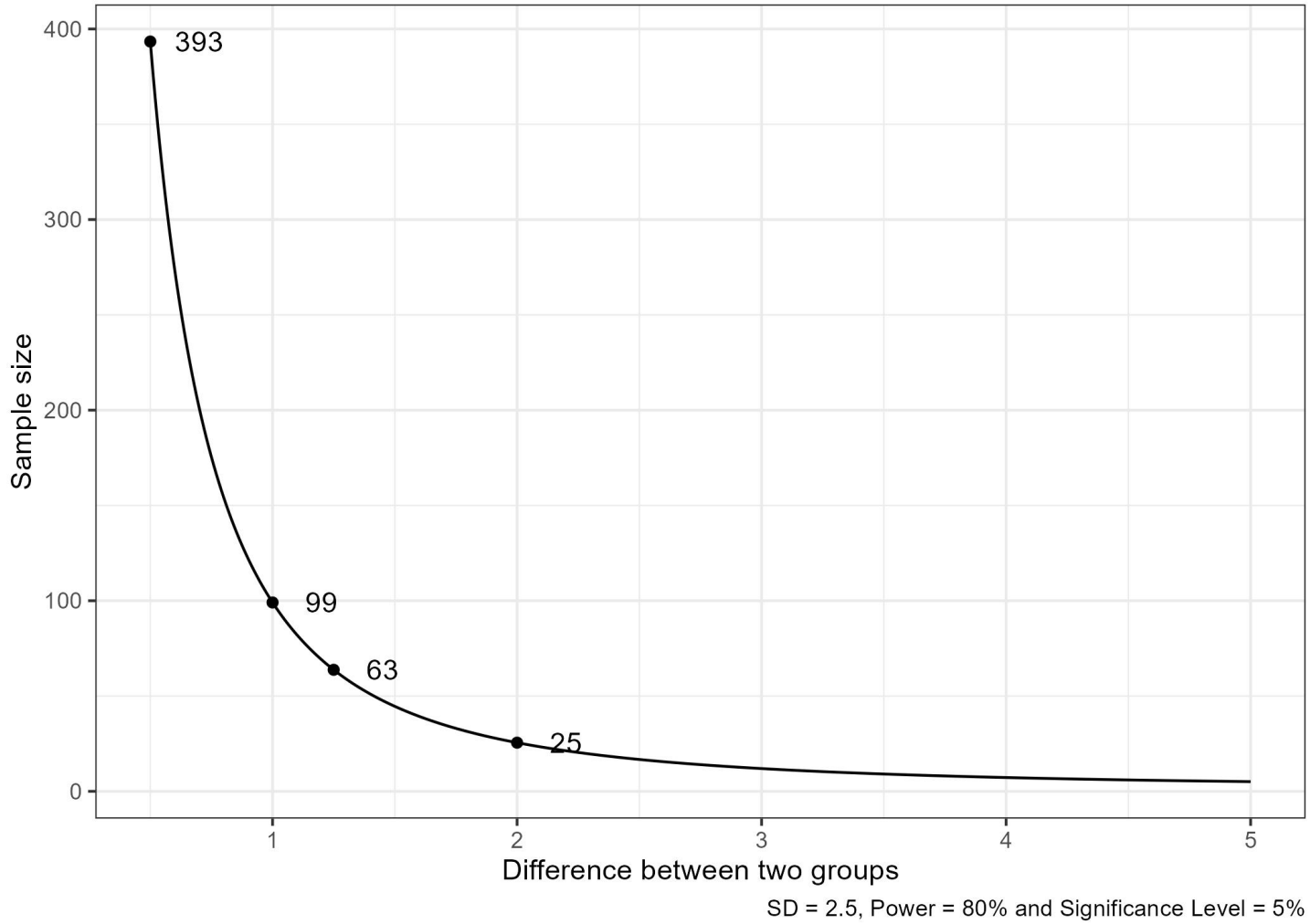
# How important is the assumption of standard deviation?



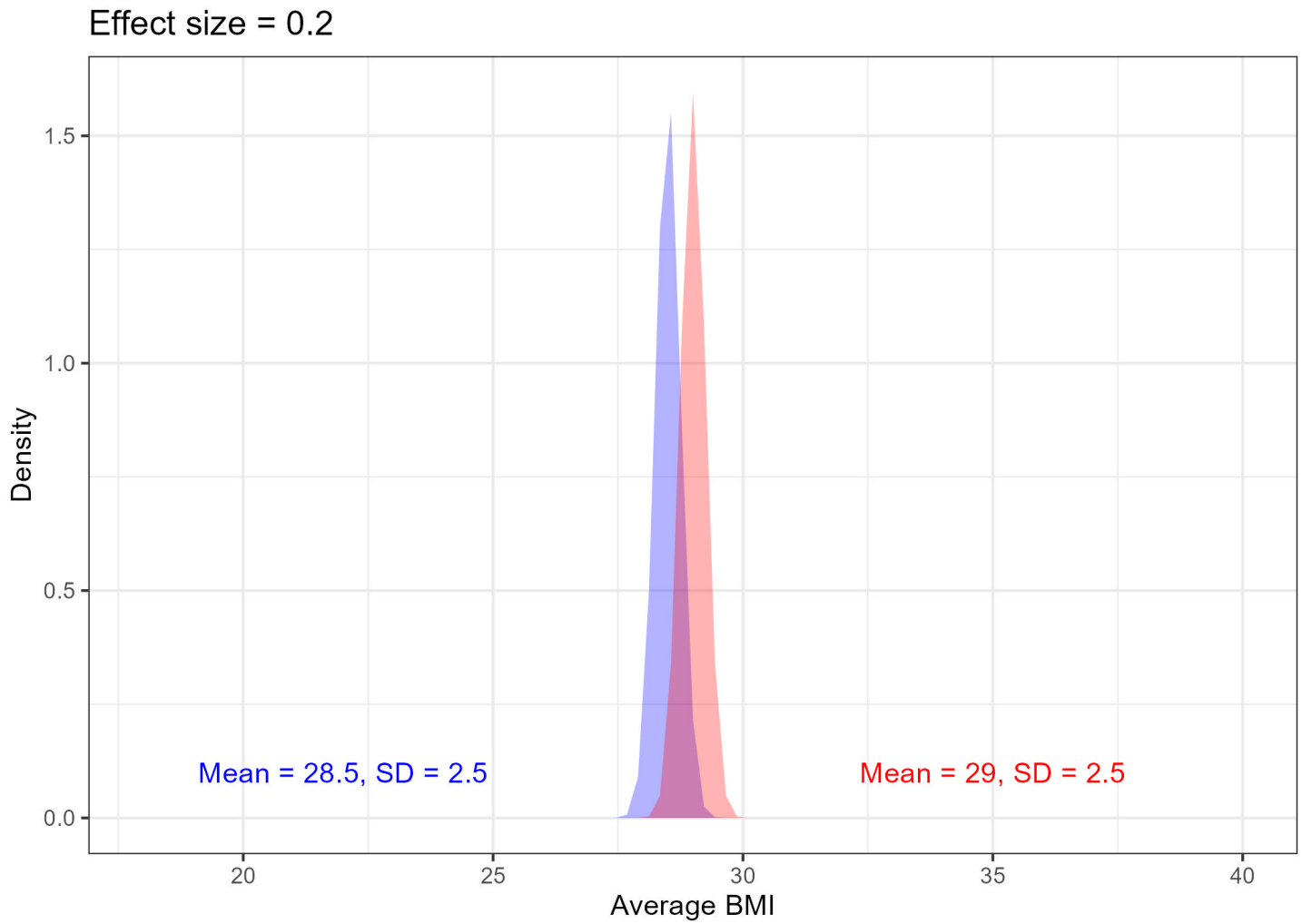
# How important is the assumption of standard deviation?



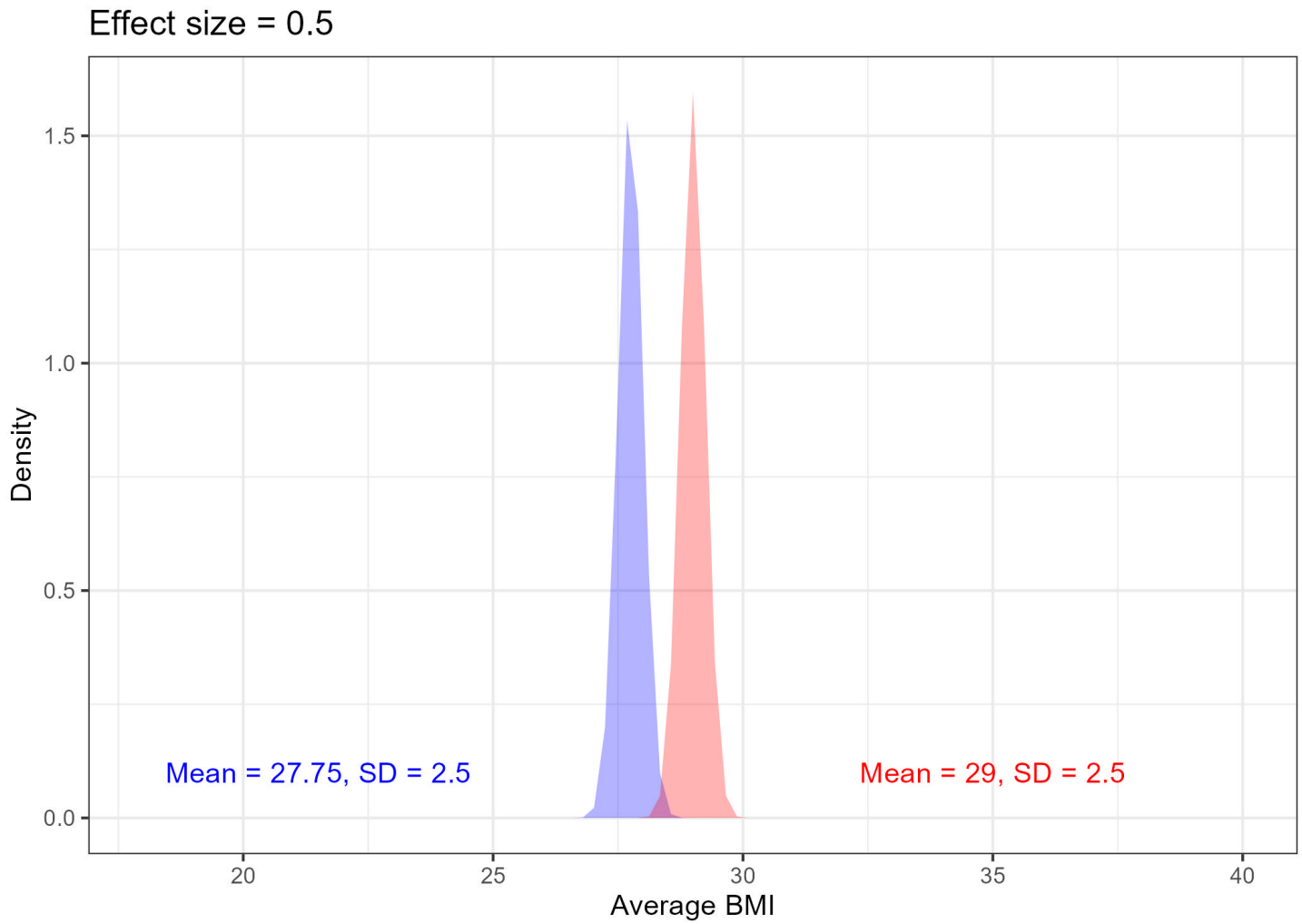
# How important is the clinical meaningful difference?



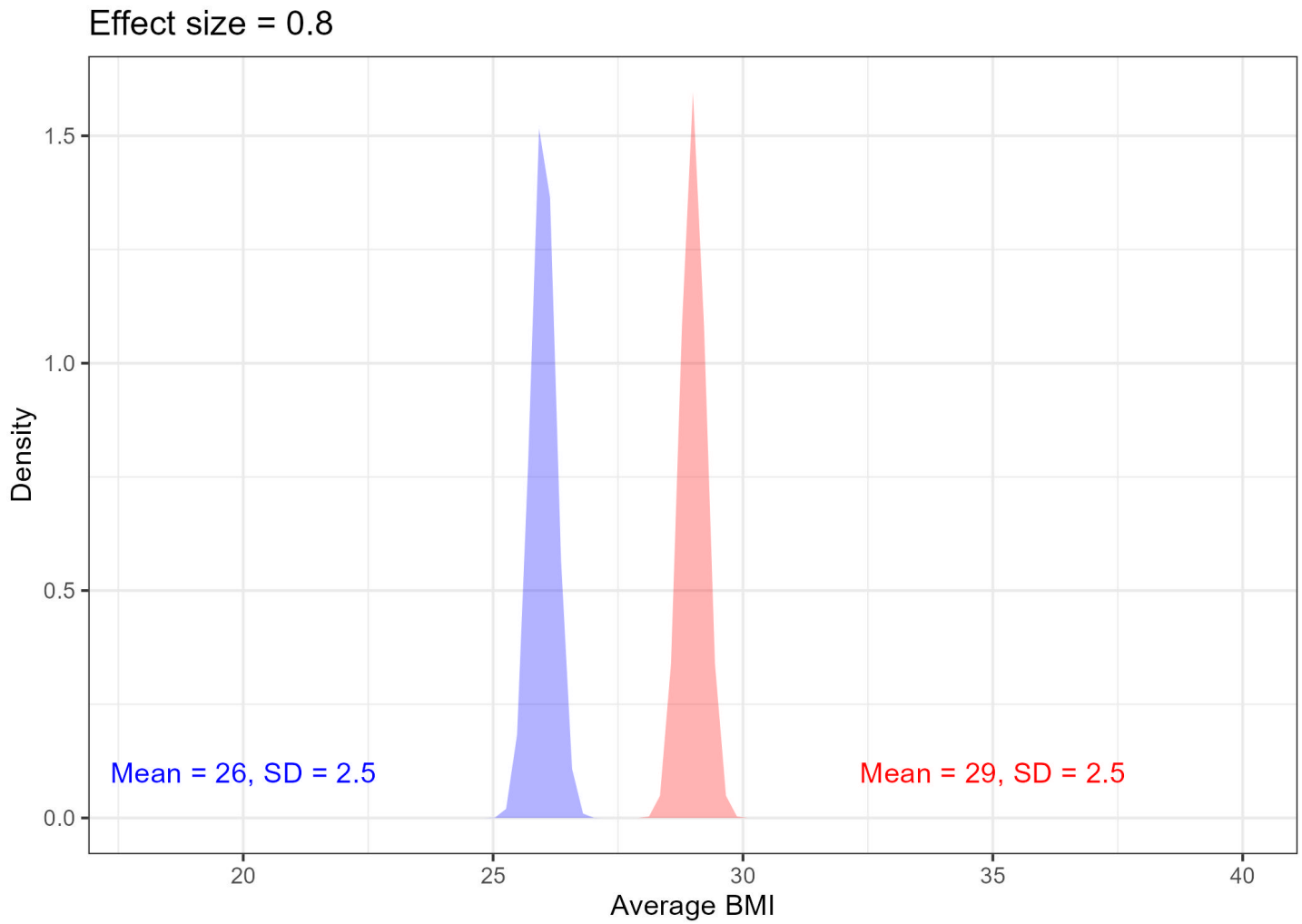
# How important is the clinical meaningful difference?



# How important is the clinical meaningful difference?



# How important is the clinical meaningful difference?



# Summary

# Power Considerations

## Sample size will be calculated based on...

- Hypothesis to be tested;
- Standard deviation (specific for t-test);
- **Minimum clinical difference;**
- Power;
- Significance level.

# Power Considerations

**Minimum detectable difference (MDF) will be calculated based on...**

- Hypothesis to be tested;
- Standard deviation (specific for t-test);
- **Sample size;**
- Power;
- Significance level.

# Power Considerations

## When should we calculate MDF instead of sample size?

- In a grant application, MDF should be calculated for secondary endpoints using the sample size defined based on the primary endpoint;

# Power Considerations

## When should we calculate MDF instead of sample size?

- In a grant application, MDF should be calculated for secondary endpoints using the sample size defined based on the primary endpoint;
- There is no clear minimum clinical difference. It still needs to be justified as achievable based on previous literature/preliminary data;

# Power Considerations

## When should we calculate MDF instead of sample size?

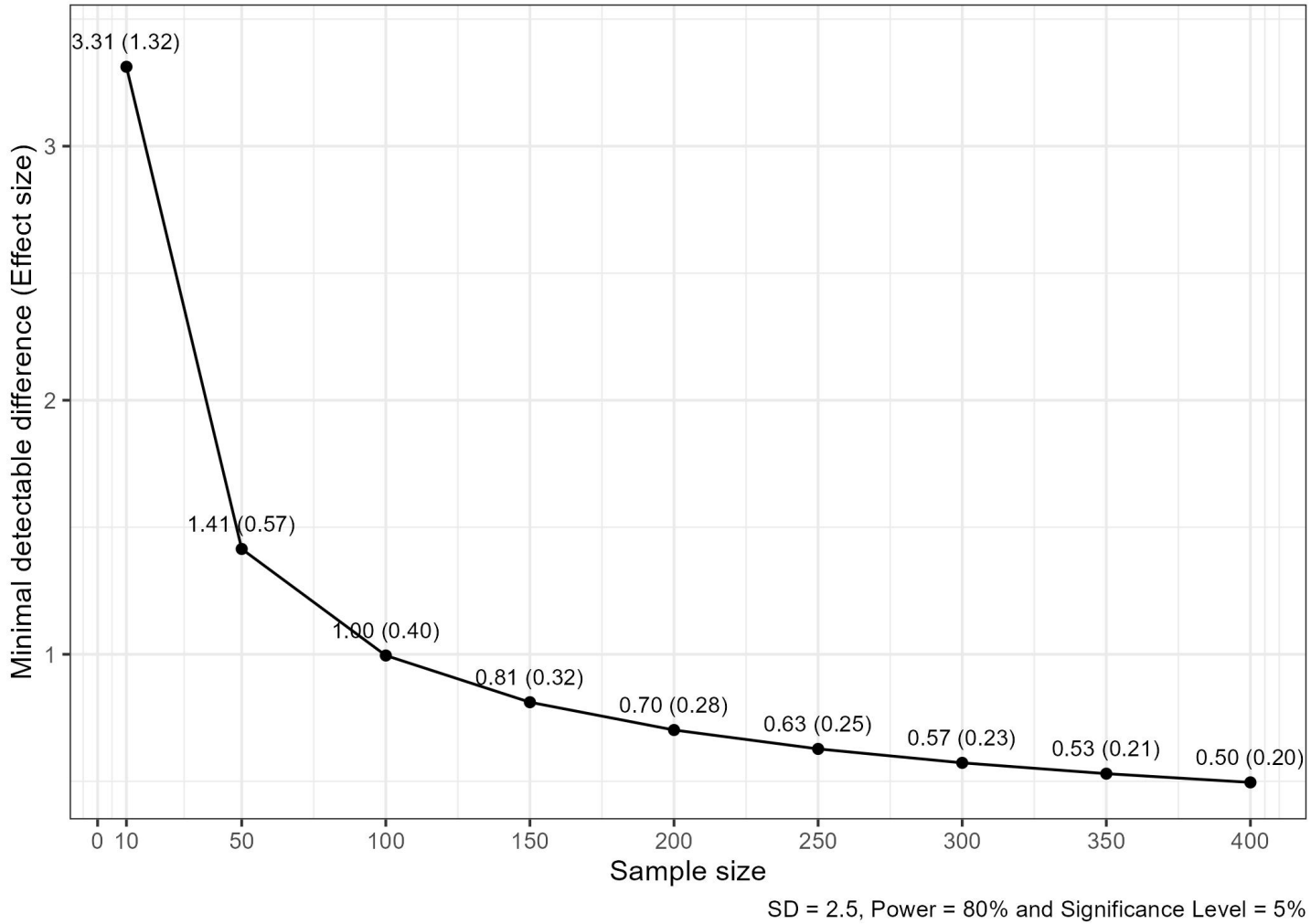
- In a grant application, MDF should be calculated for secondary endpoints using the sample size defined based on the primary endpoint;
- There is no clear minimum clinical difference. It still needs to be justified as achievable based on previous literature/preliminary data;
- Sample size is limited by budget or accrual. It still needs to be justified as clinical meaningful based on previous literature.

# Power Considerations

## Weight Loss Intervention

- Sample size is limited on 50 patients per group.

# Power Considerations



# Power Considerations

**What should an investigator bring to a meet with a statistician?**

- Well-defined biological hypothesis;
- Primary endpoint;
- Preliminary data or literature about primary endpoint;
- Maximum sample size due budget or accrual.

# Power Considerations

## Do not forget

- Sample size for simple designs have closed formulas that allow an easy calculation only if the investigator have **all** items above readily available;
- Statisticians are not able to evaluate the plausibility of assumptions based on preliminary data or literature;
- Complex problems require sample size justification based on computational simulations, which require large amounts of time.

# Study Design: Power Considerations

Questions?

## Biostatistics Core

- Request form can be found here.