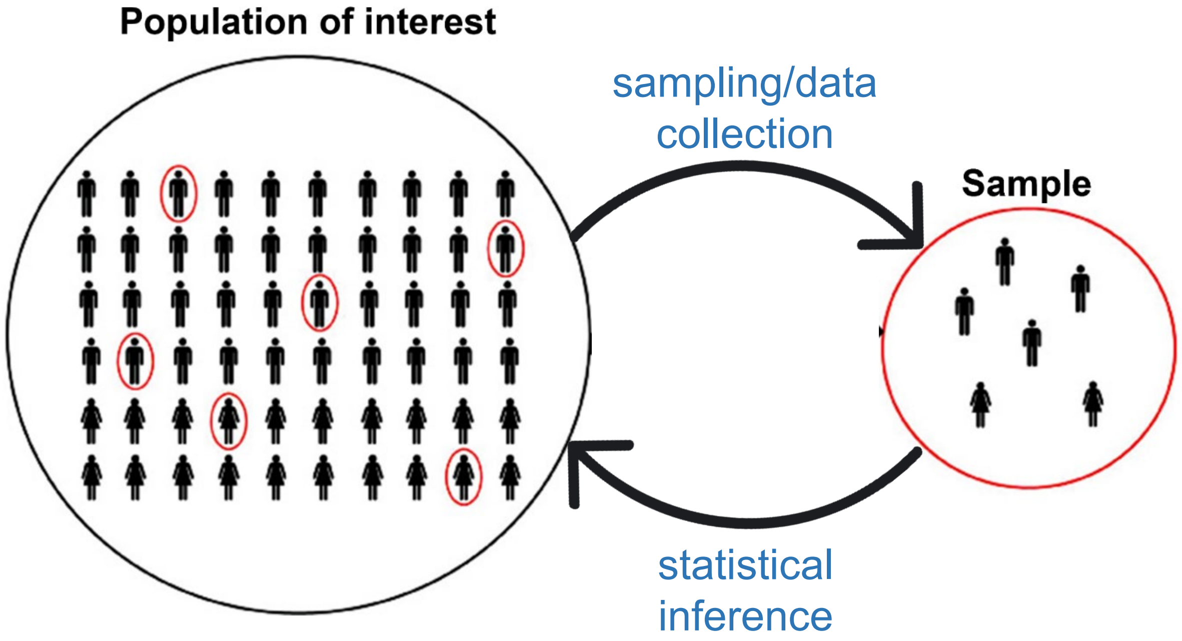
Continuous data – 1 group power and sample size

Notation

- μ = population mean
- σ = population standard deviation

- $\bar{x}$ = sample mean
- s = sample standard deviation



Types of Errors in Hypothesis Tests

		Hypothesis test decision	
		Reject H_0	Fail to reject H_0
Reality	H_0 true	**Type 1 error**	**Correct**
	H_0 false	**Correct**	**Type 2 error**

$$\alpha = P(\text{type 1 error})$$
$$\beta = P(\text{type 2 error})$$

Types of Errors

- $\alpha = P(\text{type 1 error})$
  - Probability of rejecting H_0 when H_0 is really true
  - False positive
  - Pre-specified significance level (usually $\alpha = 0.05$)
- $\beta = P(\text{type 2 error})$
  - Probability of failing to reject H_0 when H_0 is really false
  - False negative
  - $1 - \beta = \text{power}$ (usually power = 0.8 or 0.9 so $\beta = 0.2$ or 0.1)

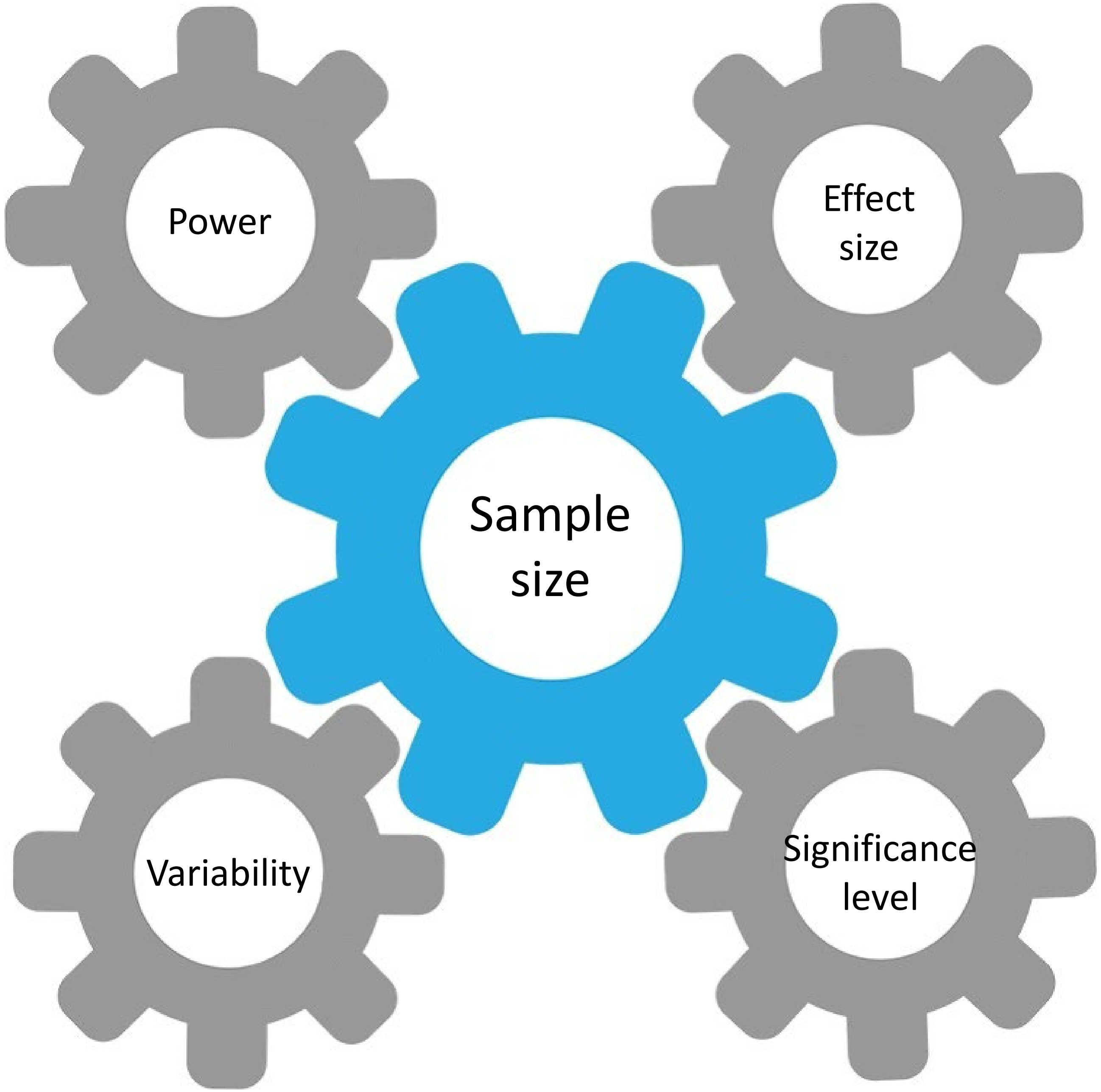
Want to minimize α and β, but it's a balance (like balancing sensitivity and specificity in a diagnostic test)

Power

- Probability of rejecting H_0 when H_0 is really false
- Think of it as the ability of your test to identify the thing that you're trying to show
- During study design: often interested in determining the number of subjects necessary to achieve a certain power
- Usually pre-specify that we want 80% or 90% power

Calculating Necessary Sample Size

- Each type of hypothesis test has a formula for calculating the necessary sample size to achieve a certain power
    - Want higher power → need more subjects
- Don’t worry about the formula – focus on what goes into it

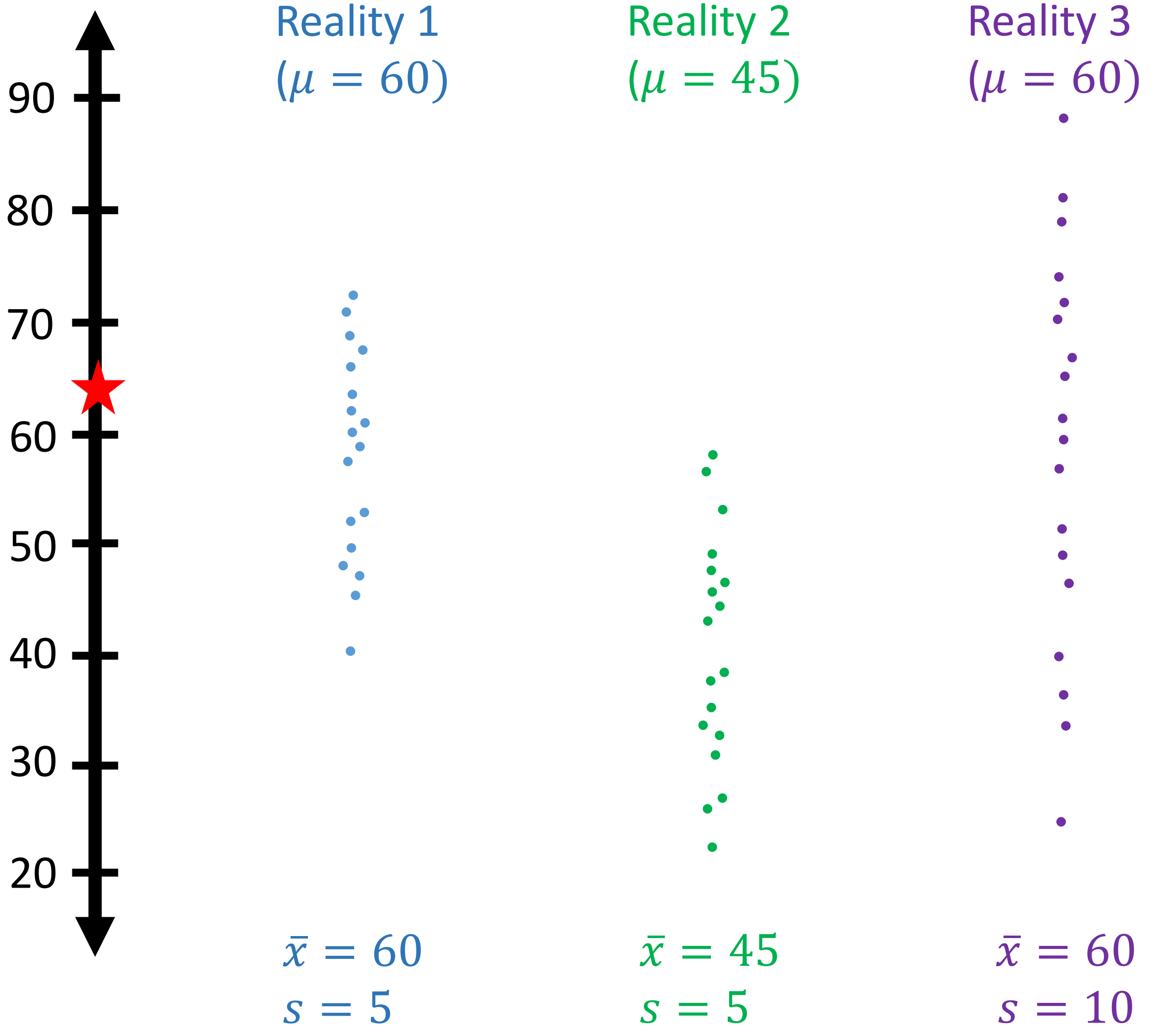

Power
Effect size
Sample size
Variability
Significance level

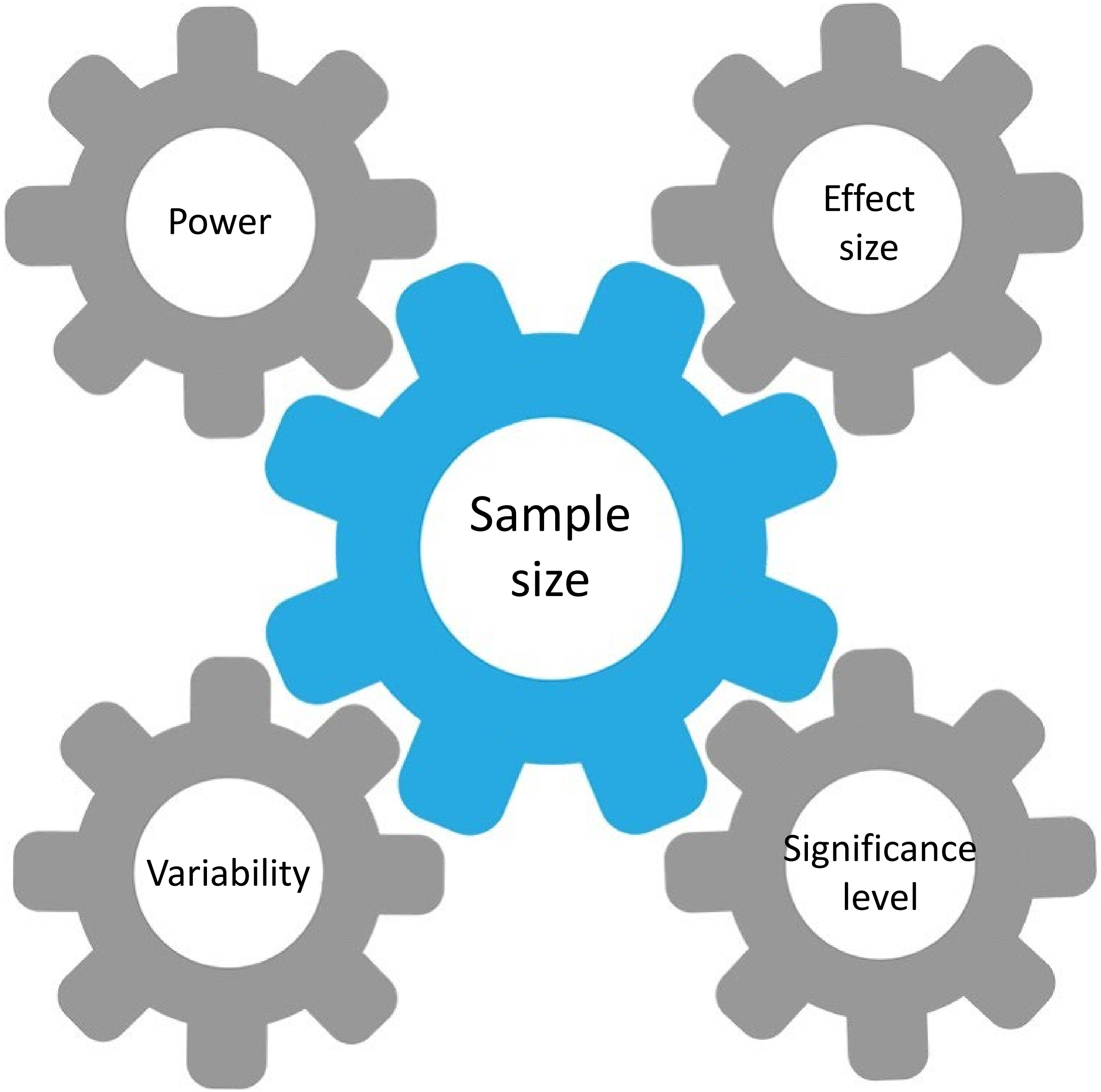
Effect Size and Variability

$H_0: \mu = 64 \qquad H_A: \mu \neq 64$

- Effect size = distance between true mean and hypothesized mean
  - Smaller effect size → need more subjects to detect effect

- Variability of observations
  - Larger variability → less sure of results → need more subjects to detect effect





Sample size

Power

Effect size

Variability

Significance level

Significance Level (α)

- Lowering the threshold for significant results means that it’s harder to pass that threshold
  - Overall, fewer things will be identified as significant
  - This is good when there really isn’t an effect (H_0 is really true)
    - Reduces false positives
  - This is bad when there really is an effect (H_0 is really false)
    - Lowers power
- To maintain a certain level of power, if we make the cutoff for significance more stringent (lower α), we need to recruit more subjects

		Hypothesis test decision	
		Reject H_0	Fail to reject H_0
Reality	H_0 true	**Type 1 error**	**Correct**
	H_0 false	**Correct**	**Type 2 error**

Factors Affecting Necessary Sample Size

- Power
  - Want higher power → need more information → larger sample size necessary
- Effect size (distance between true mean and hypothesized mean)
  - Smaller difference → harder to detect the difference → larger sample size necessary
- Variability of observations
  - Larger variability → less sure of information in sample → larger sample size necessary
- Significance level
  - Smaller α → threshold for rejection is more stringent → reject less often → need more subjects to maintain given power

Sample Size Calculation

- Power
- Effect size (distance between true mean and hypothesized mean)
- Variability of observations
- Significance level

These four numbers go into your sample size calculator/statistical software to calculate the necessary sample size.

“In order to have [___%] power to detect a mean difference of [___], assuming a standard deviation of [___] and using a significance level of [___], we would need to recruit [___] subjects.”

Study Design: Where to Get the Numbers

Talk to a biostatistician for help with these!

- Significance level and power
  - Pre-specified (you get to choose what you want to use)
- Variability
  - Need population standard deviation
  - If it’s not known, get estimate from literature (can use a previous study measuring something similar) or make best guess
- Effect size
  - We have a hypothesized mean (the number we’re trying to test in the hypotheses), but we don’t know the sample mean before we collect data
  - Get estimate from literature (can use a previous study measuring something similar) or calculate the necessary sample size for a range of possible effect sizes

Example: Exercise Intervention

- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Researchers think that their exercise intervention will alter calorie consumption by 100 calories/day, on average. How many subjects do they need to recruit to achieve a power of 80% (using a significance level of $\alpha = 0.05$)?

Power = 0.8

Effect size = 100

Standard deviation = 250

Significance level = 0.05

Example: Exercise Intervention

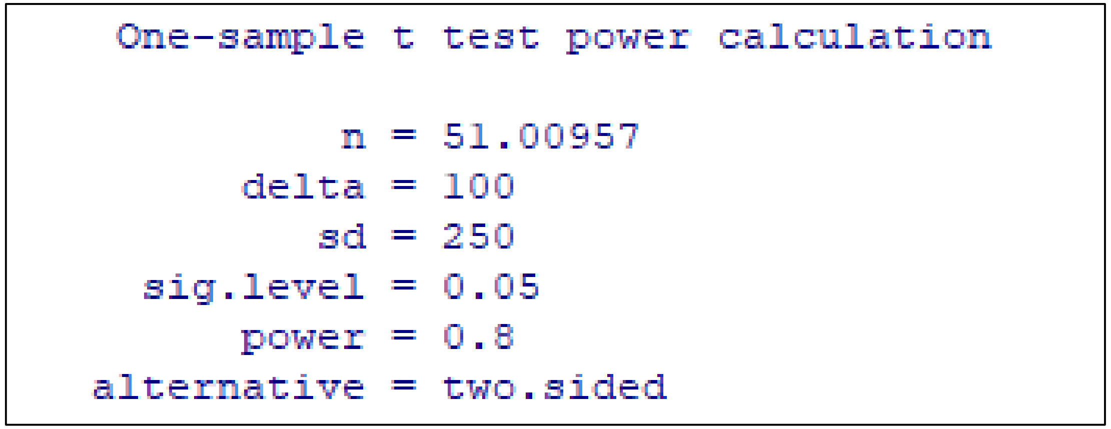
- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Researchers think that their exercise intervention will alter calorie consumption by 100 calories/day, on average. How many subjects do they need to recruit to achieve a power of 80% (using a significance level of $\alpha = 0.05$)?

Power =

Effect size =

Standard deviation =

Significance level =



```
     One-sample t test power calculation

              n = 51.00957
          delta = 100
             sd = 250
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Example: Exercise Intervention

- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Researchers think that their exercise intervention will alter calorie consumption by 100 calories/day, on average. How many subjects do they need to recruit to achieve a power of **90%** (using a significance level of $\alpha = 0.05$)?

Power =

Effect size =

Standard deviation =

Significance level =

Example: Exercise Intervention

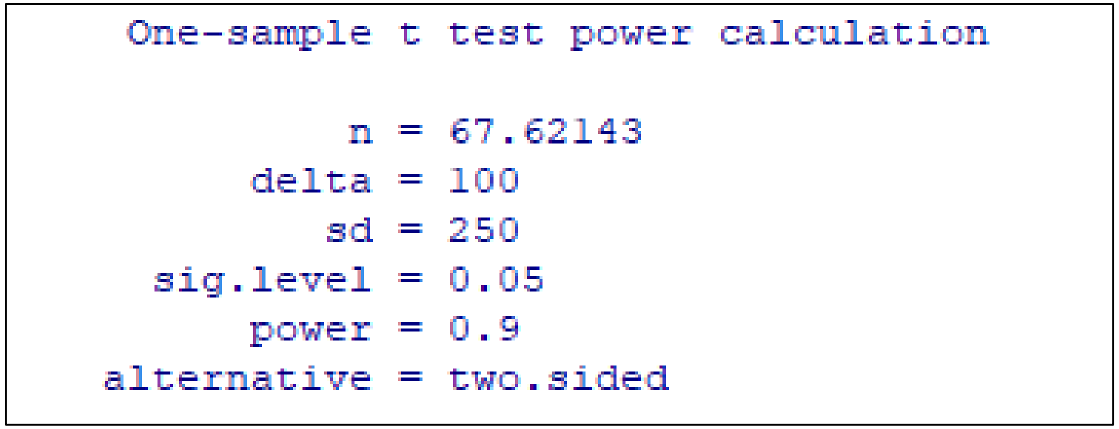
- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Researchers think that their exercise intervention will alter calorie consumption by 100 calories/day, on average. How many subjects do they need to recruit to achieve a power of **90%** (using a significance level of $\alpha = 0.05$)?

Power =

Effect size =

Standard deviation =

Significance level =



```
     One-sample t test power calculation

              n = 67.62143
          delta = 100
             sd = 250
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

Example: Exercise Intervention

- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Now suppose that the exercise intervention only alters calorie consumption by **50** calories/day, on average. How many subjects do they need to recruit to achieve a power of 90% (using a significance level of $\alpha = 0.05$)?

Power =

Effect size =

Standard deviation =

Significance level =

Example: Exercise Intervention

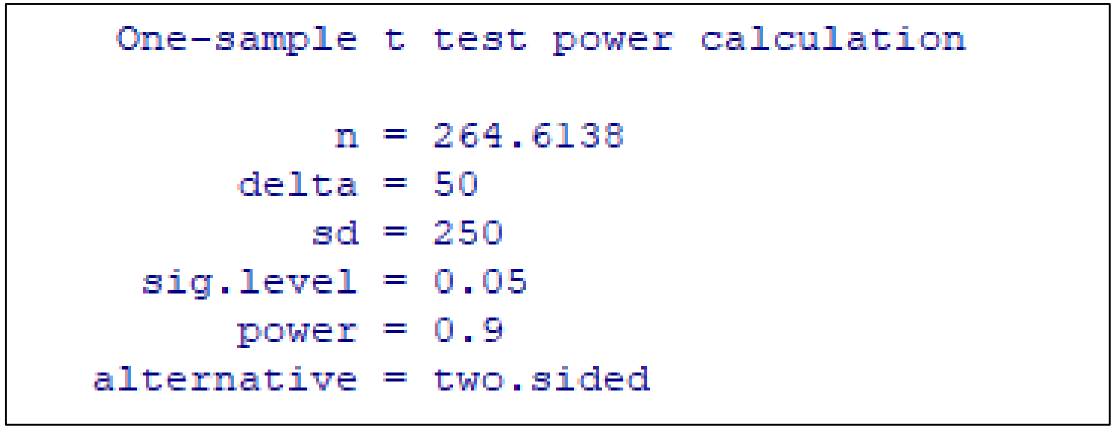
- Suppose that in the population, people consume an average of 2,000 calories/day (SD=250 calories/day). Now suppose that the exercise intervention only alters calorie consumption by **50** calories/day, on average. How many subjects do they need to recruit to achieve a power of 90% (using a significance level of $\alpha = 0.05$)?

Power =

Effect size =

Standard deviation =

Significance level =

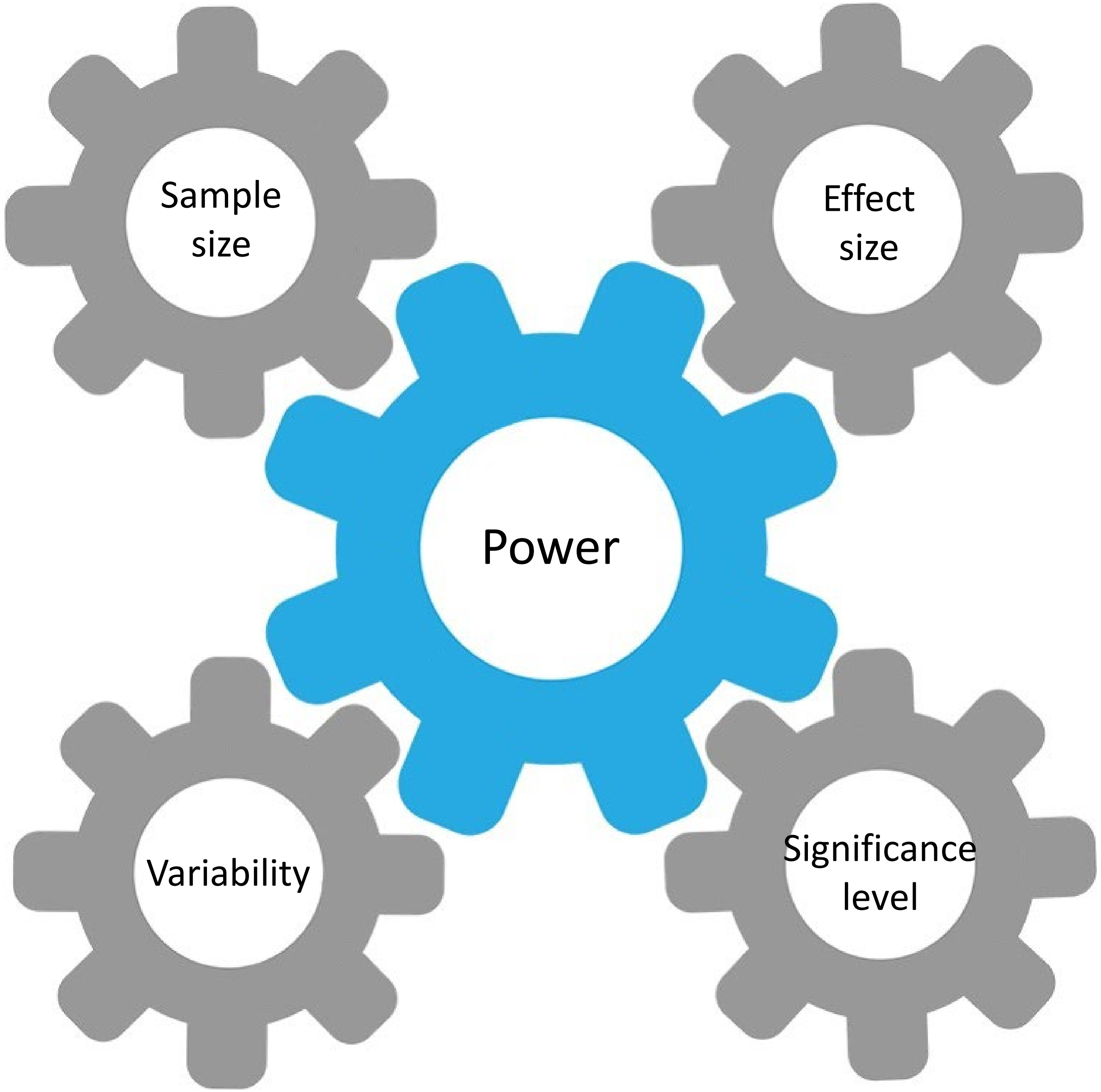


```
     One-sample t test power calculation

              n = 264.6138
          delta = 50
             sd = 250
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

Calculating Power

- You can also calculate the power of a statistical test for a pre-determined sample size
- If your power isn't high enough, you may not be able to reject H_0 **even when there really is a difference**
- Each type of hypothesis test has a formula for calculating the power that you would have given a certain sample size
- Don't worry about the formula – focus on what goes into it


Sample size
Effect size
Power
Variability
Significance level

Factors Affecting Power

- Sample size
  - Larger sample size → have more information → higher power
- Effect size (distance between true mean and hypothesized mean)
  - Bigger difference → easier to detect the difference → higher power
- Variability of observations
  - Smaller variability → more sure of information in sample → higher power
- Significance level
  - Larger α → threshold for rejection isn't as stringent → reject more often → higher power (and also more false positives)

Example: Veterans

- Suppose you want to perform a hypothesis test to assess whether the average resting heart rate of veterans is lower than that in the general population. In the general population, the average resting heart rate is 70 beats per minute (SD=7 bpm). We are going to collect information on 25 veterans to test this claim (using α=0.05). If the average resting heart rate in veterans is actually 68 beats per minute, what power would we have to detect this difference?

Sample size =

Effect size =

Standard deviation =

Significance level =

Example: Veterans

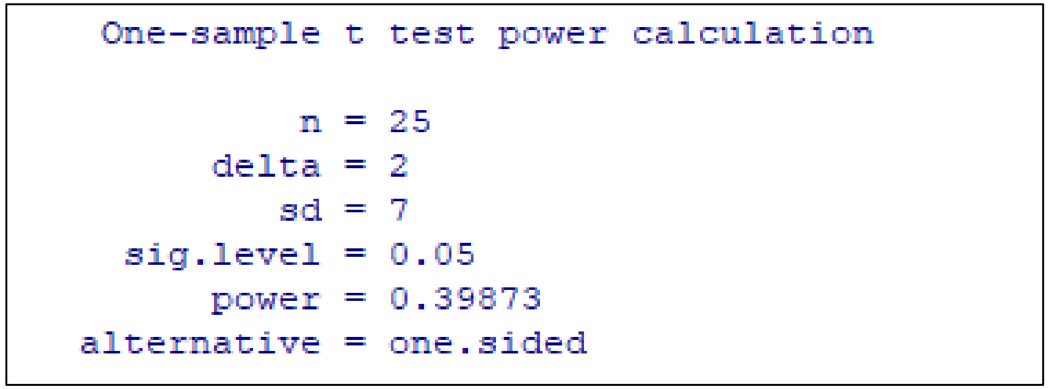
- Suppose you want to perform a hypothesis test to assess whether the average resting heart rate of veterans is lower than that in the general population. In the general population, the average resting heart rate is 70 beats per minute (SD=7 bpm). We are going to collect information on 25 veterans to test this claim (using α=0.05). If the average resting heart rate in veterans is actually 68 beats per minute, what power would we have to detect this difference?

Sample size =

Effect size =

Standard deviation =

Significance level =



```
     One-sample t test power calculation

              n = 25
          delta = 2
             sd = 7
      sig.level = 0.05
          power = 0.39873
    alternative = one.sided
```

Underpowered

- We say a study is **underpowered** when there aren’t enough subjects to detect an existing effect
- The probability of making a type 2 error is too high
  - Type 2 error: an effect/difference exists, but your test is not significant (you fail to reject H_0)
- May have severe consequences
  - Example: pharma company stops developing a drug that really does work
- “Fail to reject H_0” does NOT mean that there isn’t an effect
  - We just don’t have enough evidence to say that there is an effect
  - Maybe there really isn’t an effect – or maybe there is and our study is underpowered to detect the effect

Important Points

- Types of errors
- Concept of power
  - Definition
  - Why is it important?
  - Consequences of an underpowered study
- Relationship between power, sample size, effect size, variability, and significance level
- Power/sample size calculations
  - Where to get the information to input