BREAK
THROUGH
TECH

Unwrapping Customer Delight

Milestone #2 Meeting: The Design Phase

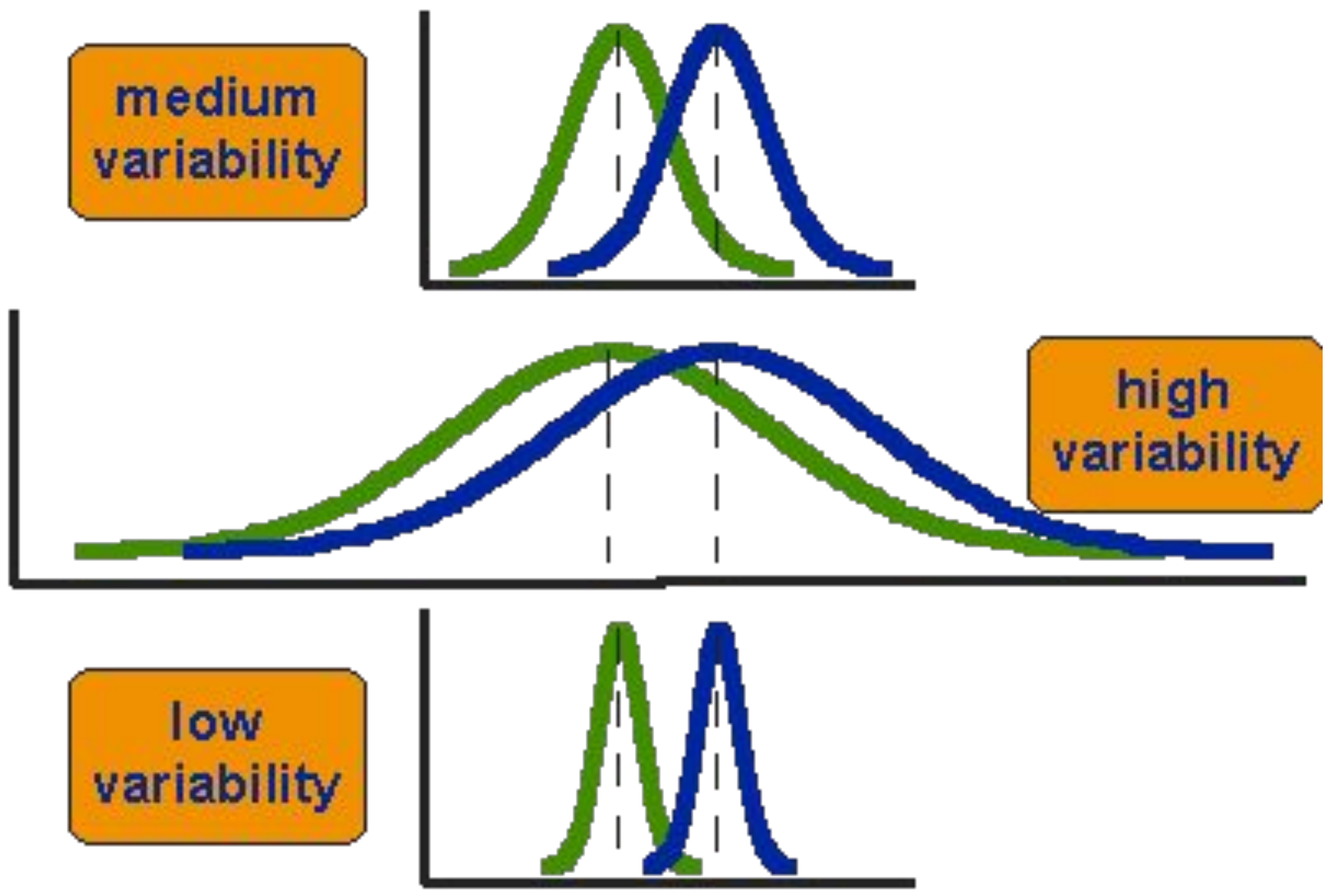
The Estée Lauder Companies

September 19, 2025

Recap: Why Variance Matters

From our last meeting, we learned that:

- Statistical power is our ability to detect a real effect.
- High variance (noise) in our outcome metric makes it harder to see the treatment effect.
- To increase power, we must reduce variance.



Recap: Why Variance Matters

From our last meeting, we learned that:

- Statistical power is our ability to detect a real effect.
- High variance (noise) in our outcome metric makes it harder to see the treatment effect.
- To increase power, we must reduce variance.

Noise/Variance

$$Coefficient\ of\ Deternination \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$Sum\ of\ Squares\ Total \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$Sum\ of\ Squares\ Regression \rightarrow \quad SSR = \sum (y' - \bar{y}')^2$$

$$Sum\ of\ Squares\ Error \rightarrow \quad SSE = \sum (y - y')^2$$

outcome metric (e.g revenue)

model output

What is Power Analysis?

- This is the main action of our **experiment design phase.** It is a planning tool used before an experiment starts.
- Power analysis helps you determine the sample size needed to detect an effect of a certain size with a given degree of confidence.
  - i.e. "How many users would I need in my experiment?"

Why is it crucial?

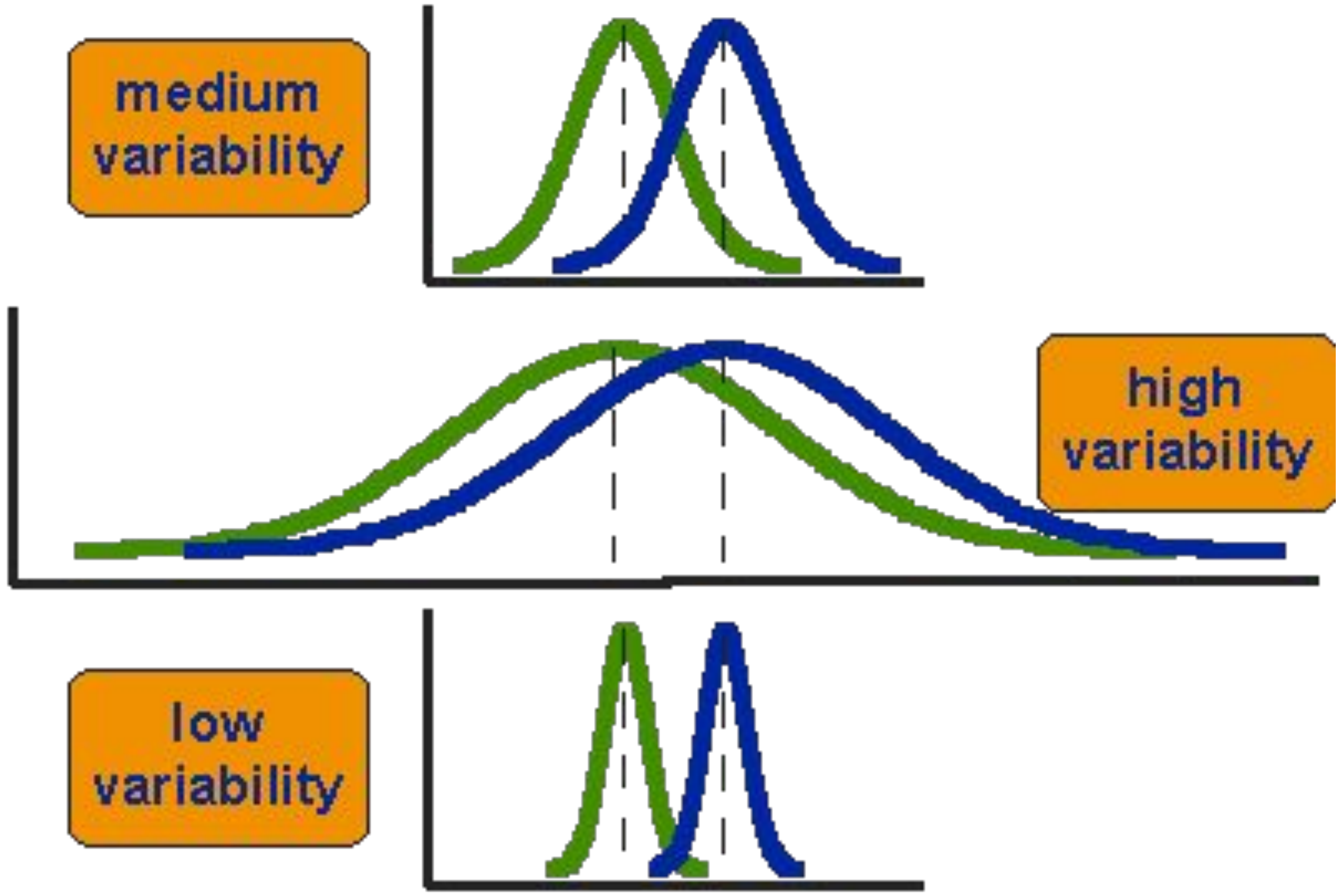
- **Too few users:** You might miss a real effect (a "false negative"). Your experiment is inconclusive.
- **Too many users:** You waste resources, time, and potentially expose too many users to a negative experience.

What is Power Analysis?

- **Statistical Power**: The probability of detecting an effect when one actually exists.
- Hypothesis test of a regression coefficient:
  - H_0: $\beta_T = 0$ (null)
  - H_a: $\beta_T \neq 0$
- Industry standard: 80-90% power

Decision made	H_0 **true**	H_0 **false**
Reject H_0	Type I error Probability $= \alpha$	Correct decision Probability $= 1 - \beta$
Do not reject H_0	Correct decision Probability $= 1 - \alpha$	Type II error Probability $= \beta$

Power



The 4 Levers of Experimental Design

These four components are interconnected. If you know any three, you can calculate the fourth.

1. **Minimum Detectable Effect size (MDE)**: The minimum desired size of change to detect. In other words, the threshold above which the change is meaningful.
2. **Significance Level (α):** Your tolerance for a "false positive". Typical value is 5%.
3. **Power (1-β)**: Your desired probability of finding a real effect. Typical value is 80%.
4. **Sample Size (N):** The number of users in your experiment (what we're solving for!).

Trade-offs:

- For a **smaller effect size**, you need a **larger sample size**
- For **more power**, you need a **larger sample size**

The 5 Levers of Experimental Design

These four components are interconnected. If you know any three, you can calculate the fourth.

1. **Minimum Detectable Effect size (MDE)**: The minimum desired size of change to detect. In other words, the threshold above which the change is meaningful.
2. **Significance Level (α):** Your tolerance for a "false positive". Typical value is 5%.
3. **Power (1-β)**: Your desired probability of finding a real effect. Typical value is 80%.
4. **Sample Size (N):** The number of users in your experiment (what we're solving for!).
5. **Variance**: The unexplained variability in your response variable.

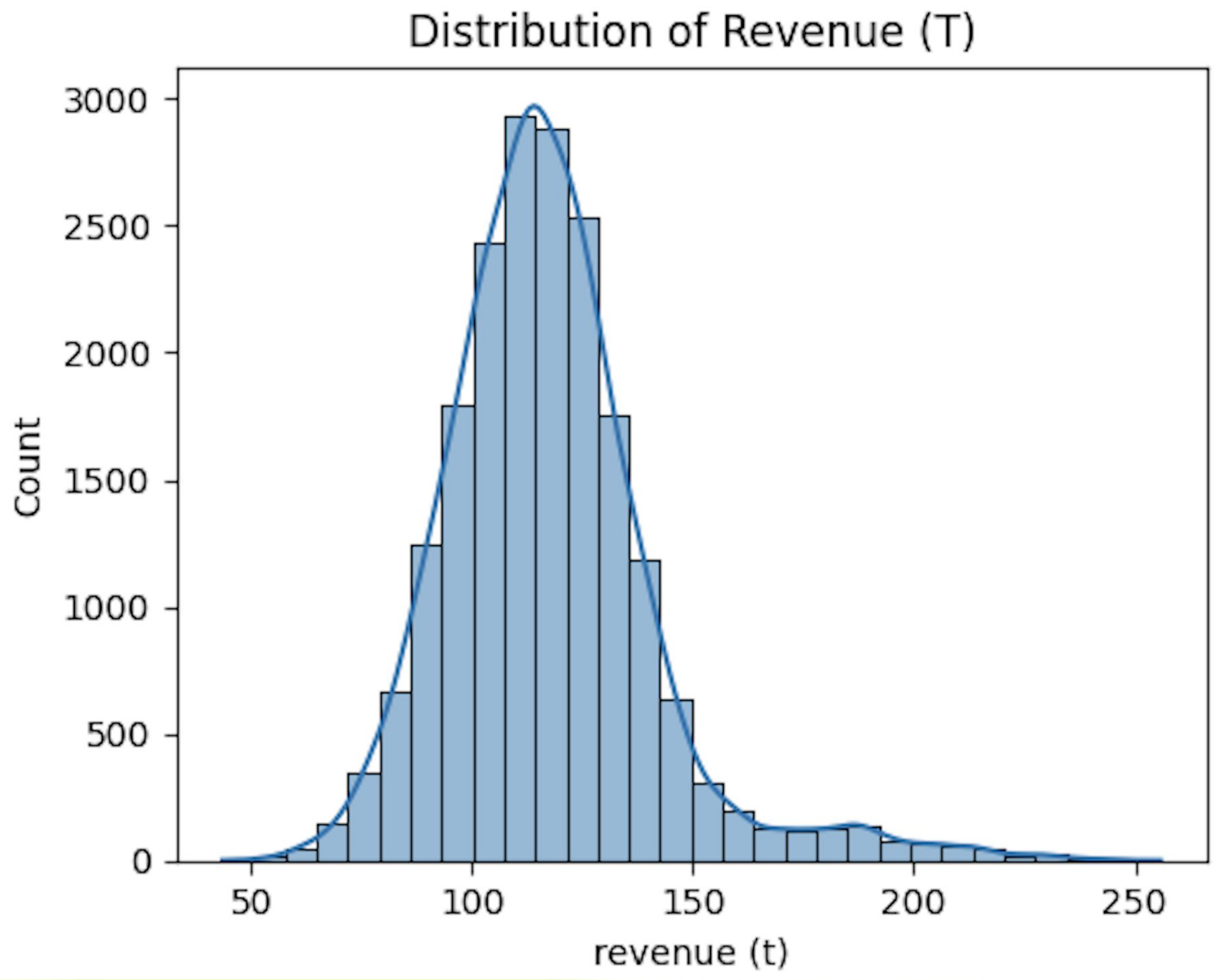
Trade-offs:

- For a **smaller effect size**, you need a **larger sample size**
- For **more power**, you need a **larger sample size**
- For **more power** or a **smaller sample size**, you need to **lower unexplained variance**

Method 1: Standard Power Analysis

This solves the power equation for an **independent, two-sample t-test.**

- It calculates the required sample size based **<u>only</u>** on the overall variance of your response variable.
- Our response variable is “revenue (t)”.
- **Inputs:**
  - Power
  - Significance Level (α)
  - MDE (standardized)
- **Output**:
  - Sample size



Method 1: Standard Power Analysis

Use the `statsmodels.stats.power` library:

1. Calculate the **standardized effect size**:
   - `effect_size = difference / standard_deviation`
   - `difference:` Your MDE must be in absolute terms, i.e. calculated as a percentage of mean “revenue (t)”.
   - `standard_deviation:` For the standard t-test, this is just the standard deviation of “revenue (t)”.
2. Call the solver with the standardized effect size, power, and alpha.

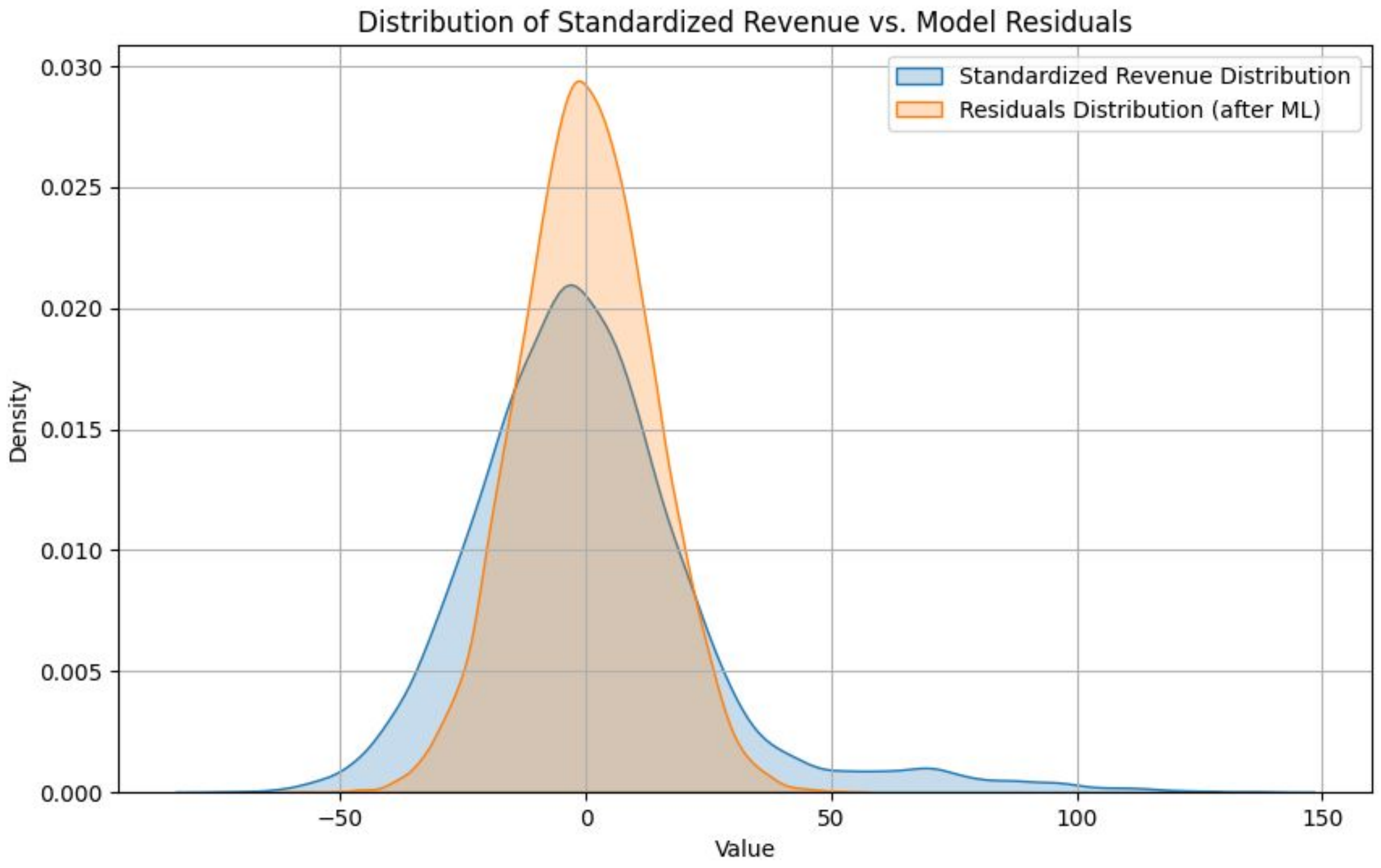
```
analysis = TTestIndPower()

effect_size = diff / target.std()
sample_size = analysis.solve_power(
    effect_size=effect_size,
    power=power,
    alpha=alpha,
    alternative="two-sided",
)
```

Method 2: MLRATE Power Analysis

This solves the same power equation, but this time using the remaining unexplained variance after an MLRATE adjustment.

- The process is nearly identical, but you'll calculate a new, larger `effect_size`.
- The `difference` (MDE) stays the same.
- The `standard_deviation` is now the standard deviation of your OLS model's residuals.

```
effect_size = difference / standard_deviation
```



Method 2: MLRATE Power Analysis

This solves the same power equation, but this time using the remaining unexplained variance after an MLRATE adjustment.

- The process is nearly identical, but you'll calculate a new, larger `effect_size`.
- The `difference` (MDE) stays the same.
- The `standard_deviation` is now the standard deviation of your OLS model's residuals.

```
effect_size = difference / standard_deviation
```

Method 1: Captures the overall variability in y (**SST**)

Method 2: Captures any remaining variability in y not explained by G (**SSE**)

$$y = \beta_0 + \beta_1 G(x) + \varepsilon$$

Noise/Variance

Coefficient of Determination → $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Sum of Squares Total → $SST = \sum (y - \bar{y})^2$

Sum of Squares Regression → $SSR = \sum (y' - \bar{y'})^2$

Sum of Squares Error → $SSE = \sum (y - y')^2$

Method 2: MLRATE Power Analysis

We will fit the following OLS model to explain away some of the variance in "revenue (t)", using a set of covariates X from t-1:

$$y = \beta_0 + \beta_1 G(x) + \varepsilon$$

Steps:

1. Use cross-fitting to generate out-of-sample predictions for each customer:
   a. Randomly split the data into two equal groups (can use sklearn.model_selection.KFold)
   b. Fit a regression model (e.g. RandomForest, GradientBoostedTree, etc.) on each group
   c. Use the model of each group to predict "revenue (t)" in the other group. Call this array G.
2. Fit the above OLS model using `statsmodels.api`
3. Extract the residuals (as a whole) from the fitted OLS object
4. Use the residuals' standard deviation in the `effect_size` division
5. Call the `solve_power()` function with this new effect size

```
import statsmodels.api as sm

model = sm.OLS(y, sm.add_constant(X))
results = model.fit(cov_type="HC0")
residuals = results.resid
```

MILESTONE #3 TASKS

Power Analysis

- **The goal is to determine the required sample size** for the upcoming experiment using the two methods discussed here. Assume the following requirements:
    - Statistical power: 90%
    - Significance level: 5%
    - MDE: 1%
- **Please do the following**:

1. Calculate the sample size using the standard t-test power analysis method.
2. Calculate the sample size using the MLRATE variance-reduced method.
3. Compare: Present both results. The MLRATE sample size *should be smaller*.
4. Recreate the residual distribution plot in slide #10. You can standardize the revenue vector (center it at 0) by subtracting the residual mean from each user.

Project milestones and timeline

These are the milestones for your Challenge Project. They include the CRISP-DM process steps you learned about in your ML Foundations course. In addition, there is an educational component in the front-end.

MILESTONE 1

- Review material from "Helpful resources" slide
- Q&A with challenge advisors

MILESTONE 2

- EDA on pre-experiment data
- Present EDA findings and data preprocessing

MILESTONE 3

- Power analysis
  - Determine the variance-reduced sample size for the upcoming RCT
  - Compare to standard t-test sample size

MILESTONE 4

- You will have received the experiment data
- Perform EDA on experimental data
- Present EDA findings and data preprocessing

MILESTONE 5

- Estimate ATE using two approaches:
  - First by using the standard t-test
  - Then by controlling for pre-experimental revenue

MILESTONE 6

- Estimate MLRATE
- Compare to:
  - Standard t-test ATE
  - Pre-experiment revenue controlled ATE

MILESTONE 7

- Deliver final project presentation: RCT, MLRATE, key findings
- Present quantified gift impact and campaign recommendations
- Final Q&A and project recap