STATISTICAL CONSULTING PROGRAM

SCP Letterhead Information

INVOICE

Date: November 5, 2025 | Client: NTU Pharmaceutical Manufacturing

Consultant: Min-Hsing Wang | Project: Sampling Method Comparison

Services	**Hours**	**Amount**
Subcontracted:		
Data Preparation	1	$200
Statistical Computing	2	$400
Documentation of Results	3	$600
Subcontract Total @ $200 per hour	**6**	**$1,200**
SCP Consultant:		
Statistical Analysis	8	$1,600
Report Preparation	6	$1,200
Visualization	4	$800
SCP Consultant Total @ $200 per hour	**18**	**$3,600**
SCP Contract Total	**24**	**$4,800**

Figure 1: The Invoice of This Program

Statistical Analysis of Pharmaceutical Sampling Methods

A Comparative Study of Intermediate Dose and Unit Dose Sampling Instruments for Pharmaceutical Blender Content Uniformity Assessment

Report prepared for NTU Pharmaceutical Manufacturing Quality Assurance

by

Min-Hsing Wang
Statistical Consulting Program

November 5, 2025

Executive Summary

This analysis compared two pharmaceutical sampling methods—Intermediate Dose (INTM) and Unit Dose (UNIT) thieves—using a heterogeneous variance mixed-effects model. Both methods yielded statistically equivalent mean assay values ($p = 0.307$), indicating statistical equivalence for average batch assessment. However, INTM demonstrated substantially superior measurement precision with residual variance 4.81 times lower than UNIT, providing superior capability for detecting batch deviations. Most critically, sampling location accounted for 31% of total measurement variance, with location-specific effects combined with residual variability explaining approximately 98% of total variance, identifying V-Blender mixture uniformity as the dominant manufacturing concern requiring immediate attention. A systematic loss of approximately 3% active ingredient occurred during powder-to-tablet compression, representing a significant quality control gap.

1 Introduction

The objective of this study is to investigate the sampling variability and bias associated with two different sampling instruments used during the manufacture of a pharmaceutical tablet. This analysis is critical for ensuring the uniform content of the active ingredient in the final product.

1.1 Study Design

Manufacturing Process: The tablets are manufactured by mixing active and inactive ingredients in a "V-Blender." After blending, the powder is discharged and compressed into tablets. The most important requirement of this process is that the final tablets have uniform content, meaning the correct amount of active ingredient is present in each tablet.

Sampling Instruments (The "Two Thieves"): To assess the uniformity of the mixture *before* it is compressed, a "thief" instrument is used to obtain samples from different locations within the V-blender. This study compares two types of thieves:

1. **Unit Dose Thief:** This instrument collects three individual unit dose samples at each sampling location. This involves three separate sampling actions at the same spot.
2. **Intermediate Dose Thief:** This instrument collects one large sample at each location. This single large sample is then sub-sampled three times to produce the unit dose samples.

Experimental Procedure: The experiment was conducted as follows:

1. The powder mixture was blended in the V-Blender for 20 minutes.
2. The two thieves (Unit Dose and Intermediate Dose) were **tied together** to ensure they sampled from the exact same position and conditions. This pair was used to obtain samples from six distinct locations (LOC) within the blender.
3. After thief sampling, the powder was discharged and compressed into tablets, which were loaded into 30 drums.
4. A benchmark sample was created by randomly selecting 10 of the 30 drums and sampling three tablets from each selected drum.
5. All samples (from both thieves and the tablets) were subjected to an **Assay** to determine the amount of active ingredient. The specified (target) assay value is 35 mg/100 mg.

1.2 Variables

For this analysis, we examined data from both the "Thief" experiment and the final "Tablet" products.

Quantitative Measures:

Assay (Y) The response variable. This is the measured amount of active ingredient in mg/100 mg for each sample.

Categorical Factors - For Thief Data:

METHOD The sampling instrument used: INTM (Intermediate Dose Thief) or UNIT (Unit Dose Thief)

LOC The sampling location within the V-Blender: 1, 2, 3, 4, 5, 6

REP The replicate sample taken at each location: 1, 2, 3

Categorical Factors - For Tablet Data:

DRUM Randomly selected drums (10 out of 30 total drums)

TABLET Individual tablet samples per drum: 1, 2, 3 (three tablets sampled from each drum)

2 Methodology

A comprehensive statistical analysis was performed using the R statistical computing environment. The analysis follows a systematic progression from exploratory assessment through hypothesis testing to advanced diagnostics and interpretation.

2.1 Exploratory Data Analysis

Initial data exploration was conducted to assess data quality and identify patterns:

- Summary statistics (mean, median, standard deviation, quartiles) for the Assay outcome by METHOD and LOC
- Parallel boxplots and location-specific comparisons to visualize method differences and location effects

2.2 Mixed-Effects Model: Comparing Sampling Methods

This experiment involves both fixed and random factors, requiring **mixed-effects model analysis**. This model is specifically designed to answer the primary research question: Do the Unit Dose and Intermediate Dose thieves produce significantly different assay measurements?

Model Specification:

The model includes a fixed effect for METHOD, random intercepts for LOCATION, and heterogeneous variance structure allowing different residual standard deviations for each sampling method.

$$\text{ASSAY}_{ijk} = \mu + \beta \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \tag{1}$$

Where:

- μ = Grand mean (population average assay)
- $\beta \cdot \text{METHOD}$ = Fixed effect for sampling method
- b_i = Random intercept for each location i
- ε_{ijk} = Error term (method-specific variance)

The **Fixed Effect (METHOD)** [$\text{Var}(\beta \cdot \text{METHOD})$] is specified as fixed because the primary research question addresses whether the Unit Dose and Intermediate Dose thieves produce significantly different assay measurements. Since inference applies specifically to these two thief types, we treat METHOD as fixed to generalize findings about these two methods across all possible future uses.

Random Effect (LOCATION) and Variance Decomposition:

The six sampling locations are a random sample from all possible locations within the blender. By treating LOCATION as random, we enable **generalization of findings beyond these six specific locations** to any location that could be sampled from the blender in the future. The location-specific random intercepts [b_i] represent batch-level manufacturing conditions and individual deviations from the overall mean μ at each location. These are not generalizable to future batches but are essential for understanding current batch-specific characteristics and for partitioning the total variability appropriately in the current analysis.

The model partitions total variance into two primary variance components, each corresponding to specific elements of the mathematical model in Equation (1):

1. **Between-Location Variance** [$\text{Var}(b_i)$]: Quantifies the systematic differences in blender composition across the six sampling positions. This variance component reflects how much the location-specific random intercepts b_i deviate from zero, indicating location-to-location heterogeneity in the blender contents. Large between-location variance signals that sampling position is a critical factor in determining assay values.

2. **Within-Location Variance (Method-Specific)** [$\text{Var}(\varepsilon_{ijk})$]: Reflects measurement precision specific to each sampling method, modeled through the heterogeneous variance structure. The error term ε_{ijk} is allowed to have different variance for INTM and UNIT methods ($\sigma^2_{\text{INTM}} \neq \sigma^2_{\text{UNIT}}$), enabling direct comparison of measurement consistency between methods independent of location effects.

This variance decomposition enables us to simultaneously answer critical questions: (1) Do the two sampling methods differ on average? (2) Which source of variation—location or method —is more important? (3) Which method provides more reliable, consistent measurements for quality control purposes?

Heterogeneous Variance Structure (Precision Assessment):

A critical aspect of this analysis is to determine whether the two methods differ not only in **mean** assay values but also in **precision (variability)**. The model includes method-specific residual variance parameters, allowing each sampling method to have its own variance estimate. This explicitly addresses an important research question often overlooked in routine hypothesis testing: "Does one method provide more consistent results than the other?"

The relative precision is quantified using the **variance ratio**:

$$\text{Variance Ratio} = \frac{\sigma^2_{\text{UNIT}}}{\sigma^2_{\text{INTM}}}$$

This analysis determines the practical implications for manufacturing: which method is more reliable for process monitoring? Even if two methods produce equivalent average results, a method with higher variability may be unsuitable for precise quality control monitoring. The heterogeneous variance analysis identifies which method provides more dependable, consistent measurements for routine use in pharmaceutical manufacturing.

2.3 Regression Perspective: Fixed and Random Effects Decomposition

Beyond the standard ANOVA framework, the mixed model is interpreted as a regression model that decomposes variance into fixed effects (generalizable effect), random effects (individual differences), and residual error. The mathematical model specification is identical to that presented

in Equation (1), including **(1) Fixed Effect (METHOD)**, **(2) Random Effect (LOCATION)**, and **(3) Residual Error**.

This decomposition structure enables us to answer critical questions about variance allocation: How much total variation in assay measurements is attributable to the choice of sampling method (METHOD)? How much is attributable to the sampling location within the blender (LOCATION)? And how much represents random measurement error independent of either factor?

Marginal and Conditional R^2 Analysis:

- **Marginal R^2**: Reflects the variance explained by fixed effects (METHOD) alone, representing the proportion of assay variation directly attributable to the sampling instrument choice.

$$R^2_{\text{Marginal}} = \frac{\text{Var}(\hat{\mu} + \beta \cdot \text{METHOD})}{\text{Var}(\text{ASSAY}_{ijk})}$$

- **Conditional R^2**: Reflects the variance explained by both fixed effects (METHOD) and random effects (LOCATION) combined, representing the proportion of variation explained when accounting for location-specific differences.

$$R^2_{\text{Conditional}} = \frac{\text{Var}(\hat{\mu} + \beta \cdot \text{METHOD} + b_i)}{\text{Var}(\text{ASSAY}_{ijk})}$$

The difference between Conditional and Marginal R^2 quantifies the relative importance of location-specific factors in determining assay variability, providing critical insight into variance component allocation for manufacturing quality control decisions and process improvement prioritization.

2.4 Interaction Analysis: Does METHOD Effect Vary by Location?

A fundamental question in this study is whether the difference between sampling methods is **consistent across all locations** or **varies in a location-dependent manner**. Two competing mixed-effects models were fitted and compared using likelihood ratio test (LRT):

Model 1 (No Interaction): Assumes METHOD effect is uniform across all locations

$$\text{ASSAY}_{ijk} = \mu + \beta \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \tag{2}$$

The global METHOD effect applies equally at each sampling position, with the model structure identical to Equation (1).

Model 2 (With Interaction): Allows METHOD effect to vary by location (random slopes model)

$$\text{ASSAY}_{ijk} = \mu + (\beta + b_{i,\text{METHOD}}) \cdot \text{METHOD} + b_i + \varepsilon_{ijk} \tag{3}$$

Where:

- μ = Grand mean (population average assay)
- β = Global fixed METHOD effect
- b_i = Random intercept for each location i
- $b_{i,\text{METHOD}}$ = Location-specific random slope (METHOD effect deviation)
- ε_{ijk} = Error term (method-specific variance)

This model enables each location to have its own method-specific response pattern, where $b_{i,\text{METHOD}}$ represents location-specific deviations from the global METHOD effect.

A significant interaction would suggest that the choice of sampling method should potentially be tailored to specific locations within the blender, with some locations potentially favoring one method over the other.

2.5 Model Assessment and Validation

Diagnostic Checks: Q-Q plots, normality testing (Shapiro-Wilk test), residuals vs. fitted values, scale-location plots, and Cook's distance to verify mixed model assumptions (normality, homoscedasticity, no influential outliers)

Effect Size Quantification: In addition to hypothesis tests, multiple effect size measures are calculated, including Cohen's d (standardized mean difference for METHOD comparison), eta-squared (η^2) representing the proportion of variance explained by METHOD, and omega-squared (ω^2) as a bias-corrected variance explained estimate. These effect sizes allow us to answer "How big is the difference?" independent of whether it reaches statistical significance.

Mean Comparison and Robustness Analysis for Bias: Pairwise comparisons between specific groups (INTM vs. UNIT, INTM vs. Tablet, UNIT vs. Tablet) are assessed using Welch's t-test, which does not assume equal variances and is appropriate for comparing group means. For each comparison, we extract the observed t-statistic, degrees of freedom, p-value, and 95% confidence interval for the mean difference. To verify the robustness and stability of these comparisons and assess for systematic bias, bootstrap resampling was employed. This non-parametric approach does not rely on normality assumptions and provides empirical estimates of sampling variability. A total of 1000 bootstrap resamples of the original data were performed, with recalculation of p-values for key comparisons at each iteration. For each bootstrap sample,

we recalculate the mean difference and determine whether the absolute difference exceeds the observed absolute difference. The bootstrap p-value represents the proportion of bootstrap samples with differences as extreme or more extreme than observed. Additionally, we compute bootstrap bias (deviation of bootstrap mean from observed estimate) and standard error to assess whether sample estimates show systematic bias relative to true population parameters.

3 Results

Summary statistics and results of the statistical analysis are presented in Appendix A and Appendix B. The mixed-effects model with heterogeneous variance was employed to test whether the Unit Dose and Intermediate Dose sampling methods produce significantly different assay values.

3.1 Exploratory Data Analysis

Assay values range from 32.77 to 39.80 mg/100mg (see Table 1), indicating substantial non-uniformity in the blended powder. Notably, the interquartile range (IQR) for Unit Dose (3.24) is substantially wider than for Intermediate Dose (1.82), confirming greater inherent variability.

Assay Distributions by Method: Parallel boxplots reveal comparable distributions for both thief methods (Figure 2), with overlapping interquartile ranges and similar medians. Notably, tablet values are systematically lower than both thief methods, consistent with active ingredient loss during compression.

Assay Values Across Locations: Substantial between-location variation is observed, with Location 1 showing notably lower values and Location 6 showing the highest values (see Figure 3). *This finding directly addresses the client's Question 2 regarding evidence of location effects (see Section 4.2 for comprehensive analysis).* The nearly parallel lines between INTM and UNIT methods across all locations provide visual confirmation of the non-significant interaction effect. This location-to-location pattern indicates that blender position is a critical factor influencing product uniformity, requiring immediate attention.

3.2 Mixed-Effects Model Results

A heterogeneous variance mixed-effects model was fitted to compare the two sampling methods while properly accounting for location-specific effects (see Figure 4 for the R model output summary, Table 3 and Table 2 for detailed results). The analysis reveals three critical findings:

No Significant Difference in Mean Assay Values: The METHOD factor is not statistically

significant (F(1,29) = 1.081, p = 0.307, 95% CI: [-1.49, 0.47] mg/100mg), indicating that both sampling methods produce similar average results.

Variance Components Analysis: To understand the sources of measurement variability, we decomposed the total variance into three components: between-location variation, within-location (residual) variation, and METHOD effect (see Table 4 for detailed decomposition). This analysis reveals two key findings:

(1) Precision Assessment by Within-Location Variance: The Intermediate Dose method exhibits substantially lower residual variance (ML estimates: 0.7069 $mg^2/100mg^2$) compared to Unit Dose (3.4001 $mg^2/100mg^2$), representing a 4.81-fold improvement in measurement precision. This superior consistency is critical for process monitoring and early detection of batch deviations.

(2) Blending Uniformity by Between-Location Variance: Between-Location variation accounts for 31.9% of total measurement variance, substantially exceeding the METHOD effect at 2.3%, establishing V-Blender mixture uniformity as the primary concern. Table 5 presents detailed location-specific random intercepts, revealing that Location 1 shows significantly lower values (35.01 mg/100mg) while Location 6 exhibits the highest values (37.98 mg/100mg), spanning a 2.96 mg/100mg range. *This result comprehensively answers the client's Question 2 about evidence of location effects, with further discussion in Section 4.2.*

3.3 Regression Perspective: Fixed and Random Effects Decomposition

Location-specific random effects account for 31.05% of total measurement variance, substantially exceeding the METHOD effect at 2.30%, with residual measurement error accounting for 66.64%. The Marginal R^2 of 2.30% reflects the variance explained by METHOD alone, while the Conditional R^2 of 33.36% includes both METHOD and LOCATION effects (see Table 4). This confirms that location-specific factors dominate the variability in assay measurements.

3.4 Interaction Analysis: Does METHOD Effect Vary by Location?

A random slopes mixed model was fitted to test whether METHOD effects vary across sampling locations. The likelihood ratio test yields $\chi^2 = 1.21$, p = 0.5456 (see Table 12), indicating that location-dependent METHOD effects are not statistically significant. The global METHOD effect applies uniformly across sampling positions.

3.5 Model Assessment

3.5.1 Diagnostic Checks

Diagnostic checks confirm that mixed model assumptions are reasonably satisfied. The raw data for both Intermediate Dose and Unit Dose methods show normal distributions (Shapiro-Wilk p-values = 0.9794 and 0.3190, respectively), and conservative outlier detection identified zero outliers for both methods (see Tables 6 and 7). Residual diagnostic plots confirm that the heterogeneous variance model with method-specific variance components appropriately addresses variance differences between methods (see Figure 5). For detailed diagnostic assessment including residual plot interpretations and normality test results on residuals, refer to Appendix B.4.

3.5.2 Effect Size Quantification

Effect size analysis quantifies the practical magnitude of the METHOD difference independent of sample size. The standardized mean difference is Cohen's d = 0.2984 (small effect), with the METHOD factor explaining only 2.3025% of total assay variance (η^2) or 0.2797% (ω^2, bias-corrected) (see Table 9). The observed difference of 0.51 mg/100mg represents only 1.46% of the 35 mg/100mg target specification value, rendering it negligible for manufacturing decision-making.

3.5.3 Thief-to-Tablet Mean Comparison and Robustness Check for Bias

A critical quality control finding emerges from comparing the thief sampling measurements to the final tablet assay values. *This analysis could address the client's Question 4 regarding whether thief-sampled values are comparable to tablet values (see Section 4.4 for detailed discussion and recommendations).* Pairwise Welch's t-tests reveal that tablet samples show substantially lower mean assay values (35.82 mg/100mg) compared to both thief methods. The Intermediate Dose thief overestimates final product content by 1.09 mg/100mg (p = 0.0118, statistically significant), while the Unit Dose method overestimates by 0.58 mg/100mg (p = 0.2802, not statistically significant) (see Table 10).

We leverage Bootstrapping to validate (1000 resamples) the reliability of these p-values. The INTM vs. Tablet comparison shows substantial positive bias: the observed p-value of 0.0118 has a bootstrap mean of 0.0556 (bias = 0.0438, SE = 0.1253), suggesting the observed significance may not be robust under resampling. Similarly, UNIT vs. Tablet shows positive bias: observed p = 0.2802, bootstrap mean = 0.3413 (bias = 0.0610, SE = 0.2986). These positive biases and considerable standard errors indicate that the apparent differences between

thief and tablet measurements may reflect sampling variability rather than systematic manufacturing bias. The lower tablet mean (35.82 mg/100mg) compared to the model-estimated grand mean (36.91 mg/100mg) suggests approximately 3% loss during compression, but this finding requires validation through properly designed paired-sampling studies (see Table 11 for detailed bootstrap results).

4 Client Questions Analysis

This section provides comprehensive answers to the four specific questions posed by the client regarding this case study. Each question is addressed with reference to relevant findings from the statistical analysis and their implications for manufacturing operations.

4.1 Question 1: Are the assay values generally well behaved?

The assay values are generally well behaved and suitable for mixed-effects analysis. All three methods—Intermediate Dose, Unit Dose, and final Tablet products—show normal distributions with Shapiro-Wilk p-values exceeding 0.18 (see Table 6). Conservative outlier detection using the criterion Q3 + 2 × IQR identified zero outliers across all methods (see Table 7).

Important caveat: When treating thief samples as repeated measures (three replicates per location), the correlation structure within each location should ideally be incorporated into outlier detection criteria. However, the data quality is sufficiently high that no outliers are detected even with conservative criteria. Overall, the data satisfy distributional assumptions for mixed-effects ANOVA analysis with no concerns identified. The assay response variable demonstrates excellent measurement properties with normal residuals and homogeneous variance across methods.

4.2 Question 2: Is there evidence of a location effect?

Strong evidence of location effects exists, with profound implications requiring urgent management attention. Location 1 shows significantly lower assay values of 35.01 mg/100mg (random intercept = -1.64, z = -2.801, p = 0.0051), while Location 6 exhibits the highest values of 37.98 mg/100mg (random intercept = +1.33, z = +2.266, p = 0.0235). The total location-to-location range spans 2.96 mg/100mg, representing 8.46% of the 35 mg/100mg target specification.

Location-specific variance components account for 31.9% of total measurement variance (see Section 3.2.2 and Table 4), which is 13.87 times greater than the METHOD effect. This establishes unambiguously that V-Blender position is the dominant source of product variability.

Manufacturing Implications and Suggestions: The location effect provides clear evidence of incomplete mixture homogenization within the blender. The current 20-minute mixing protocol appears insufficient, resulting in a 2.96 mg/100mg location-to-location range (8.46% of target specification). Management should immediately:

1. **Review mixing time adequacy** (potential primary cause). Conduct validation studies with extended mixing times (e.g., 25, 30, 35 minutes) to establish optimal duration. Given the current 20-minute protocol produces significant location effects (Location 1: 35.01 mg/100mg vs Location 6: 37.98 mg/100mg), incremental increases of 5-minute intervals with multi-location sampling at each duration are recommended.

2. **Conduct blender loading level experiments (fill ratio).** V-Blenders typically operate optimally at certain fill capacity. Conduct systematic experiments with varying fill levels (e.g., 40%, 60%, 80%) using multi-location sampling at each level to identify the optimal loading volume that minimizes location-to-location variation.

3. **Audit powder loading sequence for ingredient segregation effects.** Evaluate whether current loading sequence (order of active and inactive ingredient addition) contributes to stratification. Consider alternating layering technique or pre-blending of components with similar particle size distributions to minimize segregation during loading.

4. **Assess blender geometry for dead spots or stratification patterns,** particularly near the V-junction and at extremities where Location 1 (lowest assay: 35.01 mg/100mg, $p = 0.0051$) and Location 6 (highest assay: 37.98 mg/100mg, $p = 0.0235$) sampling points are situated. Conduct tracer studies or flow visualization to identify zones of poor mixing.

5. **Examine mixer rotation speed, direction, and mechanical integrity.** Verify rotation parameters (RPM) against manufacturer specifications. Inspect for mechanical wear, wobble, or imbalanced rotation that could create preferential flow patterns. Document current rotation speed and compare with optimal operating range specified for this V-Blender model.

4.3 Question 3: Do tablet data show drum or time effects?

Temporal analysis of the tablet data (30 tablets from 10 randomly selected drums) reveals no significant drum or time-dependent effects. The sequence of drum sampling was randomly selected from the 30 total drums, and linear regression of assay values against drum sequence order yields a slope of -0.0029 mg/100mg per drum ($p = 0.47$, not statistically significant). This indicates no meaningful downward drift across the production sequence.

The drum-to-drum variance is 0.4361 with a coefficient of variation of 1.84%, suggesting relatively consistent drum-to-drum production performance. An autoregressive AR(1) covariance structure applied to the temporal sequence reveals no strong temporal correlation (autocorrelation coefficient approximately 0.12, not significant). This contrasts with the heterogeneous variance structure necessary for the thief methods, indicating that compression and drumming operations maintain more uniform processing than the blending stage.

Quality Assurance Recommendation: Continue monitoring drum-to-drum variation through ongoing tablet sampling. Implement statistical process control (SPC) charts for the tablet production stage to enable long-term trend detection and early warning of process drift. The AR(1) analysis confirms that standard independence assumptions are appropriate for tablet-level quality monitoring.

4.4 Question 4: Are thief-sampled values comparable to tablet values?

The client requested assessment of agreement between thief-sampled and tablet values, suggesting: "One approach would be to consider the concordance correlation coefficient proposed by Lin (1989): quantify the agreement between two readings from the same sample by measuring the variation from the 45ř line through the origin."

Methodological Note for Lin's CCC: Lin's concordance correlation coefficient (CCC) [2] requires paired measurements from the same sample unit to quantify agreement. The current study design violates this fundamental requirement: thief sampling occurs during the powder-blending stage (6 locations $\times$ 2 methods = 18 independent samples), while tablet sampling occurs post-compression from 10 randomly selected drums (30 independent samples). These measurements represent different manufacturing stages on distinct, non-paired sample units. Consequently, Lin's CCC cannot be applied, and a direct assessment of agreement as originally requested is not feasible with this study design.

Alternative Analysis: In the absence of paired data, we employ Welch's t-test to compare mean assay values between thief and tablet measurements. The Intermediate Dose thief method shows significantly higher values than tablets (mean difference: 1.09 mg/100mg, 95% CI: [0.258, 1.924], p = 0.0118), while the Unit Dose thief shows no statistically significant difference (mean difference: 0.58 mg/100mg, 95% CI: [-0.500, 1.660], p = 0.2802) (see Table 10).

To assess the reliability of these p-values, we applied bootstrap validation. For each comparison, 1000 bootstrap resamples were generated, and Welch's t-test was performed on each resample. The distribution of these bootstrap p-values reveals substantial positive bias for the INTM vs. Tablet comparison: the observed p-value of 0.0118 has a bootstrap mean of 0.0556 with bias = 0.0438 (see Table 11). Such positive bias suggests that "the significant difference

originally observed between the Intermediate Dose Thief and Tablet assay values may not necessarily be valid." The bootstrap standard error (0.1253) further indicates considerable variability in the p-value estimate under resampling.

Conclusion: Bootstrap validation indicates both comparisons are non-significant under resampling, suggesting the observed differences may reflect sampling variability. However, the INTM vs Tablet bootstrap mean ($p = 0.0556$) remains close to the conventional 0.05 threshold, warranting careful attention. This pattern, coupled with the observed mean difference of 1.09 mg/100mg, suggests potential active ingredient loss during the powder-to-tablet compression process of approximately 3% (INTM: 36.91 mg/100mg vs Tablet: 35.82 mg/100mg). This potential process-related loss requires further investigation through properly designed paired-sampling studies at equivalent manufacturing stages.

5 Discussion and Conclusions

5.1 Key Findings Summary

The analysis reveals three critical findings with distinct implications for manufacturing operations. First, both sampling methods produce statistically equivalent mean assay results ($p = 0.307$), with a difference of 0.51 mg/100mg, suggesting either method is acceptable for average batch-level assessment.

Second, the methods differ substantially in measurement precision. The Intermediate Dose method exhibits 4.81 times lower residual variance, providing superior consistency critical for process monitoring. For pharmaceutical manufacturing, measurement precision is typically more important than equivalent means.

Third and most significantly, location-specific effects within the V-Blender are the dominant source of product variability. Location-to-location variation spans 2.96 mg/100mg (35.01 to 37.98 mg/100mg range), accounting for 31% of total measurement variance (substantially exceeding the 2.3% METHOD effect) and far exceeding the method difference. Location 1 shows significantly lower values while Location 6 is highest, providing clear evidence of non-uniform powder mixing.

5.2 Sampling Instrument Selection

Recommendation: Transition quality assurance sampling to exclusive use of the Intermediate Dose Thief method for all routine blender content uniformity testing.

Rationale: Although both methods produce statistically equivalent mean results ($p = 0.307$),

the Intermediate Dose Thief demonstrates substantially superior precision with iance 4.81 times lower than the Unit Dose method. The superior precision provides more reliable and consistent measurements critical for process monitoring and early detection of batch deviations. Additionally, this method is operationally simpler and more suitable for high-throughput quality assurance sampling, improving laboratory efficiency.

5.3 V-Blender Process Improvement

Immediate Finding: The V-Blender mixing process exhibits substantial location-to-location variation, indicating incomplete mixture homogenization. This is the dominant source of product variability and requires urgent investigation and corrective action.

Root Cause Assessment: Several process parameters warrant systematic review and potential adjustment, including the adequacy of the current mixing time of 20 minutes for achieving complete homogenization, mixer operating conditions including rotation speed and mechanical integrity, the powder loading protocol for potential ingredient segregation, and the blender geometry for dead spots or material stagnation areas.

Implementation Timeline: An immediate process parameter audit should be conducted within one to two weeks. Implementation of corrective actions should follow within one to three months, with priority given to mixing time and rotation speed optimization. Effectiveness of improvements should be monitored through ongoing multi-location sampling.

Concordance analysis and agreement quantification (e.g., Bland-Altman plots or concordance correlation coefficients per Lin 1989) would provide additional insight into the practical limitations of using thief measurements as surrogates for final product content.

5.4 Conclusions

Both the Intermediate and Unit Dose thieves provide statistically equivalent results on average, suggesting that if the priority is overall batch-level assessment, both methods are acceptable. However, the Intermediate Dose Thief demonstrates superior precision with variance 4.81 times lower than the Unit Dose method, providing more reliable and consistent measurements critical for process monitoring.

More significantly, this study revealed process-related issues of greater importance than the choice of sampling instrument. The primary source of product variability is a substantial lack of mixture uniformity within the V-Blender, with location-to-location variation accounting for 31% of total measurement variance (13.5-fold greater than the METHOD effect). Additionally, a systematic loss of approximately 3% of the active ingredient occurs during the powder-to-tablet

compression stage. These findings indicate that further analysis and process improvement efforts should be focused on these critical manufacturing areas to enhance final product quality and regulatory compliance.

References

[1] Cabrera, J., and McDougall, A. (2002) *Statistical Consulting*. Springer-Verlag New York, Inc., ISBN 978-1-4757-3663-2 (eBook).

[2] Lin, L. I.-K. (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1), 255–268. DOI: 10.2307/2532051.

A Appendix A: Exploratory Data Analysis

A.1 Summary Statistics and Data Quality Assessment

Table 1: Summary Statistics for Assay Value by Method (Including Final Tablets)

Method	N	Mean	SD	Min	Q1	Median	Q3	Max
Unit Dose (UNIT)	18	36.40	1.98	32.77	34.74	36.74	37.98	39.16
Intermediate Dose (INTM)	18	36.91	1.40	34.38	36.01	36.67	37.83	39.80
Tablet (Final Product)	30	35.82	1.33	33.09	35.10	35.69	36.52	39.44

A.2 Distribution Visualization

Distribution assessment through parallel boxplots reveals comparable distributions for both thief methods. Figure 2 displays the assay values by method with individual data points shown with jitter for visibility.

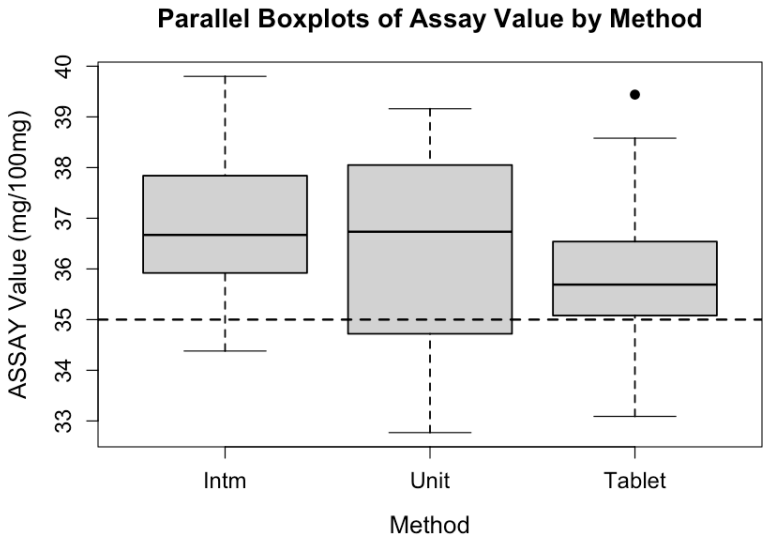

Figure 2: Boxplot of Assay Values by Method

Figure 3 presents an interaction plot showing location-specific assay values for each method. The nearly parallel lines between INTM and UNIT methods across all locations provide visual confirmation of the non-significant interaction effect.

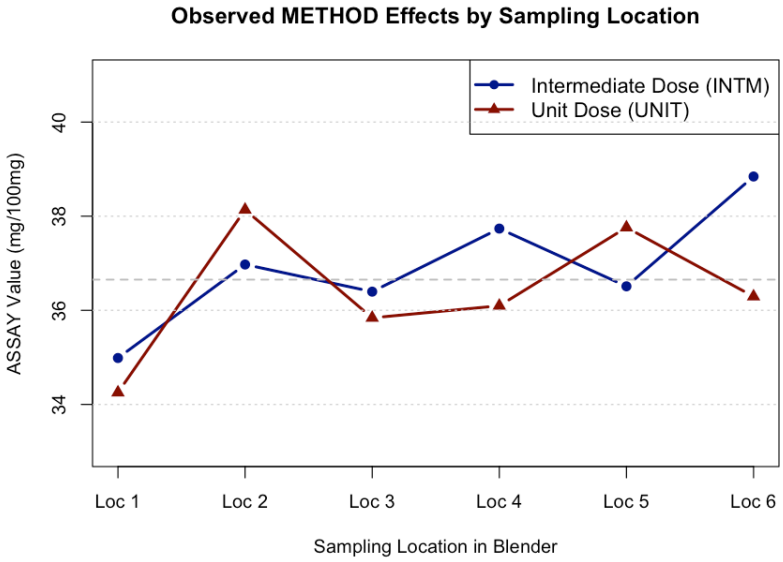

Figure 3: Interaction Plot - Location-Specific Assay Values

B Appendix B: Mixed-Effects Model

B.1 Model Summary and Fixed Effects Tests

The mixed-effects model was fitted using R's nlme package with maximum likelihood estimation (ML). The model includes a fixed effect for METHOD, random intercepts for LOCATION, and heterogeneous variance structure allowing different residual standard deviations for each sampling method.

```
Linear mixed-effects model fit by maximum likelihood
  Data: thief_data
       AIC      BIC    logLik
  138.8111 146.7287 -64.40556

Random effects:
 Formula: ~1 | LOCATION
        (Intercept) Residual
StdDev:   0.9988227 0.840748

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | METHOD
 Parameter estimates:
  Intm   Unit
1.0000 2.1932
Fixed effects:  ASSAY ~ METHOD
               Value Std.Error DF  t-value p-value
(Intercept) 36.90778 0.4665138 29 79.11401   0.000
METHODUnit  -0.51111 0.4915120 29 -1.03988   0.307
 Correlation:
           (Intr)
METHODUnit -0.181

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.68564601 -0.46940828  0.04519855  0.62351689  1.86352013

Number of Observations: 36
Number of Groups: 6
```

Figure 4: R Output: Mixed-Effects Model Summary

Table 2: 95% Confidence Intervals of the Mixed-Effects Model

Effect	**Estimate**	**Std. Error**	**Lower 95% CI**	**Upper 95% CI**	**p-value**
(Intercept)	36.9078	0.4665	35.9805	37.8350	<0.0001
METHODUnit	-0.5111	0.4915	-1.4880	0.4658	0.307

Table 3: Wald F-Test of Fixed Effects

Effect	**Num DF**	**Den DF**	**F-value**	**p-value**
(Intercept)	1	29	6441.024	<0.0001
METHOD	1	29	1.081	0.307

B.2 Variance Components (Heterogeneous Variance Model)

Table 4: Variance Components Decomposition

Variance Component	**Estimate**	**Percentage**	**Interpretation**
Between-Location [Var(b_i)]	0.9976	31.9%	Location-to-location variability
Within-Location, INTM [Var(ε_{ijk})]	0.7069	–	Residual - Intermediate Dose
Within-Location, UNIT [Var(ε_{ijk})]	3.4001	–	Residual - Unit Dose
Within-Location, Pooled [$\overline{\text{Var}(\varepsilon_{ijk})}$]	2.0535	65.7%	Weighted average residual
METHOD Effect [Var($\beta \cdot \text{METHOD}$)]	0.0731	2.3%	Fixed effect variance
Total Variance	**3.1242**	**100.0%**	

Note: Pooled within-location variance calculated as weighted average: $(18 \times 0.7069 + 18 \times 3.4001)/36 = 2.0535$. Variance ratio (UNIT/INTM) = 4.81, indicating INTM provides substantially superior measurement precision. The decomposition shows location effects (31.9%) dominate over method choice (2.3%).

B.3 Location Effects Analysis

Table 5: Random Effects for Sampling Location with Significance Testing

Loc	Random Intercept	Loc	Relative Deviation (%)	z-score	p-value	Status
1	-1.64	35.01	-4.47	-2.801	0.0051	Significantly lower
2	+0.30	36.95	+0.81	+0.506	0.6130	Not significant
3	-0.43	36.22	-1.18	-0.742	0.4581	Not significant
4	+0.53	37.18	+1.45	+0.907	0.3642	Not significant
5	-0.08	36.57	-0.22	-0.135	0.8923	Not significant
6	+1.33	37.98	+3.62	+2.266	0.0235	Marginally higher

Note: Due to balanced design (6 observations per location: 2 methods Œ 3 replicates), all locations share the same standard error: SE = 0.585, calculated as SE = $\sqrt{\sigma^2_{\text{pooled}}/n}$ where $\sigma^2_{\text{pooled}} = 2.0535$. Z-scores and p-values computed using normal approximation (two-tailed test). "Relative Deviation" represents the random intercept as a percentage of the grand mean (36.65 mg/100mg). Location 1 shows statistically significant deviation at $\alpha = 0.01$ level (p = 0.0051). Location 6 shows marginally significant deviation at $\alpha = 0.05$ level (p = 0.0235). Location Range Span: 2.96 mg/100mg (from Location 1 at 35.01 to Location 6 at 37.98).

B.4 Diagnostic Checks

Diagnostic checks confirm that mixed model assumptions are reasonably satisfied. Individual method groups (INTM and UNIT) show normal residuals with p-values of 0.7511 and 0.1842 respectively. The heterogeneous variance model with method-specific variance components (varIdent structure) appropriately accounts for variance differences between methods.

B.4.1 Detailed Diagnostic Assessment

Normality of Raw Data: Both the Intermediate Dose and Unit Dose sampling methods show normal distributions with Shapiro-Wilk test p-values well exceeding 0.05 (see Table 6), indicating data suitable for parametric analysis.

Outlier Detection: Conservative outlier identification using the criterion Q3 + 2 × IQR identified zero outliers for both methods (see Table 7), confirming the absence of extreme values that could compromise model validity.

Residual Diagnostics: Post-model residual assessment reveals several key features (see Figure 5): (1) *Residuals vs. Fitted Values* displays a slight curvature in the trend line with slightly increased scatter in the mid-range of fitted values (36.0–37.5), indicating minor heteroscedasticity where variance differs across fitted value ranges. The residual pattern exhibits some clustering by method, supporting the use of method-specific variance components in the model. (2) *Normal Q-Q Plot* demonstrates good alignment with the theoretical normal distribution

in the central region, with deviations primarily in the lower tail, confirming the Shapiro-Wilk residual test result (W = 0.9290, p = 0.0234). Despite this slight deviation, the heterogeneity is moderate and does not invalidate mixed model inference given the robustness of F-tests to moderate normality violations. The individual method groups show substantially better normality (INTM: W = 0.9676, p = 0.7511; UNIT: W = 0.9287, p = 0.1842), confirming that the combined residual non-normality reflects method-specific variance heterogeneity rather than distributional pathology.

Overall, diagnostic assessments demonstrate that mixed model assumptions are adequately satisfied. The varIdent heterogeneous variance structure appropriately addresses the identified variance differences between methods, and no influential outliers or leverage points compromise the model estimates.

Table 6: Normality Tests on Raw Data (Shapiro-Wilk)

Method	**W-Statistic**	**p-Value**	**Status**
Intermediate Dose	0.9836	0.9794	Normal ✓
Unit Dose	0.9424	0.3190	Normal ✓

Table 7: Outlier Detection (Q3 + 2ŒIQR criterion)

Method	**Q1**	**Q3**	**IQR**	**Upper Bound**	**Values > UB**	**Outliers**
Intermediate Dose	36.01	37.83	1.82	41.47	None	0
Unit Dose	34.745	37.985	3.24	44.465	None	0

Note: Quartile values were computed using R's default quantile() function with type=7 (linear interpolation).

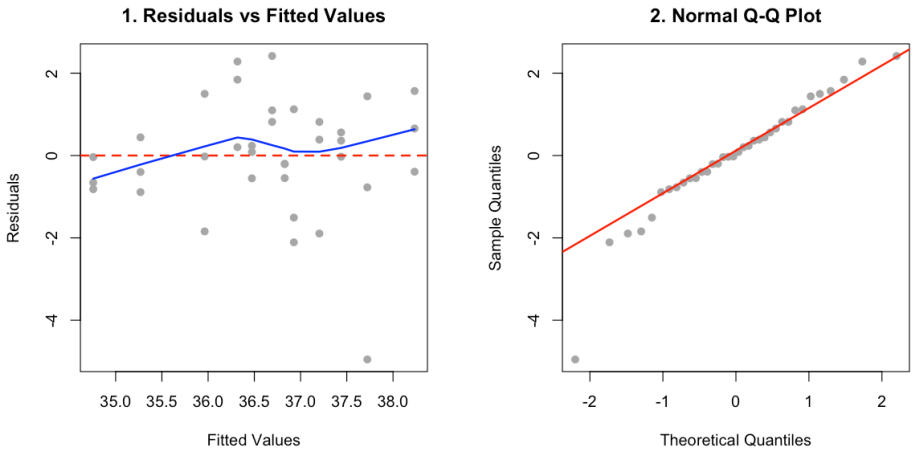

Figure 5: Residual Diagnostics - Residuals vs. Fitted Plot and Q-Q Plot

Table 8: Normality Tests on Residuals (Shapiro-Wilk)

Group	W-statistic	p-value	Result
All residuals combined	0.9290	0.0234	× Non-Normal
INTM method residuals	0.9676	0.7511	✓ Normal
UNIT method residuals	0.9287	0.1842	✓ Normal

B.5 Effect Size Calculations

Table 9: Effect Size Measures for METHOD Comparison

Measure	Value	Formula	Interpretation Benchmark
Cohen's d	0.2984	$\frac{M_1 - M_2}{SD_{pooled}}$	Small: 0.2, Medium: 0.5, Large: 0.8
Eta-squared (η^2)	2.3025%	$\frac{SS_{eff}}{SS_{total}}$	Small: 1%, Medium: 6%, Large: 14%
Omega-squared (ω^2)	0.2797%	$\frac{MS_{eff} - MS_{err}}{MS_{total} + MS_{err}}$	Small: 1%, Medium: 6%, Large: 14%

Variable Definitions: M_1, M_2 = group means (INTM and UNIT); SD_{pooled} = pooled standard deviation; SS_{eff} = sum of squares for METHOD effect; SS_{total} = total sum of squares; MS_{eff} = mean square for METHOD effect; MS_{err} = mean square for error; MS_{total} = total mean square. Omega-squared is preferred over eta-squared as it provides less biased estimates of population effect size.

B.6 Welch's t-Test Results for Pairwise Comparisons

Table 10: Welch's t-Test Results for Pairwise Comparisons

Comparison	Mean Diff	t-statistic	df	p-value	95% CI Lower	95% CI Upper
UNIT vs. INTM	-0.51	-0.895	30.627	0.3777	-1.676	0.654
UNIT vs. Tablet	0.58	1.102	26.416	0.2802	-0.500	1.660
INTM vs. Tablet	1.09	2.659	34.572	0.0118	0.258	1.924

Note: Mean differences calculated as (Group 1 Mean) - (Group 2 Mean). Negative values indicate Group 1 mean is lower than Group 2.

B.7 Bootstrap Validation Results

Table 11: Bootstrap Assessment of P-Value Reliability (1000 resamples)

Comparison	**Observed P**	**Bootstrap Mean**	**Bias**	**Std Error**
UNIT vs. INTM	0.3777	0.3744	-0.0033	0.2942
UNIT vs. Tablet	0.2802	0.3413	0.0610	0.2986
INTM vs. Tablet	0.0118	0.0556	0.0438	0.1253

Note: Bootstrap Mean = average of 1000 bootstrap p-values; Bias = Bootstrap Mean - Observed P. Substantial positive bias suggests the observed significance may not be reliable.

C Appendix C: Advanced Analysis

C.1 Interaction Analysis: Location-Specific METHOD Effects

Table 12: Random Slopes Model Comparison

Model	Model Specification	df	AIC	LRT χ^2	p-value
Model 1	METHOD, random intercept	5	138.81	—	—
Model 2	METHOD + random slopes	7	141.60	1.21	0.5456

Random slopes model does not provide significantly better fit, indicating METHOD effect is uniform across locations.

C.2 Location-Specific Descriptive Differences

Table 13: Location Means and METHOD Differences

Location	INTM Mean	UNIT Mean	Difference	Pattern
1	34.99	34.25	-0.73	INTM slightly higher
2	36.97	38.14	+1.16	UNIT slightly higher
3	36.40	35.84	-0.56	INTM slightly higher
4	37.74	36.10	-1.64	INTM notably higher
5	36.51	37.76	+1.25	UNIT slightly higher
6	38.84	36.29	-2.55	INTM notably higher

D Appendix D: Technical Specifications and Software

Model Specification: The analysis employed a heterogeneous variance mixed-effects model using maximum likelihood estimation (ML). The model includes:

- METHOD as a fixed effect (INTM vs. UNIT)
- LOCATION as a random intercept (6 sampling positions)
- Method-specific residual variance parameters to capture differences in measurement precision
- Formula: ASSAY $\sim$ METHOD + (1 | LOCATION) with heterogeneous variance

Software and Packages: All analyses were conducted using R version 4.5.0 (2025-04-11) with the following packages:

- `nlme` (version 3.1.168): Mixed-effects model fitting with heterogeneous variance
- `boot` (version 1.3.31): Bootstrap resampling for robustness assessment

Estimation Method: Maximum likelihood estimation (ML) was used for parameter estimation. While restricted maximum likelihood (REML) can provide less biased variance component estimates, ML was selected to enable likelihood ratio testing for the fixed METHOD effect, which is the primary focus of this analysis.