QDA Sensory Analysis Dashboard

Statistical Methods Documentation

Version 1.0
October 2025

Comprehensive guide to statistical methods implemented
in the QDA Sensory Analysis Dashboard

Table of Contents

1. Introduction

The QDA (Quantitative Descriptive Analysis) Sensory Analysis Dashboard implements state-of-the-art statistical methods for analyzing sensory profiling data. This document provides comprehensive documentation of all statistical methods, formulas, and interpretation guidelines used in the dashboard.

1.1 Purpose and Scope

Sensory evaluation generates complex data structures with multiple sources of variation: products, panelists, sessions, and their interactions. Traditional ANOVA approaches often fail to account for the hierarchical and repeated measures nature of sensory data. This dashboard implements Linear Mixed Models (LMM) to properly handle these complexities while providing robust statistical inference.

1.2 Data Structure

The dashboard expects data in the following format:

- **Column 1:** Session - Identifies the evaluation session (factor)
- **Column 2:** Panelist - Identifies the trained assessor (factor)
- **Column 3:** Product - Identifies the product being evaluated (factor)
- **Columns 4+:** Descriptors - Sensory attributes rated on continuous scales (numeric)

This structure allows for proper modeling of the repeated measures design inherent in sensory evaluation studies.

2. Linear Mixed Models (LMM)

Linear Mixed Models are the cornerstone of modern sensory data analysis. They extend traditional linear models by incorporating both fixed effects (what we want to test) and random effects (sources of variation we want to account for).

2.1 Model Specification

The dashboard implements the following mixed model for each sensory descriptor:

$$Rating_{ijkl} = \mu + Product_i + Panelist_j + Session_k + (Panelist \times Product)_{ij} + \varepsilon_{ijkl}$$

Where:

- μ (mu): Overall mean rating
- **Product**$_i$: Fixed effect of product (what we test)
- **Panelist**$_j$: Random effect of panelist ~ $N(0, \sigma^2_{panelist})$
- **Session**$_k$: Random effect of session ~ $N(0, \sigma^2_{session})$
- **(Panelist×Product)**$_{ij}$: Random interaction ~ $N(0, \sigma^2_{interaction})$
- ε_{ijkl}: Residual error ~ $N(0, \sigma^2_{residual})$

2.2 Why This Model?

1. Accounts for Panelist Variability: Different panelists have different baseline intensities. Some rate everything higher, others lower. The random panelist effect captures this.

2. Accounts for Session Effects: Ratings can vary between sessions due to environmental factors, panelist fatigue, or product preparation variations. The random session effect models this.

3. Captures Individual Preferences: The panelist×product interaction allows for the fact that some panelists may perceive certain products differently. This is the "individual preference" component.

4. Proper Standard Errors: By accounting for all sources of variation, the model provides accurate standard errors for product comparisons, leading to valid statistical inference.

2.3 Estimation Method: REML

The dashboard uses **Restricted Maximum Likelihood (REML)** estimation. REML is preferred over Maximum Likelihood (ML) because:

- It provides unbiased estimates of variance components
- It accounts for the loss of degrees of freedom due to estimating fixed effects
- It is the standard approach recommended by ISO 11132:2021 for sensory analysis

The REML criterion being optimized is:

$$-2 \log L_{REML} = \log|V| + \log|X'V^{-1}X| + (y-X\beta)'V^{-1}(y-X\beta)$$

3. Kenward-Roger Approximation

One of the most important features of this dashboard is the implementation of the **Kenward-Roger (KR) approximation** for degrees of freedom and standard errors. This is the gold standard for small-sample inference in mixed models.

3.1 Why Kenward-Roger?

Standard mixed model inference assumes large samples and can be **anti-conservative** (too many false positives) in small samples. The Kenward-Roger method addresses this by:

1. Adjusting Standard Errors: Corrects for the variability in estimating variance components, which is ignored in naive approaches.

2. Computing Appropriate df: Uses a Satterthwaite-type approximation to calculate appropriate degrees of freedom. This is why you see decimal df values (e.g., df = 15.83) - this is correct and reflects the complexity of the model!

3. Better Type I Error Control: Maintains the nominal $\alpha = 0.05$ level even with small sample sizes, preventing false discoveries.

4. Recommended by Standards: ISO 11132:2021 specifically mentions the need for small-sample corrections like Kenward-Roger in sensory analysis.

3.2 Mathematical Details

The Kenward-Roger approximation involves two main components:

3.2.1 Adjusted Standard Errors

The variance-covariance matrix of fixed effects is adjusted:

$$\text{Var}(\beta)_{\text{adjusted}} = \Phi \cdot \text{Var}(\beta)_{\text{naive}}$$

Where Φ is an inflation factor computed from the information matrix of the variance components. This accounts for the fact that we're estimating variance components from data, not treating them as known.

3.2.2 Degrees of Freedom

The Kenward-Roger degrees of freedom use a Satterthwaite-type approximation:

$$df_{KR} = 2E / \text{Var}(E)$$

Where E is the quantity being tested. This formula ensures that the degrees of freedom reflect both the sample size and the complexity of the variance structure. The resulting df are typically non-integer and always less than or equal to the naive residual df.

4. Post-hoc Comparisons: Tukey HSD

After finding a significant product effect, we need to determine which specific products differ. The dashboard uses **Tukey's Honest Significant Difference (HSD)** test with Kenward-Roger standard errors.

4.1 Why Tukey HSD?

When comparing multiple products, we face the **multiple comparisons problem**: The probability of at least one false positive increases with the number of tests. Tukey HSD controls the **family-wise error rate** (FWER) at $\alpha = 0.05$.

Advantages of Tukey HSD:
• Controls FWER exactly for all pairwise comparisons
• More powerful than Bonferroni correction
• Appropriate for balanced and slightly unbalanced designs
• Provides simultaneous confidence intervals
• Widely accepted in sensory science (ISO standards)

4.2 Test Statistic

For comparing products i and j, the test statistic is:

$$q = (\mu\blacksquare_i - \mu\blacksquare_j) / SE_{diff}$$

Where:
• $\mu\blacksquare_i$, $\mu\blacksquare_j$ are the estimated marginal means (LSMeans) for products i and j
• SE_{diff} is the standard error of the difference, computed using Kenward-Roger adjustment
• q follows the Studentized range distribution with parameters (k, df_{KR})
• k = number of products being compared

The adjusted p-value is: $p_{adj} = P(Q \geq |q|)$ where $Q \sim q(k, df_{KR})$

4.3 Least Squares Means (LSMeans)

LSMeans (also called Estimated Marginal Means or EMMs) are the product means **adjusted for the random effects** in the model. They differ from simple arithmetic means in important ways:

LSMeans account for:
• Unbalanced designs (different numbers of observations per product)
• Missing data
• Panelist baseline differences
• Session-to-session variation

Formula: LSMeans are predictions from the fitted model at the reference levels of random effects (all random effects set to 0). This gives us the "expected" rating for an "average" panelist in an "average" session.

4.4 Compact Letter Display (CLD)

The dashboard generates a Compact Letter Display to visualize pairwise comparison results. Products sharing a letter are **not significantly different** at $\alpha = 0.05$.

Example interpretation:

Product A: "a"
Product B: "b"
Product C: "ab"

This means:
• Products A and B are significantly different (no shared letters)
• Product C is not significantly different from either A or B (shares letters with both)
• Products A and B remain significantly different from each other

5. Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms the original sensory descriptors into a new set of uncorrelated variables (principal components) that capture the maximum variance in the data.

5.1 Mathematical Foundation

PCA is based on the eigenvalue decomposition of the correlation matrix:

```
R = PΛP'
```

Where:
• **R**: Correlation matrix of LSMeans (products × descriptors)
• **P**: Matrix of eigenvectors (loadings)
• Λ: Diagonal matrix of eigenvalues (variances)

The principal components are computed as:
$PC_k = Xp_k$

Where X is the standardized data matrix and p_k is the k-th eigenvector.

5.2 Interpretation

Eigenvalues (λ):
• Represent the variance explained by each principal component
• Kaiser criterion: Retain components with $\lambda > 1$
• Cumulative variance: Aim for 70-80% in first 2-3 components

Loadings:
• Correlation between original descriptors and principal components
• |loading| > 0.7 indicates strong association
• Used to interpret what each PC represents

Scores:
• Position of products in PC space
• Products close together are similar in sensory profile
• Distance on PC1 represents differences in the main sensory dimension

5.3 Biplots

The biplot simultaneously displays:
• Product positions (scores)
• Descriptor directions (loadings)

Reading a biplot:
• Products positioned in the direction of a descriptor have high values for that attribute
• Length of descriptor vectors indicates how well they're represented
• Angle between vectors indicates correlation (small angle = positive correlation)
• Products far from origin differ most from the average sensory profile

6. Hierarchical Agglomerative Clustering (AHC)

Hierarchical clustering groups products based on their sensory similarity without pre-specifying the number of clusters.

6.1 Algorithm

1. Distance Calculation:
The dashboard uses **Euclidean distance**:

$$d(i,j) = \sqrt{[\Sigma_k (x_{ik} - x_{jk})^2]}$$

2. Linkage Method: Ward's Criterion
Ward's method minimizes the increase in total within-cluster variance when merging clusters:

$$\Delta(C_i, C_j) = \Sigma_{x \in C_i \cup C_j} ||x - \mu_{ij}||^2 - \Sigma_{x \in C_i} ||x - \mu_i||^2 - \Sigma_{x \in C_j} ||x - \mu_j||^2$$

Where μ_i, μ_j, μ_{ij} are the centroids of clusters i, j, and their union. Ward's method tends to create compact, spherical clusters of approximately equal size.

6.2 Dendrogram Interpretation

The dendrogram shows:
• Height of merge: Dissimilarity between clusters
• Vertical lines: Products or clusters being joined
• Horizontal cuts: Define number of clusters

Choosing the number of clusters:
• Look for large jumps in height
• Consider domain knowledge
• Use in conjunction with PCA results
• Typically 2-5 clusters for sensory data

7. K-Means Clustering

K-means is a partitioning method that divides products into k clusters by minimizing within-cluster sum of squares.

7.1 Algorithm

The k-means algorithm iterates between two steps:

Assignment Step: Assign each product to the nearest centroid
$C_i = \{x : ||x - \mu_i|| \leq ||x - \mu_j|| \text{ for all } j\}$

Update Step: Recompute centroids as the mean of assigned products
$\mu_i = (1/|C_i|) \Sigma_{x \in C_i} x$

The algorithm minimizes the objective function:

$$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} ||x - \mu_i||^2$$

7.2 Elbow Method

To determine the optimal k, the dashboard uses the **elbow method**:

1. Run k-means for k = 1, 2, ..., 10
2. Compute total within-cluster sum of squares (WSS) for each k
3. Plot WSS vs. k
4. Look for an "elbow" where WSS decreases more slowly

The elbow represents the point of diminishing returns - adding more clusters doesn't substantially improve the model fit.

8. Normality Testing

Mixed models assume that random effects and residuals follow normal distributions. The dashboard provides three complementary normality tests.

8.1 Shapiro-Wilk Test

The most powerful test for small to medium samples (n < 5000).

Test Statistic:

$$W = (\Sigma_i a_i x_{(i)})^2 / \Sigma_i (x_i - x\blacksquare)^2$$

Where $x_{(i)}$ are ordered observations and a_i are weights from the covariance matrix of order statistics.

- W ranges from 0 to 1
- W close to 1 indicates normality
- p < 0.05 suggests departure from normality
- Most sensitive to outliers and asymmetry

8.2 Anderson-Darling Test

Gives more weight to the tails of the distribution, making it better for detecting tail deviations.

Test Statistic:

$$A^2 = -n - (1/n)\Sigma_i (2i-1)[\log F(x_i) + \log(1-F(x_{n+1-i}))]$$

Where F is the cumulative distribution function of the standard normal.

- Better than Shapiro-Wilk for larger samples
- More sensitive to deviations in tails
- Particularly useful for detecting heavy-tailed distributions

8.3 Jarque-Bera Test

Tests for normality based on skewness and kurtosis.

Test Statistic:

$$JB = (n/6)[S^2 + (K-3)^2/4]$$

Where:
- S = skewness = $E[(X-\mu)^3]/\sigma^3$
- K = kurtosis = $E[(X-\mu)\blacksquare]/\sigma\blacksquare$
- For normal distribution: S = 0, K = 3

JB ~ $\chi^2(2)$ under the null hypothesis of normality.

- Good for large samples
- Directly tests moment assumptions
- Less powerful than Shapiro-Wilk for small samples

9. Variance Components Analysis

Variance components quantify how much of the total variability in ratings comes from different sources. This is crucial for understanding panel performance and data quality.

9.1 Decomposition

The total variance in ratings is partitioned as:

$$\sigma^2_{total} = \sigma^2_{product} + \sigma^2_{panelist} + \sigma^2_{session} + \sigma^2_{interaction} + \sigma^2_{residual}$$

Component Interpretation:

$\sigma^2_{product}$ (Product variance):
- How much products actually differ
- Large value = good product discrimination
- Typically want this to be substantial

$\sigma^2_{panelist}$ (Panelist variance):
- Baseline differences between panelists
- Some use higher/lower parts of the scale
- Moderate values are normal and acceptable

$\sigma^2_{session}$ (Session variance):
- Day-to-day variability
- Should be relatively small
- Large values may indicate panel drift or environmental issues

$\sigma^2_{interaction}$ (Panelist × Product variance):
- Individual differences in product perception
- Some panelists perceive products differently
- Moderate values indicate good training
- Very large values may indicate poor panel agreement

$\sigma^2_{residual}$ (Residual variance):
- Unexplained random variation
- Includes measurement error and within-panelist variability
- Should be the smallest component for well-trained panels

9.2 Intraclass Correlation (ICC)

The ICC measures panel consistency:

$$ICC = \sigma^2_{product} / (\sigma^2_{product} + \sigma^2_{panelist} + \sigma^2_{interaction} + \sigma^2_{residual})$$

ICC Interpretation:
- ICC > 0.75: Excellent agreement
- ICC 0.60-0.75: Good agreement
- ICC 0.40-0.60: Moderate agreement
- ICC < 0.40: Poor agreement (panel may need retraining)

High ICC indicates that most variability comes from actual product differences rather than panelist inconsistency.

10. Model Diagnostics

Model diagnostics help verify that the assumptions of the mixed model are satisfied.

10.1 Residual Plots

Residuals vs. Fitted Values Plot:
• Checks for homoscedasticity (constant variance)
• Should show random scatter around zero
• Patterns indicate model misspecification
• Funnel shape suggests heteroscedasticity

What to look for:
• Random scatter (good)
• No systematic patterns (good)
• Constant spread across fitted values (good)
• Outliers should be investigated

10.2 Q-Q Plots

Quantile-Quantile Plot:
• Tests normality of residuals
• Points should fall along the diagonal line
• Deviations indicate departures from normality

Common patterns:
• S-curve: Heavy tails
• Curve upward at both ends: Light tails
• Points below line on left, above on right: Right skew
• Points above line on left, below on right: Left skew

Note: Mixed models are relatively robust to moderate deviations from normality, especially with larger samples.

10.3 Convergence

Convergence Warnings:
The dashboard may display warnings like "Model failed to converge" for descriptors with very low variance. This is **expected and handled automatically**:

• Dashboard falls back to simple means for non-converging descriptors
• Results are still valid for descriptors with low between-product variance
• These warnings indicate descriptors that don't discriminate products well
• Not a cause for concern in typical sensory data

11. Interpretation Guidelines

11.1 Fixed Effects Table

F-value:
• Ratio of between-product variance to within-product variance
• Larger F = stronger evidence of product differences
• F > 4 typically indicates meaningful differences

p-value:
• Probability of observing the data if products were identical
• $p < 0.05$: Significant product differences ($\alpha = 0.05$)
• $p < 0.01$: Highly significant differences
• $p < 0.001$: Very highly significant differences

Significance Stars:
• *** : $p < 0.001$
• ** : $p < 0.01$
• * : $p < 0.05$
• ns : $p \geq 0.05$ (not significant)

11.2 Pairwise Comparisons

Estimate:
• Mean difference between two products
• Positive value: First product rated higher
• Magnitude indicates practical significance

Standard Error (SE):
• Uncertainty in the estimate
• Smaller SE = more precise estimate
• Kenward-Roger adjusted (more conservative)
• Similar SEs across comparisons indicate balanced design

Degrees of Freedom (df):
• Can be non-integer (this is correct!)
• Kenward-Roger approximation accounts for model complexity
• Larger df = more information for inference
• df typically between 10-50 for sensory studies

t-ratio:
• t = Estimate / SE
• Larger |t| = stronger evidence of difference
• |t| > 2 roughly corresponds to significance

Adjusted p-value:
• Tukey-adjusted for multiple comparisons
• Controls family-wise error rate
• More conservative than unadjusted p-values
• $p < 0.05$ indicates significant difference

11.3 Product Differentiation Summary

This table aggregates pairwise comparisons across all descriptors to identify which products are most different overall.

Columns:
• **Comparison:** Product pair being compared
• **Highly Differentiated:** Number of descriptors with $p < 0.001$
• **Differentiated:** Number of descriptors with $0.001 \leq p < 0.05$
• **Not Differentiated:** Number of descriptors with $p \geq 0.05$
• **Total Descriptors:** Total number of sensory attributes

Interpretation:
• High "Highly Differentiated" count = Products very different
• High "Not Differentiated" count = Products similar
• Use to prioritize which product pairs to focus on

12. References

Standards and Guidelines:

ISO 11132:2021. Sensory analysis - Methodology - Guidelines for monitoring the performance of a quantitative sensory panel. International Organization for Standardization.

Statistical Methods:

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983-997. doi:10.2307/2533558

Lenth, R. V. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.0. https://CRAN.R-project.org/package=emmeans

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99-114. doi:10.2307/3001913

Multivariate Methods:

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R* (2nd ed.). CRC Press.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244. doi:10.1080/01621459.1963.10500845

Normality Testing:

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611. doi:10.2307/2333709

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769. doi:10.1080/01621459.1954.10501232

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259. doi:10.1016/0165-1765(80)90024-5

Software:

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). shiny: Web Application Framework for R. R package version 1.7.5. https://CRAN.R-project.org/package=shiny

Appendix: Frequently Asked Questions

Q: Why are the degrees of freedom decimals?

A: This is correct! Mixed models use the Kenward-Roger approximation, which produces non-integer degrees of freedom. This reflects the complexity of the model's variance structure. Integer df would be incorrect for mixed models.

Q: Why are all standard errors similar?

A: In balanced designs (equal observations per product), standard errors are naturally similar. This indicates excellent experimental design, not an error. Variability comes from random effects being similar across comparisons.

Q: What does 'Model failed to converge' mean?

A: This occurs for descriptors with very low variance (products don't differ much). The dashboard automatically falls back to simple means. This is not a problem - it just means that particular descriptor doesn't discriminate products well.

Q: Should I transform my data?

A: Only if normality tests show strong violations AND you see patterns in residual plots. Mixed models are robust to moderate non-normality. If you do transform, use sqrt(x) for counts, log(x) for right-skewed data, or arc-sin(sqrt(x)) for proportions.

Q: How do I report these results?

A: Example: "Products differed significantly for Sweetness ($F_{2,15.83} = 12.45$, $p < 0.001$). Tukey HSD tests with Kenward-Roger adjustment revealed that Product A (LSMean = 6.2, SE = 0.3) was significantly sweeter than Product B (LSMean = 4.1, SE = 0.3, $p < 0.001$)."