On the applicability of ANOVA models for CATA data

Michael Meyners
Procter & Gamble
meyners.m@pg.com

Anne Hasted
Qi Statistics
anne@qistatistics.co.uk

Background and objective

- Check-All-That-Apply (CATA) questions return binary data from every subject on each product and attribute.
- Cochran's *Q* test[1] to compare multiple samples, pairwise comparisons using McNemar's test[2], but has limitations:
  - Incomplete and/or imbalanced data
  - Samples based on an experimental design (requires more complex models)
  - Testing more complex hypotheses (e.g. interactions of gender and samples)
- All these situations accommodated easily using classical ANOVA **if valid** for CATA data
- Cochran[1] noted similarity between results for *F*- and *Q*-test, confirmed by own experience (Fig. 1a), with p values for *F* slightly larger than those for *Q* if they are small, Fig. 1b).

Question: Can we safely use ANOVA on the binary responses from CATA, and if so under which conditions?

↳ Unclear if/how this can be addressed mathematically, therefore empirical approach chosen here

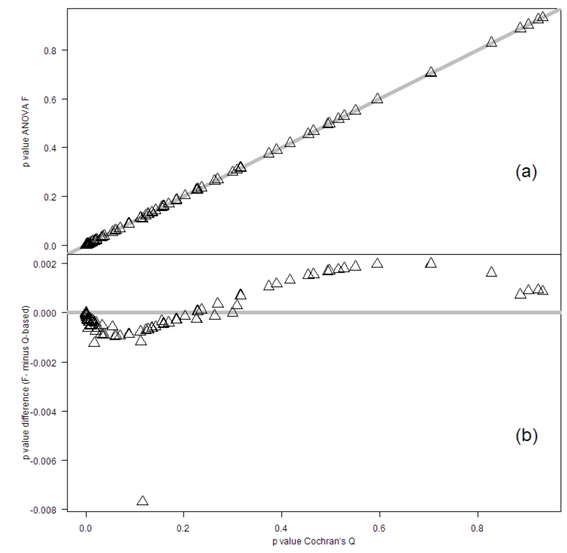

Fig. 1: p values from *Q*-test vs. those from *F*-test (a) and difference between p values (b) across various studies and attributes

Materials & Methods

- Data from 5 different CATA studies (sensory and wellbeing)
- Variety of sample sizes (54-161), # of products (5-13) and attributes (12-31)
- Randomization distribution under H_0 for every specific situation compared with respective *F*- or *t*-distribution
- Randomization test valid by design
  ⇨ if distributions coincide, ANOVA-based test (approximately) validated
- 10,000 re-randomizations for each of the following scenarios:
  - Simple 2-way ANOVA without interactions
  - Decreasing sample sizes: n = 100, 70, 50, 30, 20
  - Incomplete data with n = 150, 100, 50, 30 and rate missing 5, 10, 20, 30 and 50%

  } 10 subsamples with 1,000 randomizations each
  - 2-factorial design with interactions (subject as another factor)
- More than 20,000 Figures (tables provide elicitation counts), summarized in videos (partly embedded, click on the Figures to play, to become fully available[3])

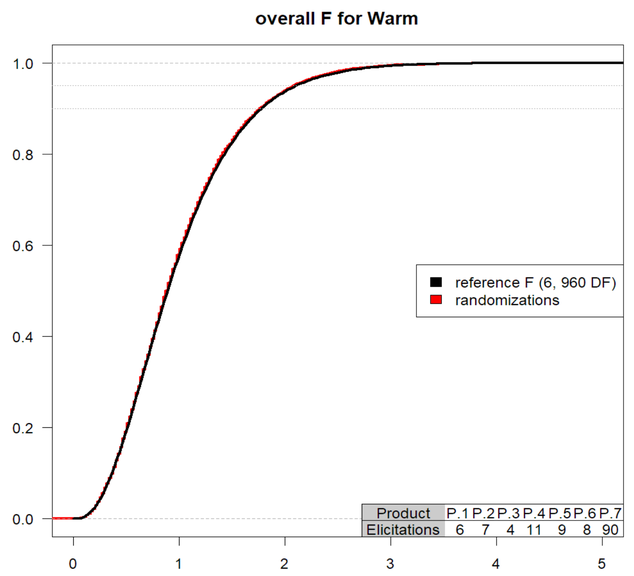

Fig./Vid. 2: empirical and parametric distribution for ANOVA-based *F*

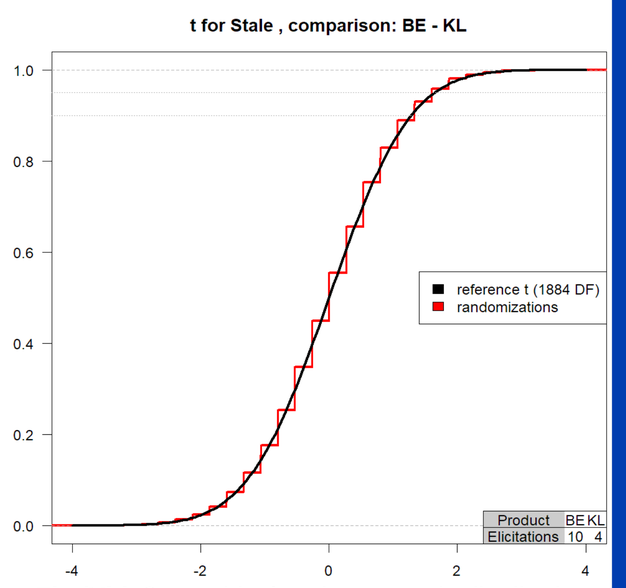

Fig./Vid. 3: empirical and parametric distribution for ANOVA-based *t*

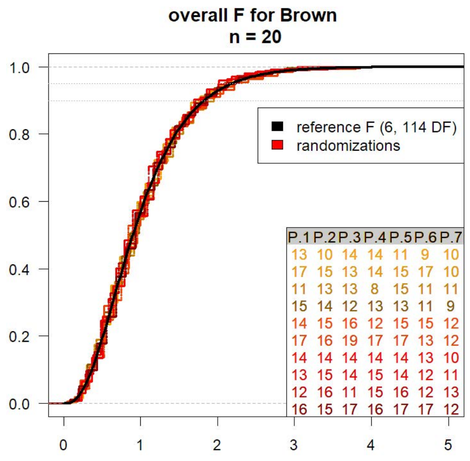

Fig./Vid. 4: empirical and parametric distribution for ANOVA-based *F* with various sample sizes

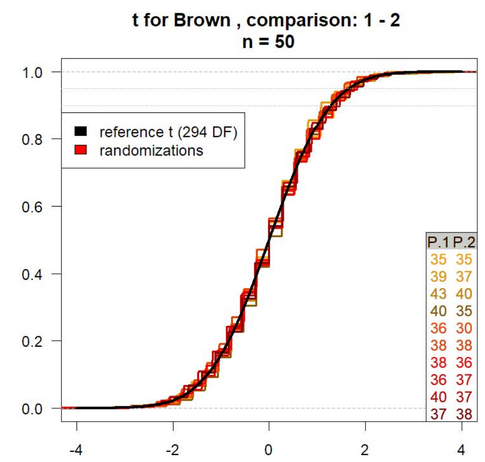

Fig./Vid. 5: empirical and parametric distribution for ANOVA-based *t* with various sample sizes

Results standard 2-way ANOVA (see Figures / Videos 2, 3)

- Good fit between distributions for *F*-test except for low overall elicitation rates
- Recommendation: ≥ 10 elicitations for most or all products
- *t*-test also with good fit, but only robust for ≥ 20 elicitations on average

Results decreasing sample sizes (see Figures / Videos 4, 5)

- Acceptable fit for sample sizes ≥ 50; below, parametric tests might be liberal
- *F*-test more robust than *t*-test (as expected, as more data overall)

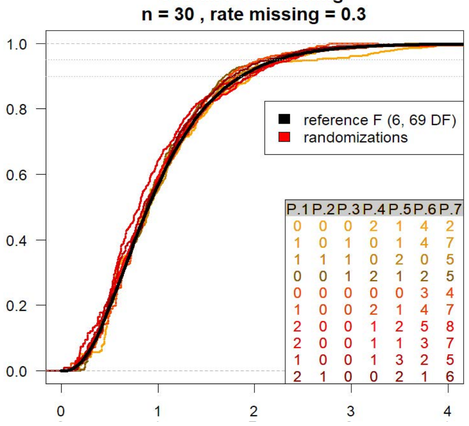

Fig./Vid. 6: empirical and parametric distribution for ANOVA-based *F* with incomplete data

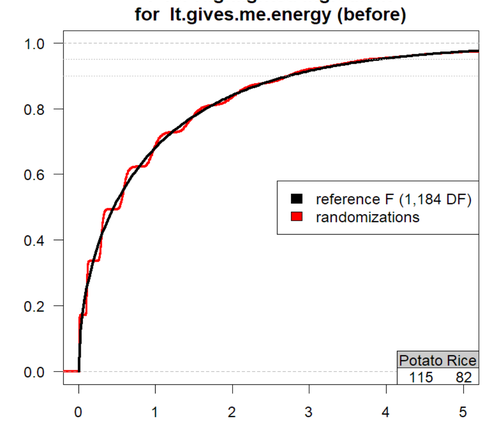

Fig./Vid. 7: empirical and parametric distribution for ANOVA-based *F* in a factorial design

Results incomplete data (see Figure / Video 6)

- *F*-test ok with n=100 and 30% missing; *t*-test borderline even with 20% missing
- With n=50, even with 30% missing values neither appears robust enough
- Note that this scenario is likely to generalize to incomplete block designs

Results factorial design (see Figure / Video 7)

- Sufficient number of elicitations per factor combination required for validity of *F*-tests for main effects as well as interactions (average ≥10)
- Smaller numbers of elicitation substantial violations of nominal level may occur

Conclusions

- **With reasonable sample sizes and number of elicitations (recommend average: ≥ 20) ANOVA provides valid tests**
- **Cochran's *Q* gives similar p values and is therefore likely valid only under similar restrictions**
- **Pairwise comparisons ideally via exact binomial test, unless not applicable in more complex designs**

References

[1] Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37*, 256–266.
[2] Meyners, M., Castura, J. C., & Carr, B. T. (2013). Existing and new approaches for the analysis of CATA data. Food Quality and Preference, 30, 309–319.
[3] Meyners, M., & Hasted, A. (2020). On the applicability of ANOVA models for CATA data. Submitted to FQAP.