Statistical Methods for USP Chapters

STEVEN WALFISH, MS MBA

Disclaimer

- This talk is neither endorsed by or approved by USP.
- All material is public domain and/or contained in USP-NF

Agenda

- <1010> - ANALYTICAL DATA—INTERPRETATION AND TREATMENT
  - Accuracy and Precision
  - TOST
- <1210> - STATISTICAL TOOLS FOR PROCEDURE VALIDATION
  - Combined Estimate of Precision and Accuracy
- <1220> - ANALYTICAL PROCEDURE LIFE CYCLE
  - Replication Strategy
- Bioassay Chapters
  - Relative Potency Estimates

Accuracy and Precision

- Accuracy is determined by the difference (d) that two procedures can differ and still be considered equivalent. This is a practical difference.

$$\left|\mu_D\right| = \left|\mu_A - \mu_B\right|$$

$$H_0 : \left|\mu_D\right| \geq \mathrm{d}$$

$$H_a : \left|\mu_D\right| < \mathrm{d}$$

Accuracy and Precision

- The goal of comparing precision of two procedures is to have a maximum ratio (k) that is practically important.

$$H_0: \frac{\sigma_A}{\sigma_B} \geq k$$

$$H_a: \frac{\sigma_A}{\sigma_B} < k$$

Example

- Specification is 45 – 55
- Current method has analytical variation of 0.75. No process variation.
- Assume d = 2 and k=2
- Prob(OOS) = 1 – Prob(45 ≤ Sample process value ≤ 55)

- $1 - \Phi\left(\frac{55-(50+d)}{\sqrt{.75*k^2}}\right) - \Phi\left(\frac{45-(50+d)}{\sqrt{.75*k^2}}\right)$
- $1 - \Phi\left(\frac{55-(50+2)}{\sqrt{.75*2^2}}\right) - \Phi\left(\frac{45-(50+2)}{\sqrt{.75*2^2}}\right)$
- 1-(0.958)-(0.0000266)=4.16%

TOST

- Most statistical tests are performed to show treatment differences.
- The null hypothesis is rejected when the test statistic is sufficiently large (usually p-value < 0.05)
- Often, we want to show comparability, especially between test methods or critical reagent lots.
- Equivalence testing (and non-inferiority testing) reverse the hypothesis where rejecting the null hypothesis shows equivalence.

Hypothesis Tests

- Traditional Difference Testing
  - H_o: $\mu_1-\mu_2 = 0$
  - H_1: $\mu_1-\mu_2 \neq 0$
- Equivalence Testing (TOST)
  - H_o: $|\mu_1-\mu_2| \geq d$
  - H_1: $|\mu_1-\mu_2| < d$
  - Where d is the equivalence margin or the maximum difference that can be considered equivalent.

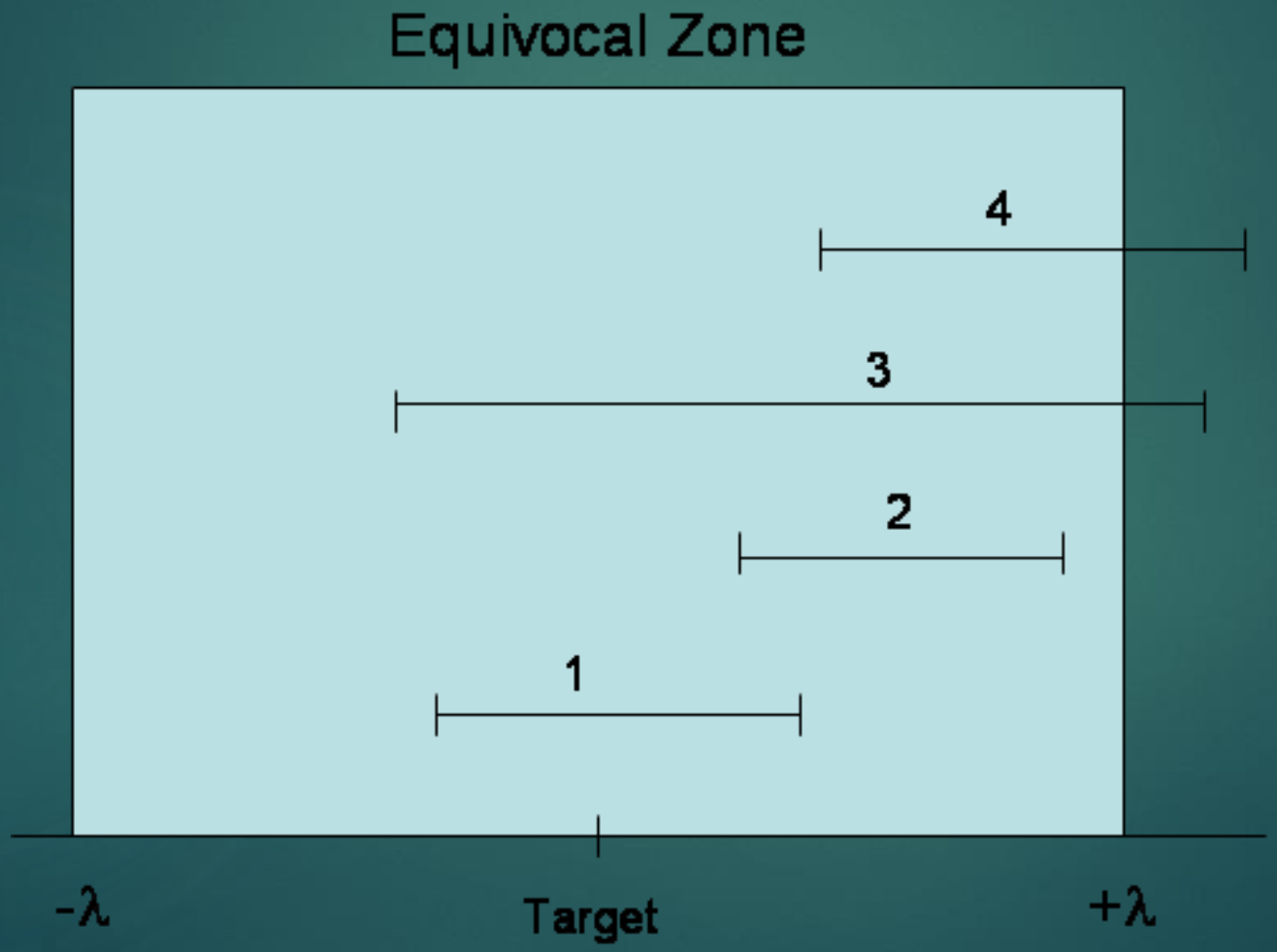
Superiority or Difference Testing

- The traditional difference test is the t-test.
- Failure to reject the null hypothesis does not imply equivalence.
- These tests are highly dependent on sample size (power) and variability
  - Small sample size can lead to failure to reject the null hypothesis, while too large a sample size can lead to rejecting the null hypothesis when there is no practical differences.

Equivalence and Non-Inferiority

- Applying equivalence test warrants that consumer risk (accept a poor assay) to be bounded by a pre-selected number *a*, e.g., 5%.
- Predetermine interval of "sufficiently similar"
- Calculate a 90% confidence interval for the measure of dissimilarity
- If the entire confidence interval falls in the similarity interval, then conclude equivalent; conclude unable to conclude equivalent
- The 90% confidence interval corresponds to a 5% false positive rate for the equivalence hypothesis

Comparison of Significance and Equivalence Tests



Combined Estimate

- The previous approach gives the upper bound on the probability of out of specification based on selecting a k and d value.
- Instead of selecting pre-specified values, you can determine the probability of out of specification based on the accuracy and precision data obtained from the validation study.
- If a computed interval falls completely inside the range from $(-\lambda + \tau)$ to $(\lambda + \tau)$, the procedure is validated for both accuracy and precision.
- Combining accuracy and precision we can compute an interval based on the inference we desire
  - A prediction interval (also referred to as an expectation tolerance interval) is used to validate the probability that the next reportable value falls in the range.
  - A tolerance interval (also referred to as a content tolerance interval) is used to validate the proportion of all future reportable values falling inside the range.

Combined Estimate Example

- An API has a specification of 98-102
- If the tolerance or prediction interval fits inside the specification, then the assay is sufficiently accurate and precise.
- If we use the prediction interval, the assay shows sufficient accuracy and precision
- This is not the case for the tolerance interval.

1	99.50
2	99.70
3	100.40
4	101.20
5	98.90
6	99.40
7	99.60
8	100.40
9	99.10
Mean	99.80
SD	0.73
PI	98.37
	101.23
TI	97.88
	101.72

Replication Strategy

- In general, replicating the aspects of the procedure that contribute to a large part of the overall variation followed by an appropriate analysis to generate the reportable value will lead to improved precision.
- Estimating the amount of variation associated with each step of the procedure during the procedure design phase can allow for determination of an optimal sample and standard replication strategy, which may enable the procedure to more easily meet the criteria of the ATP for the reportable value.
- Consideration should be given to including the acceptable variability among the individual results for the factors (e.g. sample preparation, sample injection etc.) being replicated in system suitability criteria

Replication Strategy

- Risk Assessment to acquire understanding of the overall sources of variation contributed by the procedure:
  - Propagation of error models and
  - Measurement system analysis experiments
- Sources of variation may be grouped into general categories such as:
  - run-to-run (between run) variations
  - within-run variations
- Those sources could also be broken down into component categories such as lab-to-lab, instrument-to-instrument, sample preparation, and sample introduction.
- If the amount of variation from these sources can be quantified, it may allow for the use of an optimized sample and standard replication strategy.

Statistical Models

- Relative Potency:

  From "Statistical Method in Biological Assay" by D.J. Finney (Second Edition-1971), page 61

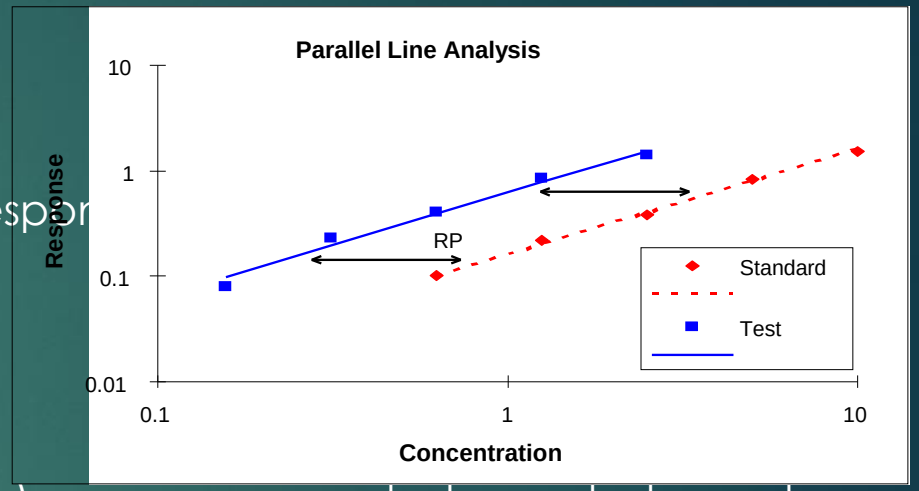
  - $F_T(z)=F_S(\rho z)$ where
  - Z = dose
  - F_T = dose response regression function for the test
  - F_S = dose response regression function for the standard
  - ρ = potency of the test relative to the standard

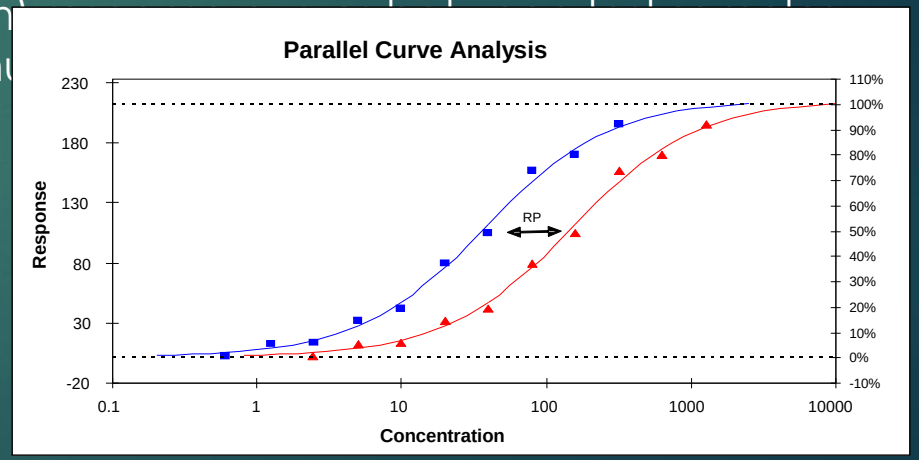
- This is a statement of the condition of "similarity":
  - Test function is a translation of the standard

Relative Potency Models

After demonstration of similarity, widely used parallel models for *in vitro* bioassay:

- Parallel line design (linear)
  - Horizontal distance in parallel linear log(concentration) respor the standard
- Parallel curve design (full curve)
  - Horizontal distance in parallel nonlinear log(concentration and the standard (slopes, upper and lower asymptotes mu





When To Use Each Model

- Parallel Line Assay
  - Linear response (or linear after transformation). Can also be used for the linear portion of a non-linear response.
- Parallel Curve (4-PL)
  - Non-linear response. Typically used for assays with asymptotes at low and high concentrations.

4-PL Example

- Constrained Model

$$Y = Z * \left(A + \frac{B - A}{1 + \left(\frac{conc}{xmid_S} \right)^D} \right) + (1 - Z) * \left(A + \frac{B - A}{1 + \left(\frac{conc}{xmid_R} \right)^D} \right)$$

  - Test of Parallelism (from unconstrained model)
  - Ratio of slopes
  - Asymptote ratio
  - Set acceptance criteria based on “impact of non-parallelism”

4-PL Example

Sample	Lower	Upper	Slope	EC50	
Reference	2035	61619	2.13	18.84	
50% RP	1827	60972	2.06	34.78	
Constrained	1956	61236	2.10	19.03	34.48
Relative Potency				55.2%	

- Ratio of slopes = 1.03
- Ratio of Lower = 1.11
- Ratio of Upper = 1.01

Acknowledging…

- Thanks to
  - You, for your attention and interest
  - USP, for supporting efforts to of developing standards with the best scientific practices