

# Statistical Methods for Setting Acceptance Criteria on Parallelism in Bioassays

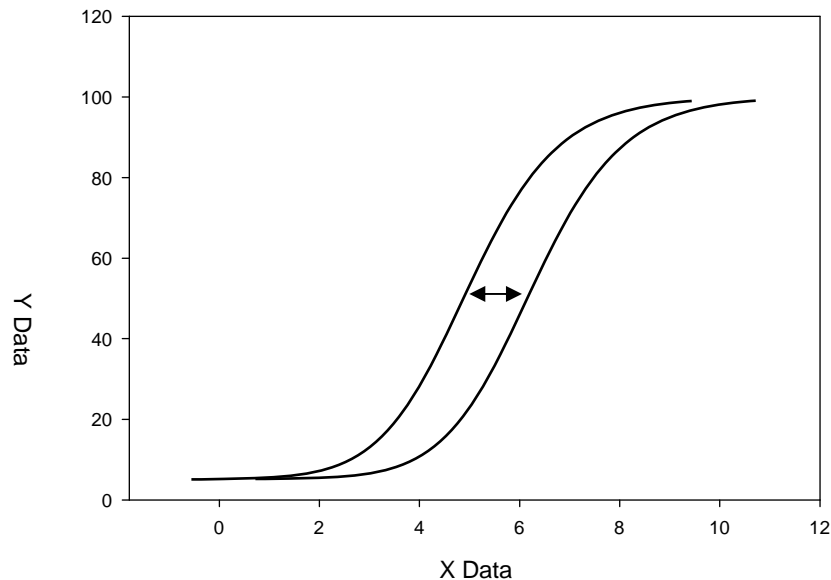
Daniel Joelsson

# Parallelism Acceptance Criteria

- Setting acceptance criteria on assay parameters is critical for proper interpretation of the results
  - Bioassays measure both potency and stability of samples
- Current methods for setting criteria on parallelism
  - Rely heavily on the expertise of an experienced biometrician
  - Often involve judgment calls on what data to include in an analysis
- New method using bootstrapping
  - Involves less subjectivity
  - Rugged and statistically defensible criteria

# Potency Assays and Parallelism

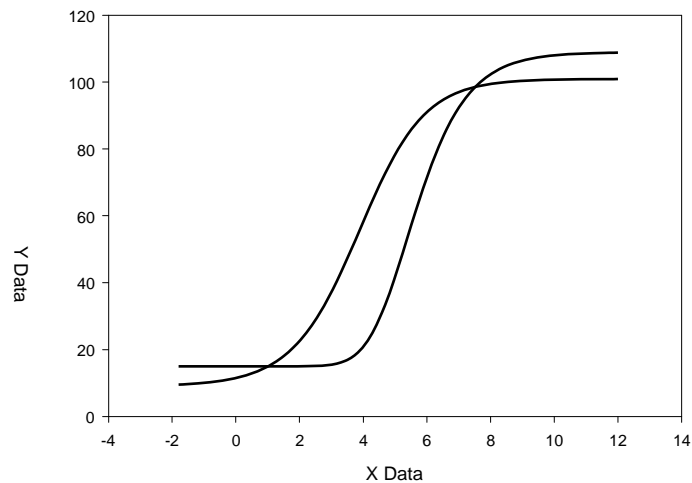
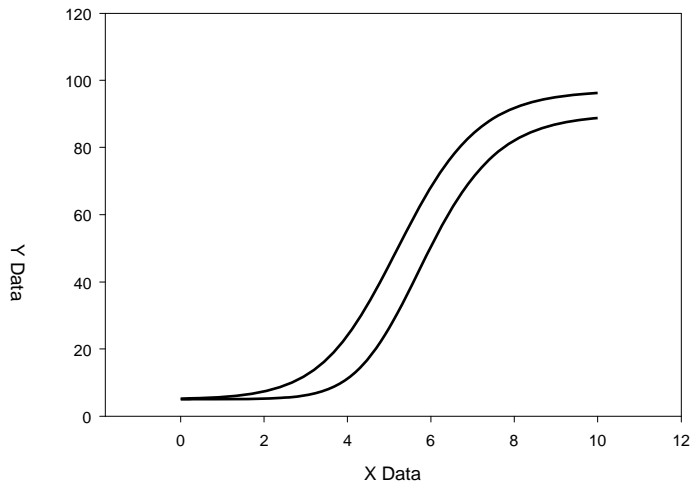
Potency is a shift in the X-axis compared to a reference standard



- Strictly speaking, curves *must* be the same exact shape for potency to have meaning
- This is usually called parallelism, but more accurately known as *similarity*

# Parallelism in the real world

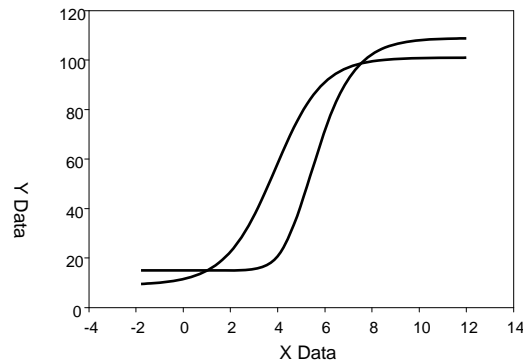
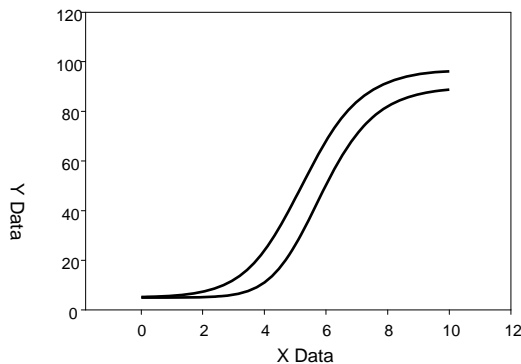
- Bioassays have high variability
- Replication commonly performed to control the mean
- Curves are samples from all possible curves



Are these curves parallel?  
How do we know?

# Measurements of parallelism

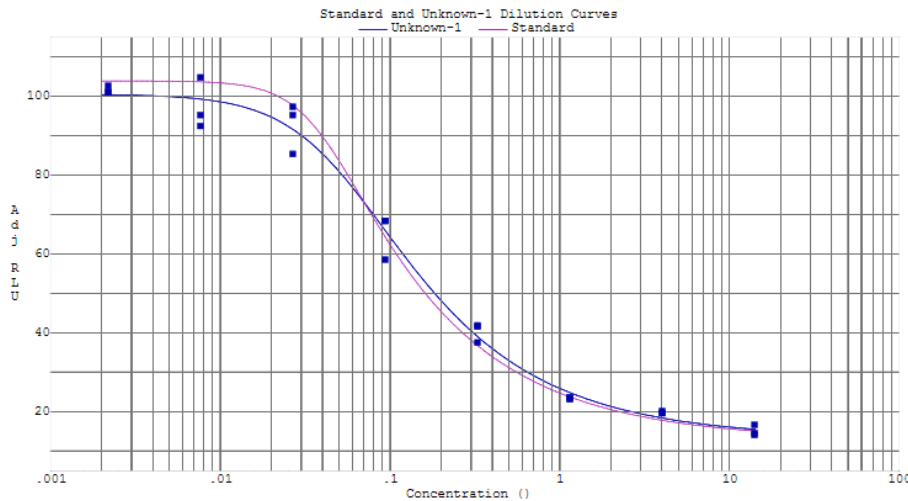
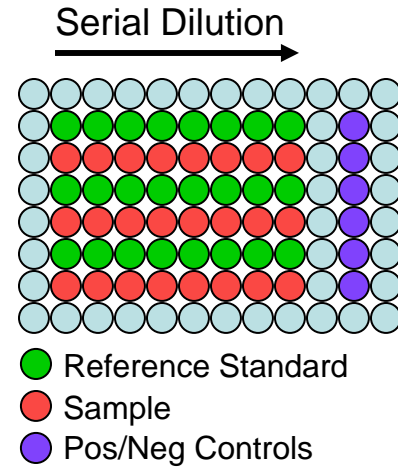
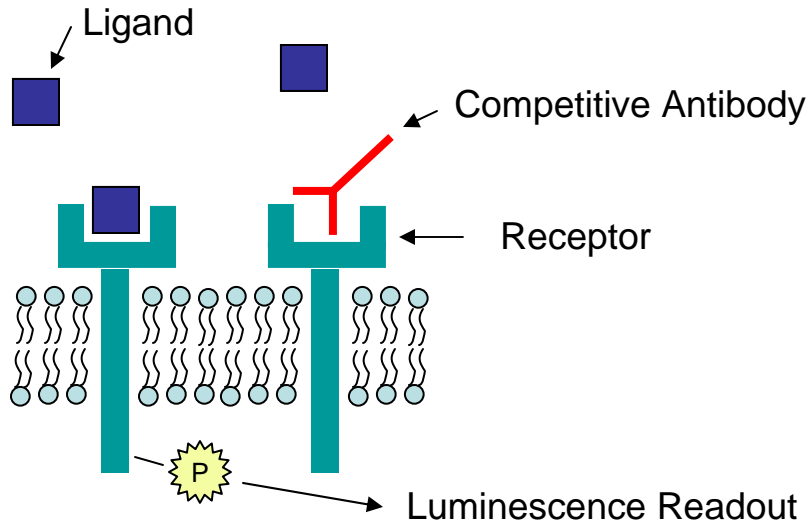
- How to measure parallelism?
- Several different methods available:
  - Slope ratios
  - Hypothesis testing
    - F test and Chi-Square test
  - Equivalence testing
- Each generates a ‘metric’ for parallelism
  - How parallel are these two curves?
- What is the cutoff?
  - How far off do the curves have to be before they are no longer parallel?



# Acceptance Criteria

- How do we set the cutoff?
- Ideally set by clinical data
  - Not realistic in early development
- Assay capability cutoffs
  - Based on 95% or 99% limits
  - How do we set these?
  - Run 2000 or so assays with the standard compared to another replicate of the standard.
    - “Parallel” by default
    - Distribution of metric due solely to assay variability

# Bioassay for a monoclonal antibody project

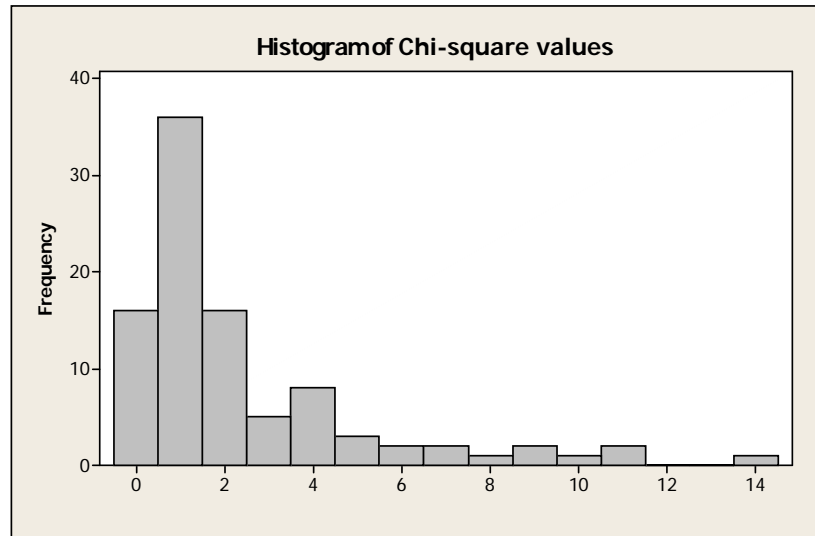


StatLIA:

Potency Calculation

Chi-Square parallelism metric

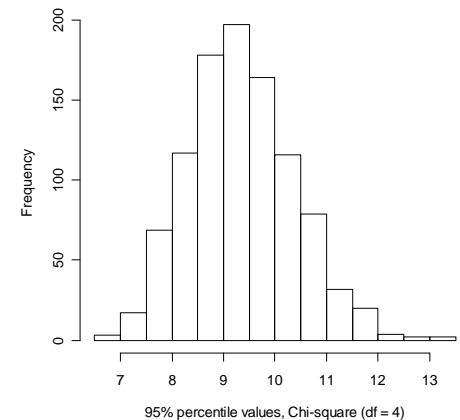
# Traditional Method: Actual data from 95 runs



‘Traditional Method’ has problems:

- Some of these runs are probably NOT parallel
- Data ‘massaging’
  - Eliminated outliers
- Low sample number
  - 95% cutoff based on only ~ 4 runs →

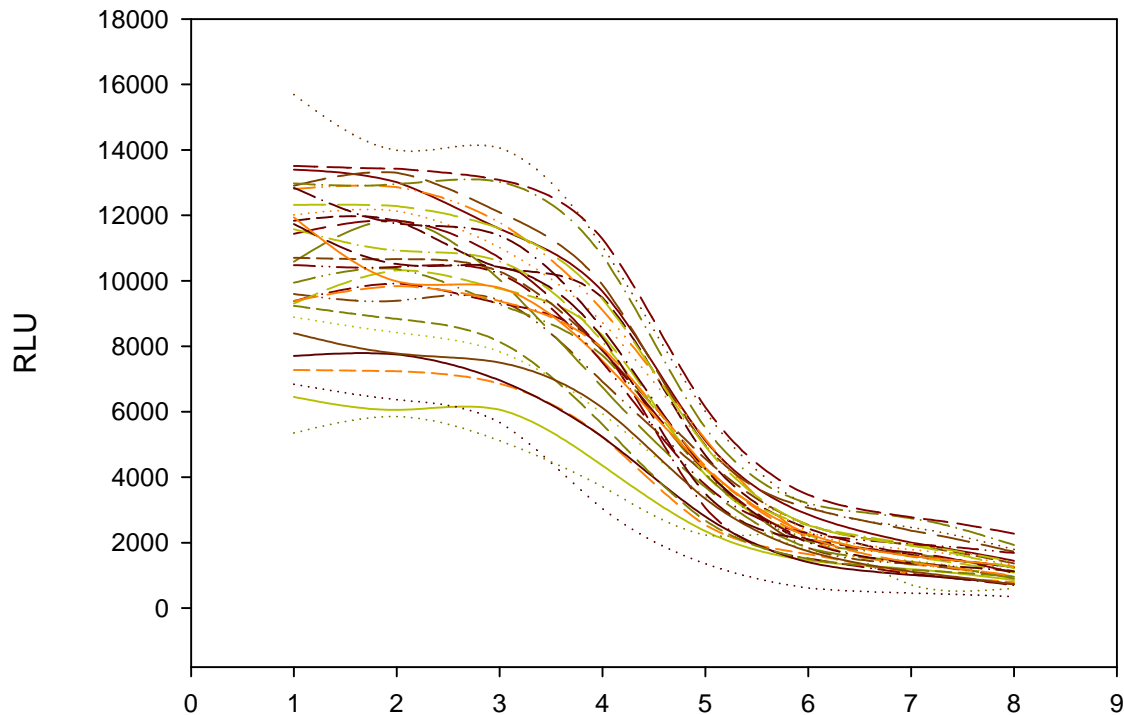
End results = 95% cutoff at 7.28



Simulated data with 100 runs.  
Large variability in 95% cutoff

# Standard Curves

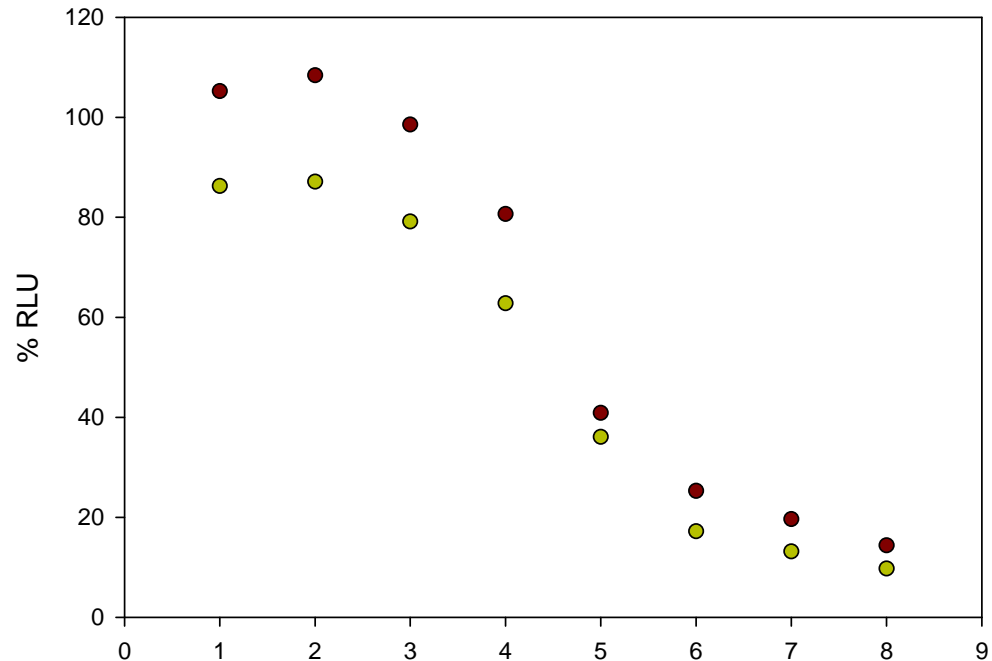
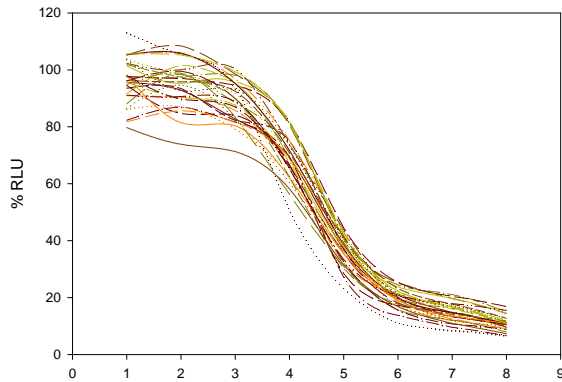
- Can we use the data we already have?
- Many replicates of the standard
  - StatLIA reference set
  - 30 representative assays
    - Why 30?





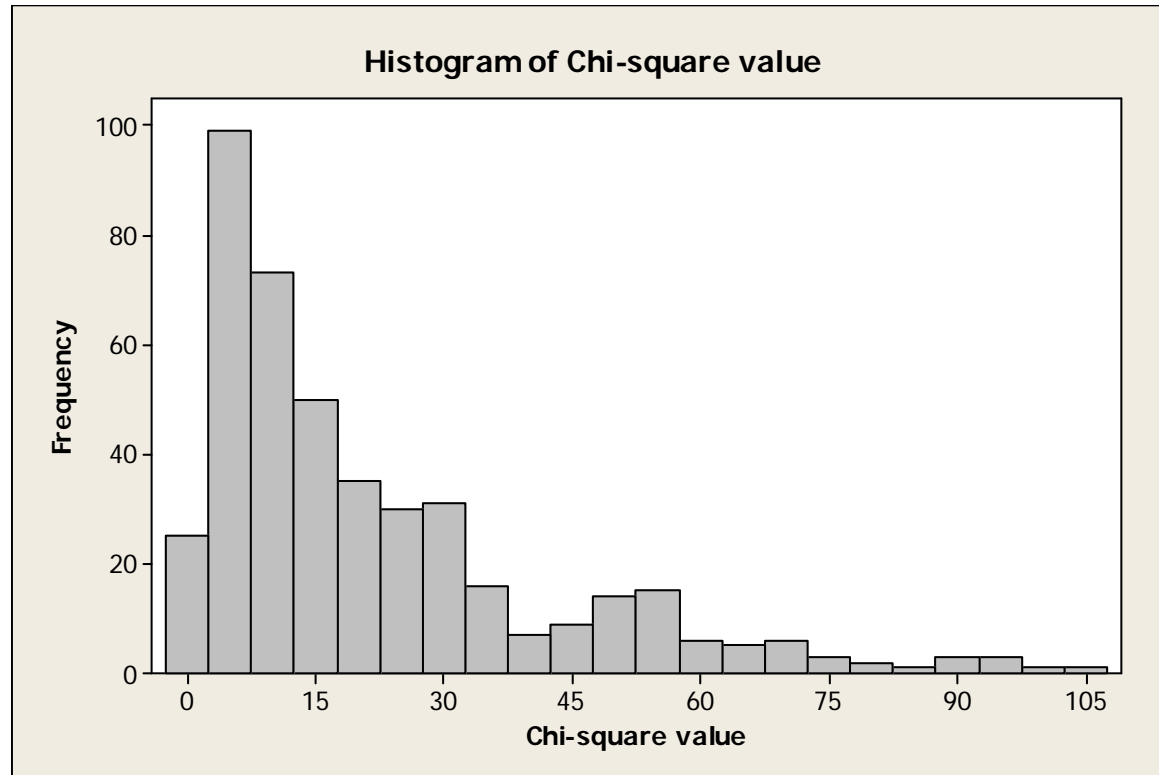
# Combining standards

Sample 2 curves



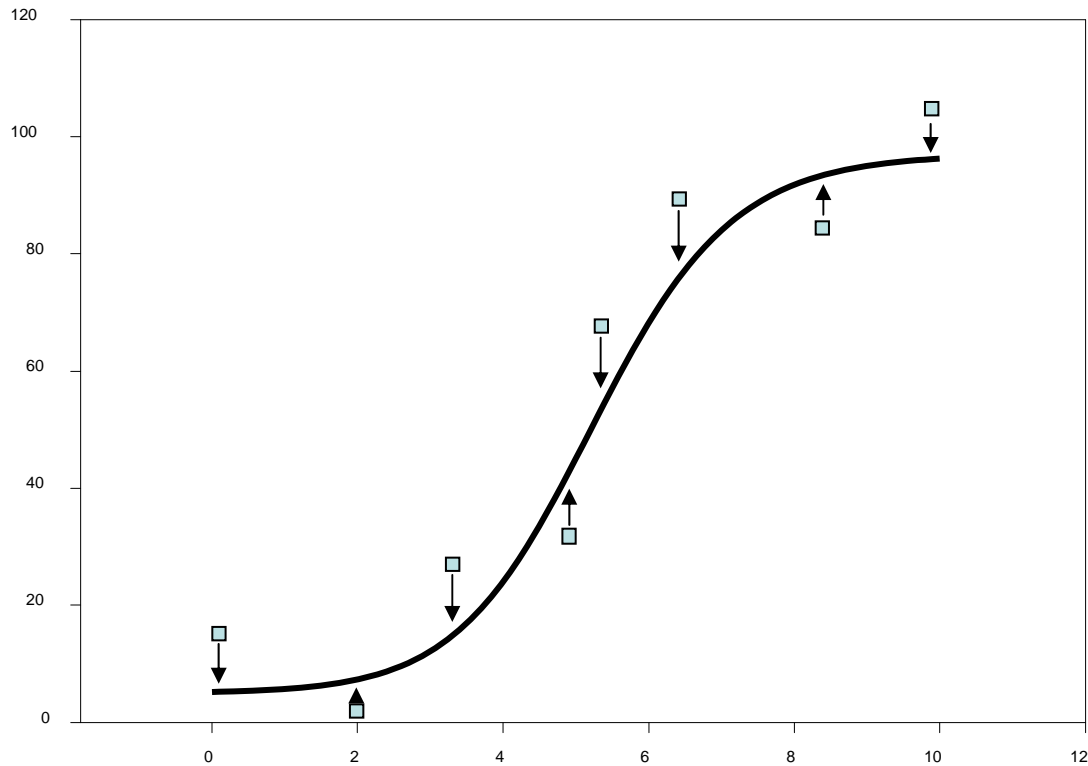
- Use StatLIA to generate chi-square value
- 435 possible combinations of curves

# Results of combinations



- Not good – bimodal distribution?
- 95% cutoff at ~62

# Residuals

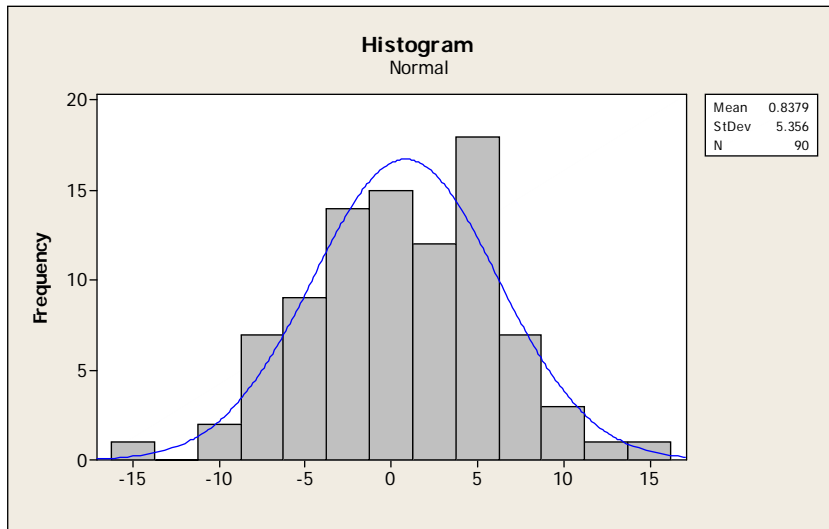


Residual = The distance from each data point to the fitted curve

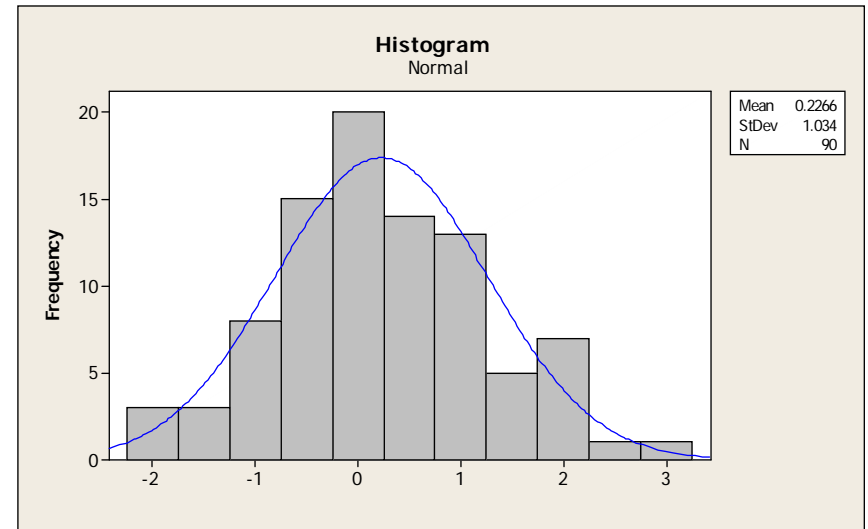
Due to the underlying variability of the assay

# Distribution of residuals

High end of the curve



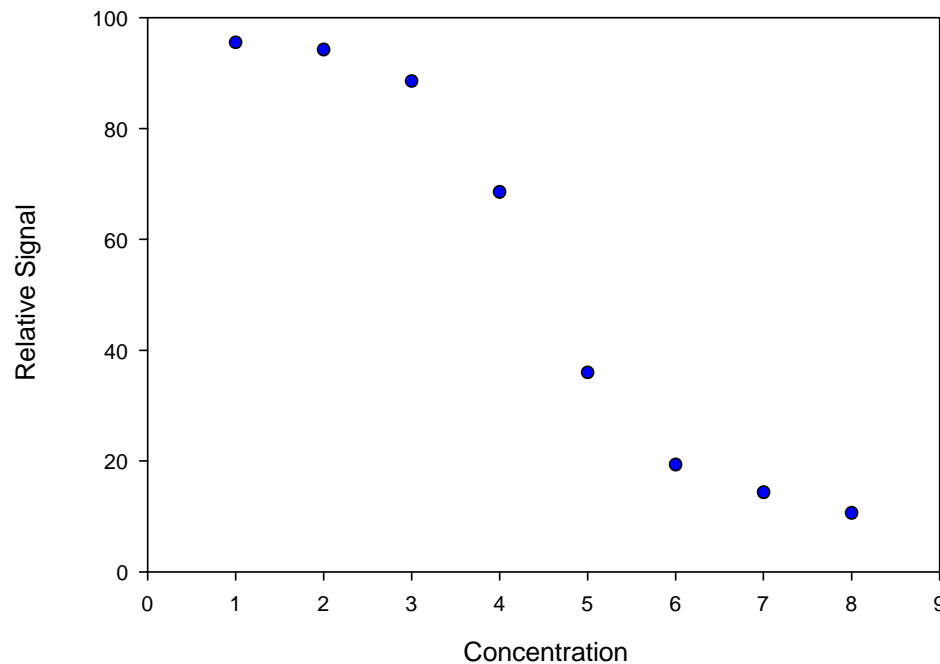
Low end of the curve



- Close to normally distributed
- Residuals larger at higher signal levels
  - Need to do weighted regression

# Bootstrap strategy

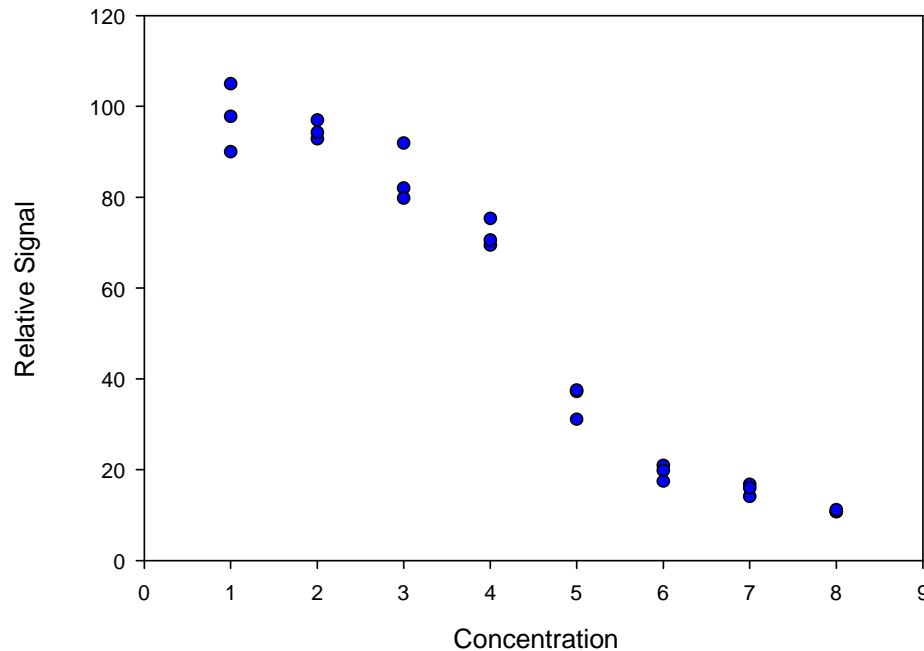
Take the mean at each concentration of the 30 reference curves



Concentration	1	2	3	4	5	6	7	8
Mean	95.52	94.26	88.56	68.55	35.99	19.37	14.39	10.64

# Bootstrap strategy

Randomly sample (with replacement) 3 of the 90 residuals at each concentration and add to the mean to make a standard curve

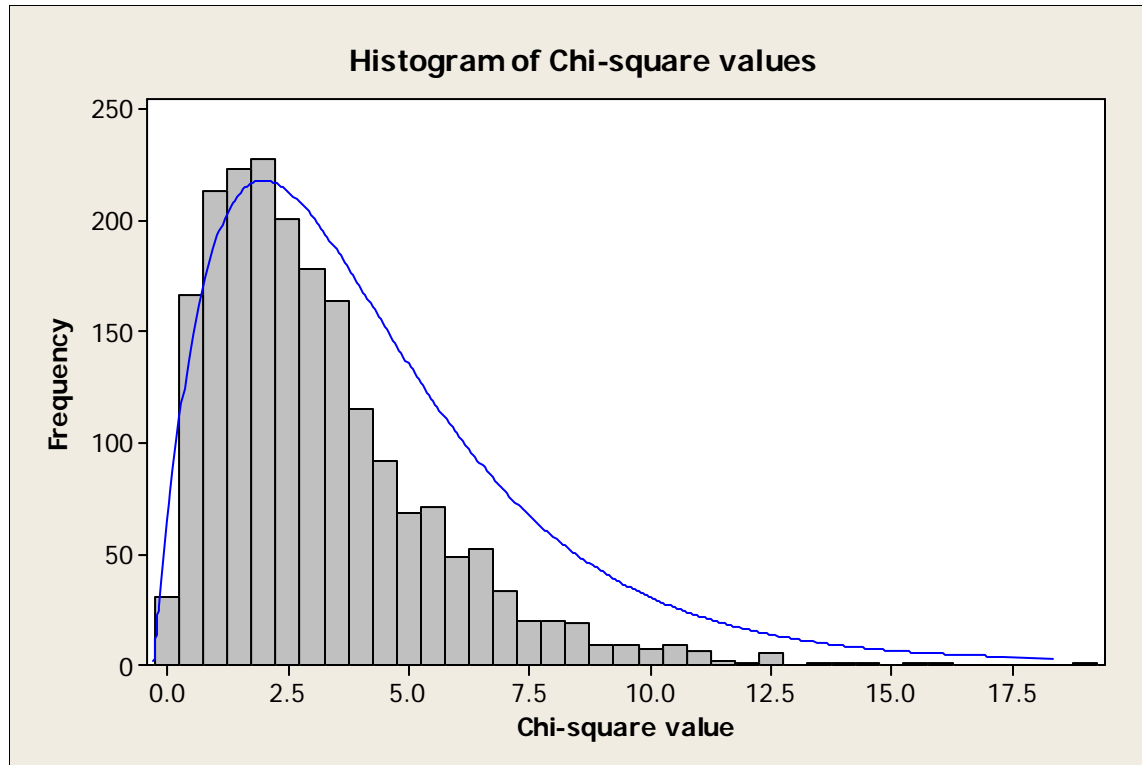


Concentration	1	2	3	4	5	6	7	8
Residual 1	9.46	-1.43	3.38	0.92	1.22	1.59	2.36	0.08
Residual 2	2.28	2.76	-6.56	6.75	1.58	0.50	-0.28	0.13
Residual 3	-5.50	0.02	-8.73	2.04	-4.89	-1.88	1.62	0.54

# Bootstrap strategy

- Using this strategy we can generate  $8 \times 10^{46}$  possible standard curves
- That's  $4 \times 10^{46}$  pairs of curves!
- Each pair generates one chi-square metric
- We generated 2000 of those possible values
  - Same mean function = parallel curves by default
  - Differences are due to assay variability

# Results



(Overlay: Chi-square with 4 df)

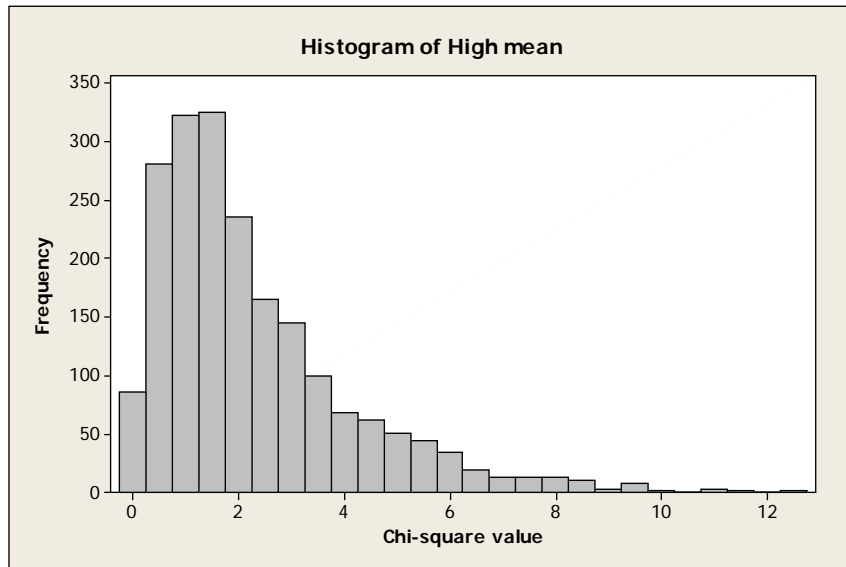
95<sup>th</sup> percentile is at 7.499

\*Estimated cutoff based on 95 runs = 7.28

\*Estimated chi-square distribution cutoff = 9.483

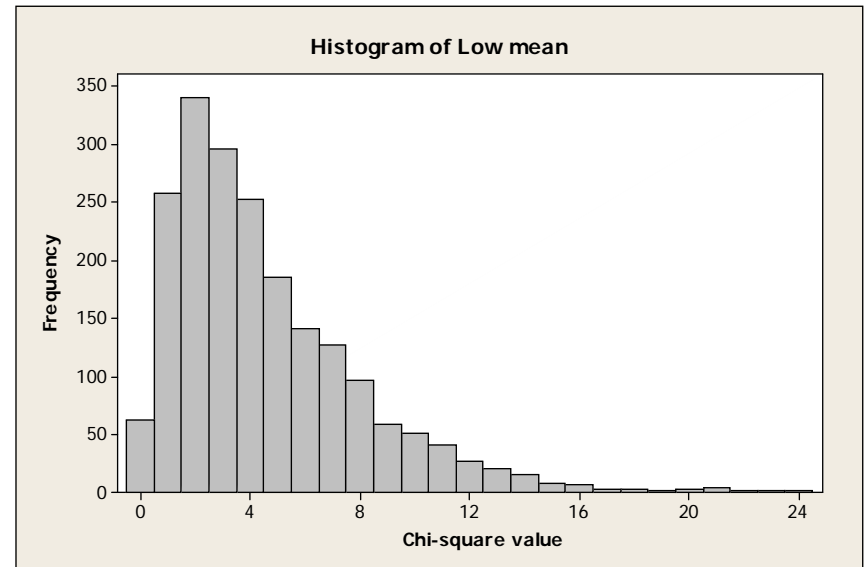
# Is the mean appropriate?

“High” mean



95% cutoff = 5.932

“Low” mean

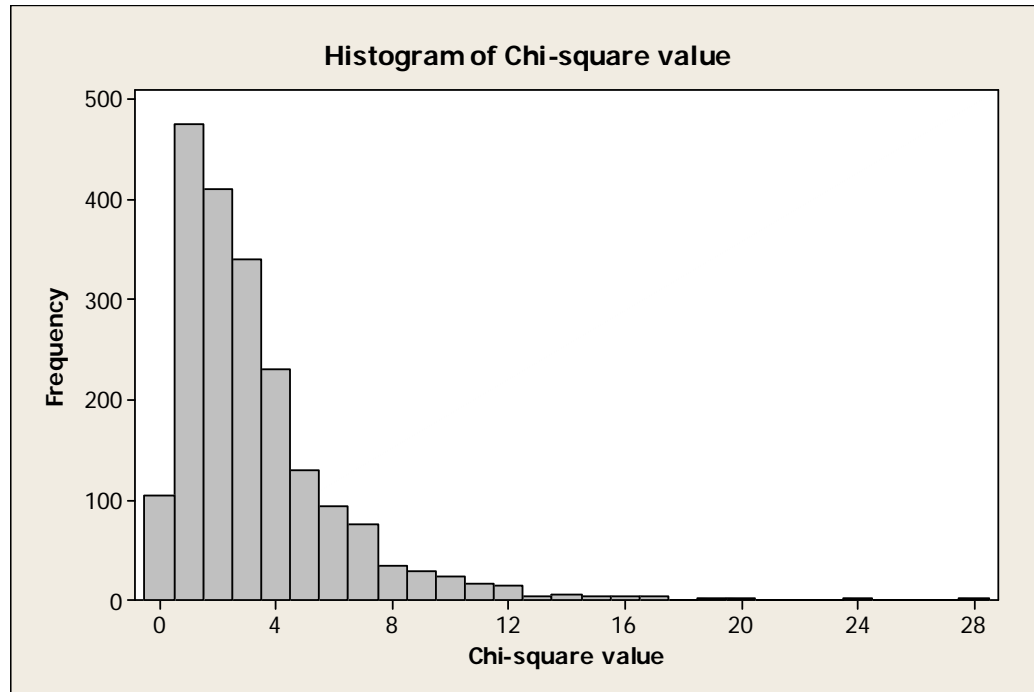


95% cutoff = 11.21

Wider distribution with a low mean function.  
Weighting plays a role.

# Sampling the mean function

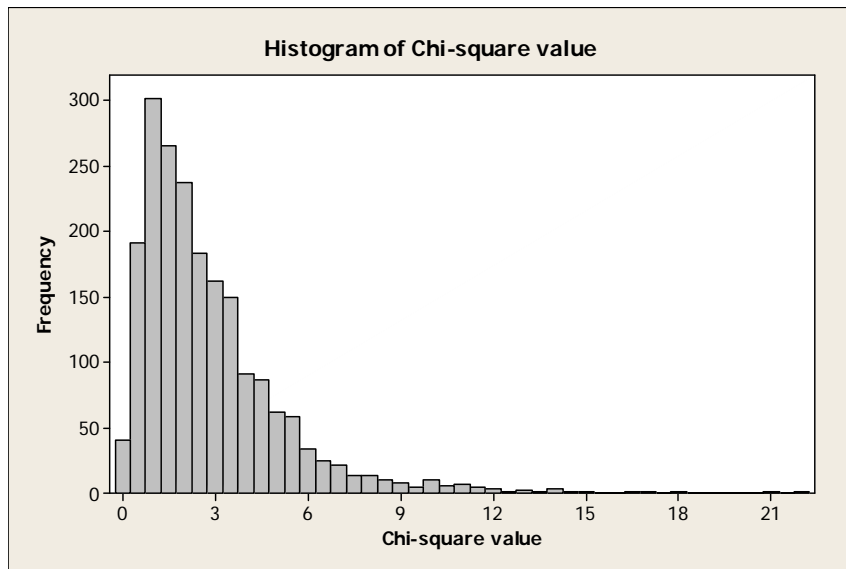
Added in an extra step – a new mean for each pair of curves  
Randomly selected one of the 30 original curves as the mean



Distribution shifted slightly to the right  
95<sup>th</sup> percentile moved to 8.701 from 7.499

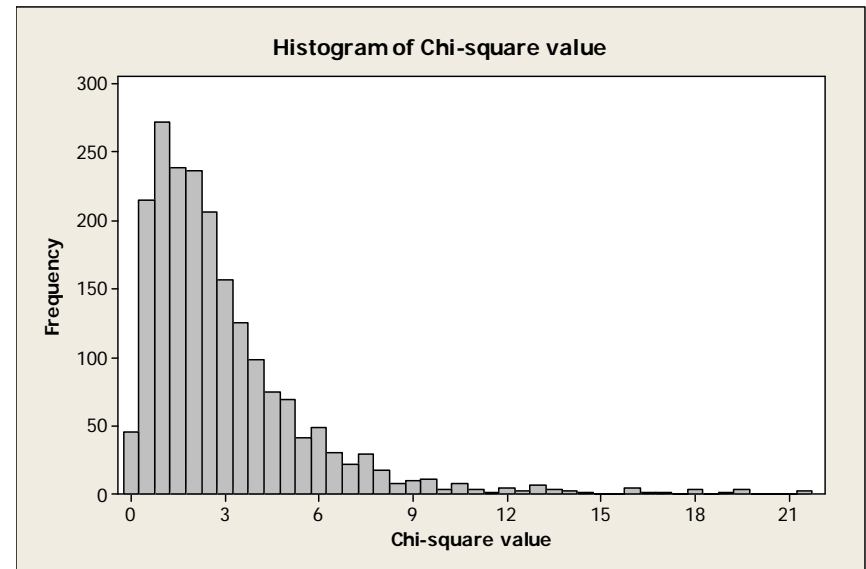
# Can we use less information?

## Bootstrapping 20 curves



95% cutoff = 7.017

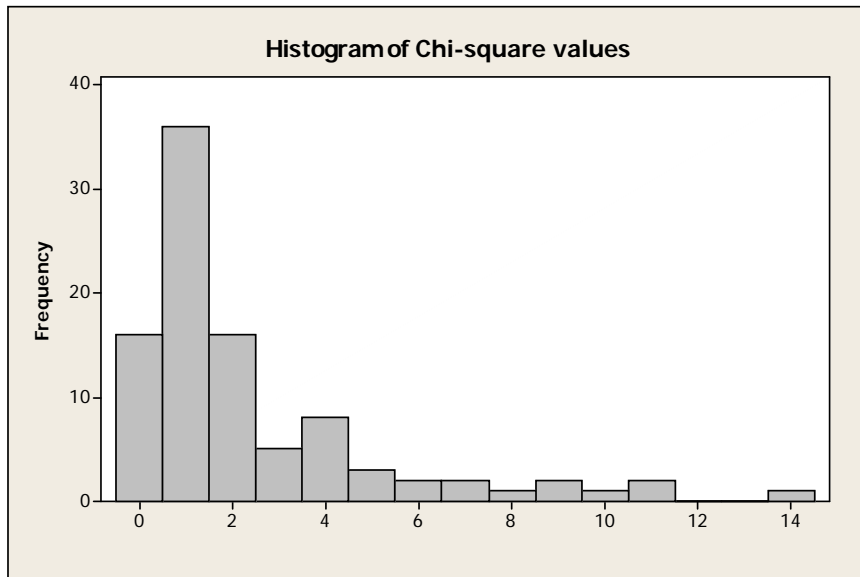
## Bootstrapping 10 curves



95% cutoff = 7.683

# Summary

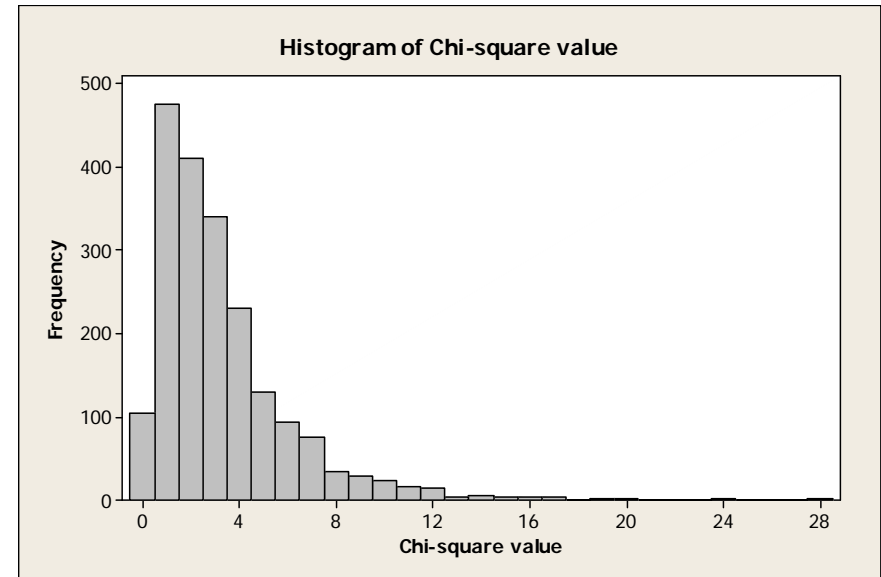
95 actual runs of the assay



95% cutoff = 7.28

Questionable data in the tail of the distribution.

2000 bootstrapped runs



95% cutoff = 8.701

Consequence – would fail 7.7% of assays instead of 5%

# Conclusions

Bootstrapping to set a parallelism acceptance criterion has several advantages over traditional methods:

- Eliminates the need for multiple assays comparing the reference standard to itself
- Eliminates the subjectivity involved with selecting the appropriate runs to include in the analysis
- Based on empirically observed data
  - no simulation necessary
- Can be used to set acceptance criteria very early in development (10 curves) and refined as more data is available
- Repeatable and defensible