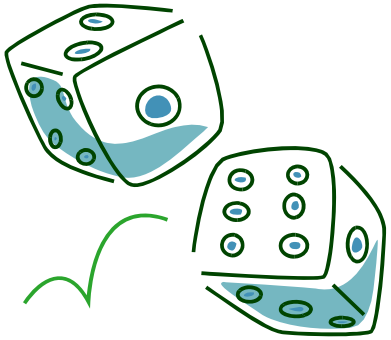
Cutoff Studies with Visually Read In Vitro Qualitative Devices



Robert Sarber, PhD.

Operations Statistician

Sekisui Diagnostics

Introduction

- Cutoff studies are required for visually interpreted qualitative assays as part of CLIA waiver and 510(k) submissions.
- While there is useful information that can be gathered from these studies, the specific performance goals required impose conditions that are often difficult to achieve in practice.
- Forces a clinical device to perform ideally in an analytical study.

FDA Qualitative Cutoff Studies

- Cutoff studies because they are a part of the CLIA Waiver requirements.
  *"The percent of positive results for the 60 aliquots (**20 replicates at 3 sites**) of the prepared, weak positive samples should be close to 95%." – "Close" is not defined.*
- FDA recently required cutoff information for 510(k) submission
  - *To best characterize the cutoff of the device and performance around the cutoff for all matrices, the sensitivity studies should demonstrate the concentrations in your study at which:*
  - *(1) all samples yielded a positive result*
  - *(2) approximately 95 percent (%) of the test results were positive.*
  - *(3) approximately 95 percent (%) of the test results were negative and*
  - *(4) no positive results were observed.*
  - *For samples with concentrations near the cutoff, we recommend you perform **at least 20 measurements per level**. Overall, you should test a sufficient number of replicates per sample so that results obtained with your device will be statistically meaningful.*
- While the 95% levels are part of the January 2008 CLIA Guidance, levels that are positive or negative 100% of the time have been added for the 510(k) recommendations.

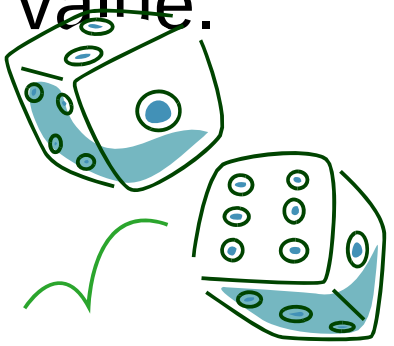
Positive/Negative Probability

- The standard deviation of a qualitative probability is related to the probability.
  - SD = square root (p x (1-p))
- The standard deviation will thus increase as the probability of occurrence decreases from 100% likelihood until ***at 50% the standard deviation is 50%!***
- The standard deviation then decreases as the probability continues to decrease.
- This error makes making and testing a 95% sample difficult.
- For small studies, a nonparametric method is needed as the formula is not valid if there are not 5 observations in each category (p, 1-p).

Making a 95% Sample

- Making the sample is simple – **verifying** it is hard
- If we get a 95% result on testing the sample, the likely range that may contain the true performance is dependent on the number of replicates tested in getting the 95% value.

	Nonparametric 95%CI
n=20	76.4-99.1%
n=60	86.3-98.3 %
n=100	88.8-97.8%
n=500	93.2-96.6%



Simulated Sample Preparation Example

- 100 random replicates at 97%, 95%, 90% and 85% true performance gave:

  99%, 98%, 95%, 83%

- Retesting the 95% level using another 100 random replicates gave a 93% value.
- True sample rate was 90%
- 20 sample performance range will be between 75-100%.

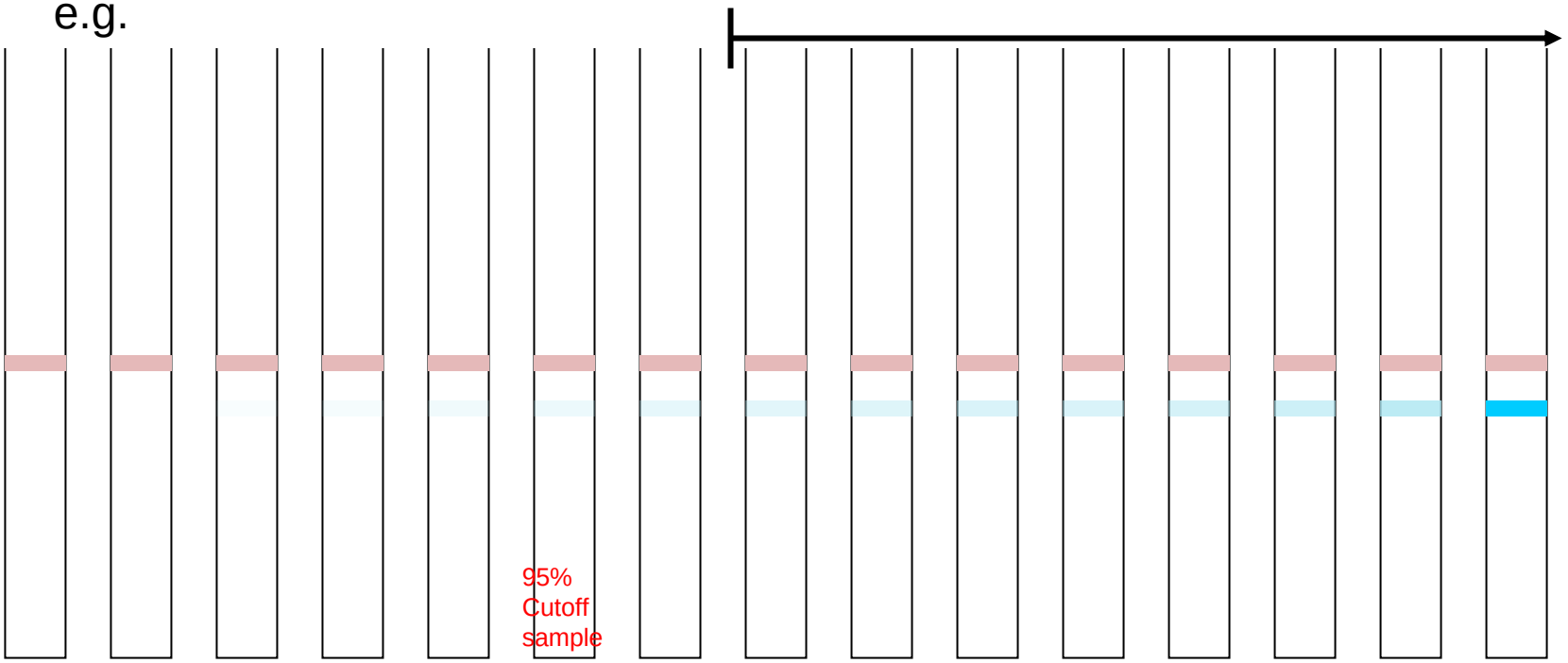
Visual Analytical Readings

- At the April Statistician's Meeting speakers remarked on the problems of reader variability and the problem of getting consistent results from visual interpretations.
- To meet the 95% testing requirements we need to find a level that is hard to see, but is seen equally (un)well by all readers.
- For a particular reader, getting variability in the readings at this level means that something in the reading process has changed, either devices, sample addition, timing or some other condition.
  - **Devices may be the *least* likely cause of variability.**

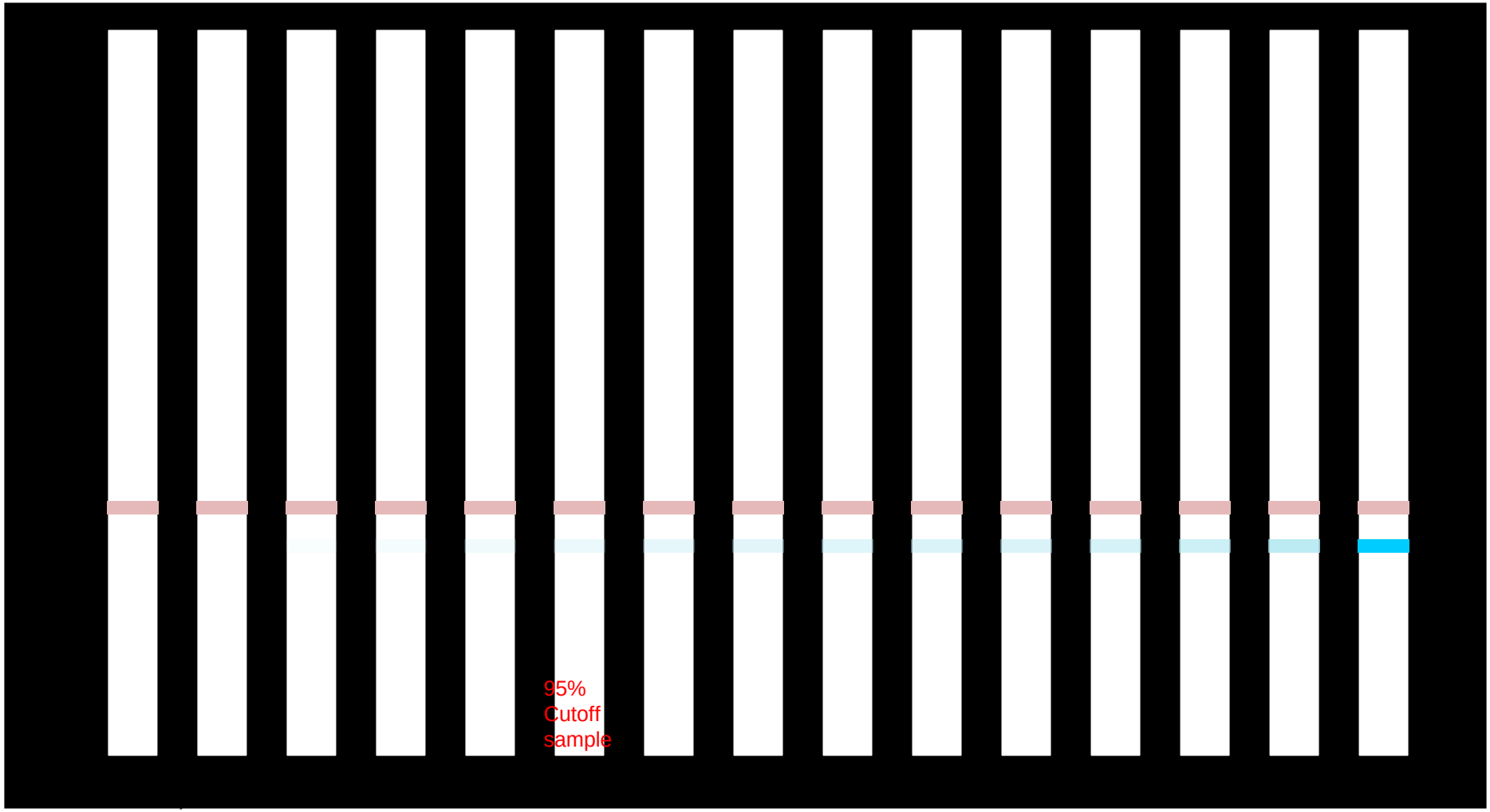
Example of Qualitative Reading Range

- As the analyte concentration falls it becomes harder to see the test line
- The level at which this occurs is dependent on the human reader and may also be affected by the number of replicates being tested – fatigue factor.
- Clinical test levels are usually well above the cutoff.

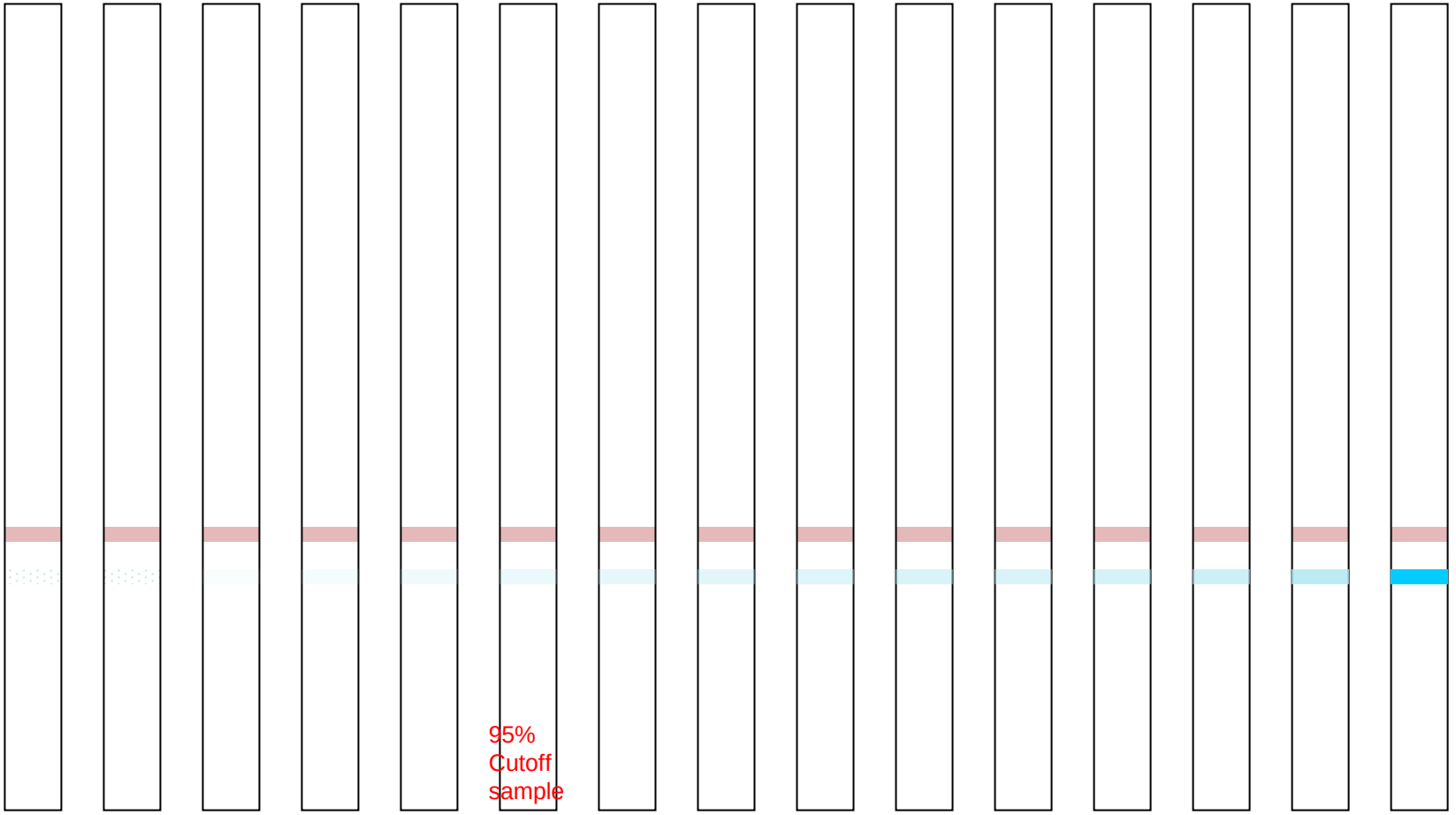
e.g.



Background May Have an Effect

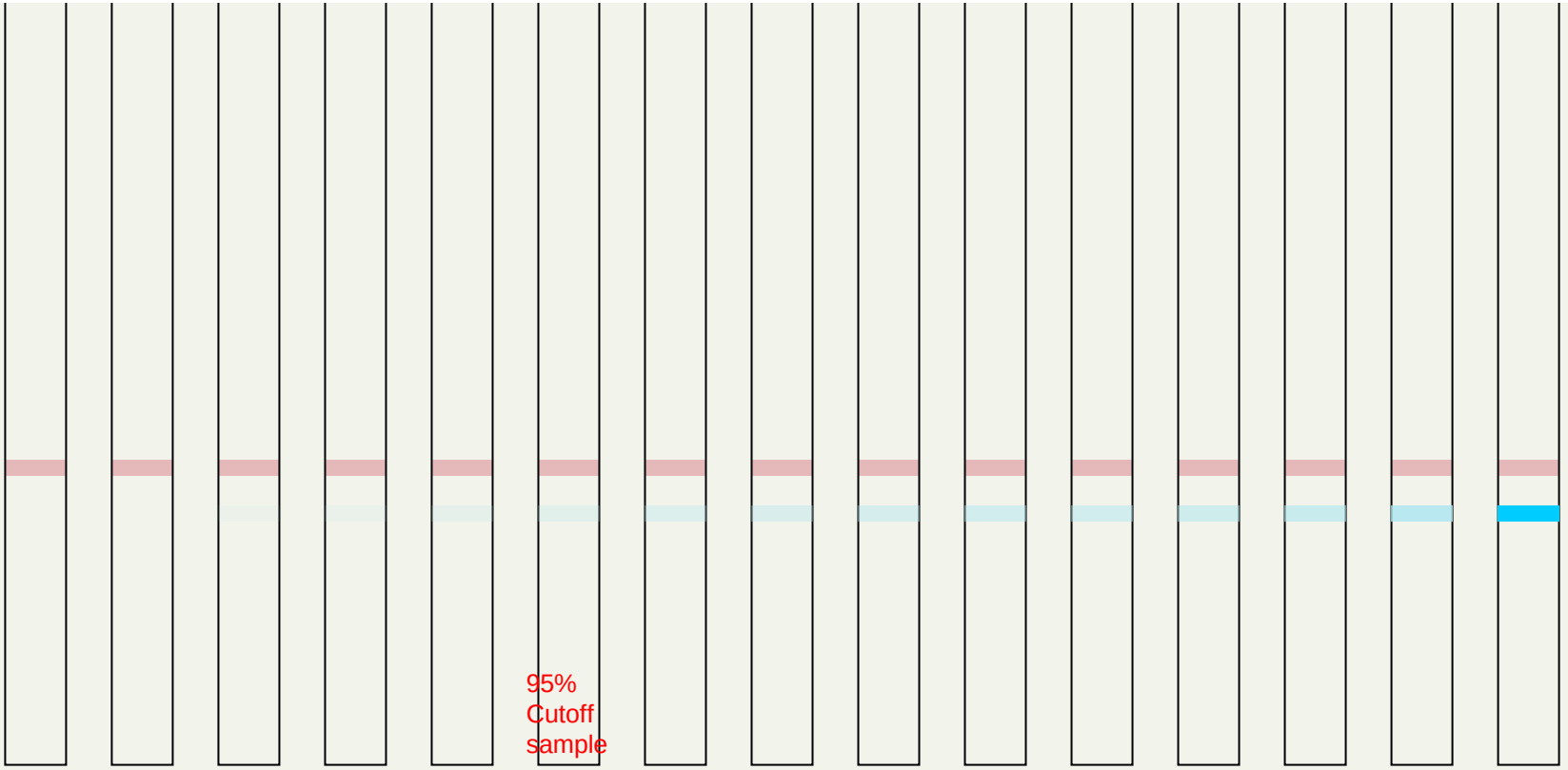


Lighting May Too

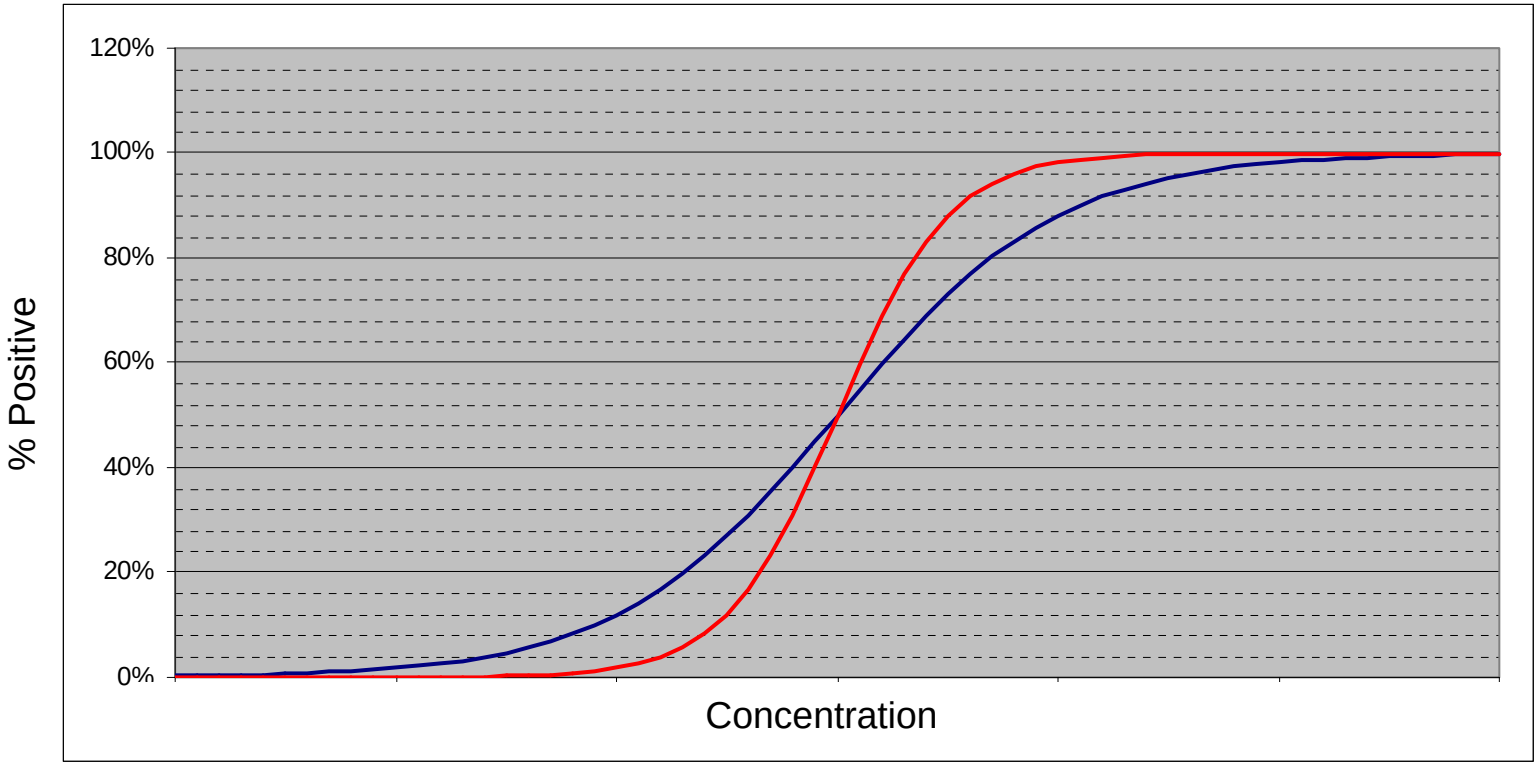


Or Other Issues

- This is for those of us who have had a cataract removed and found things are actually a lot whiter than we thought.

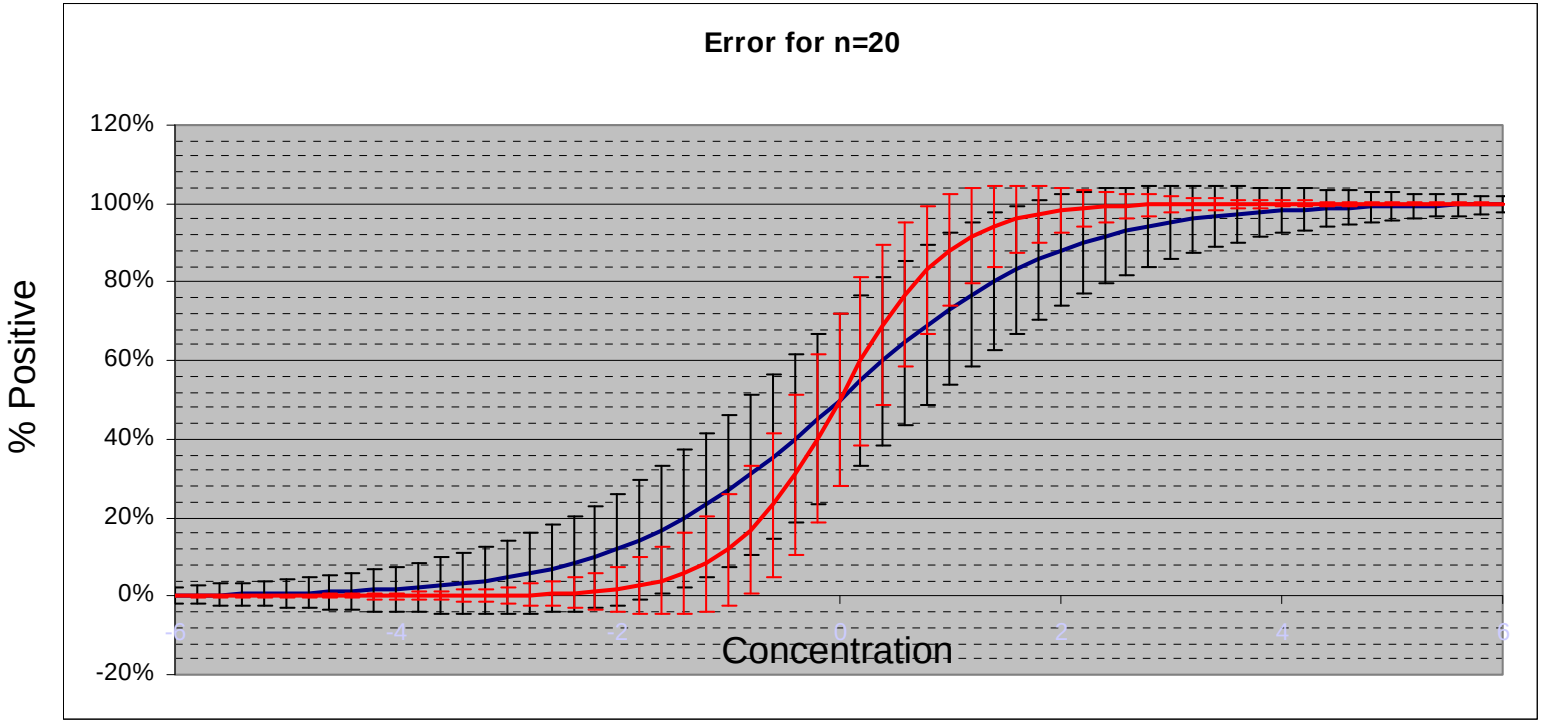


Cutoff is Dependent on the Shape of the Assay Curve



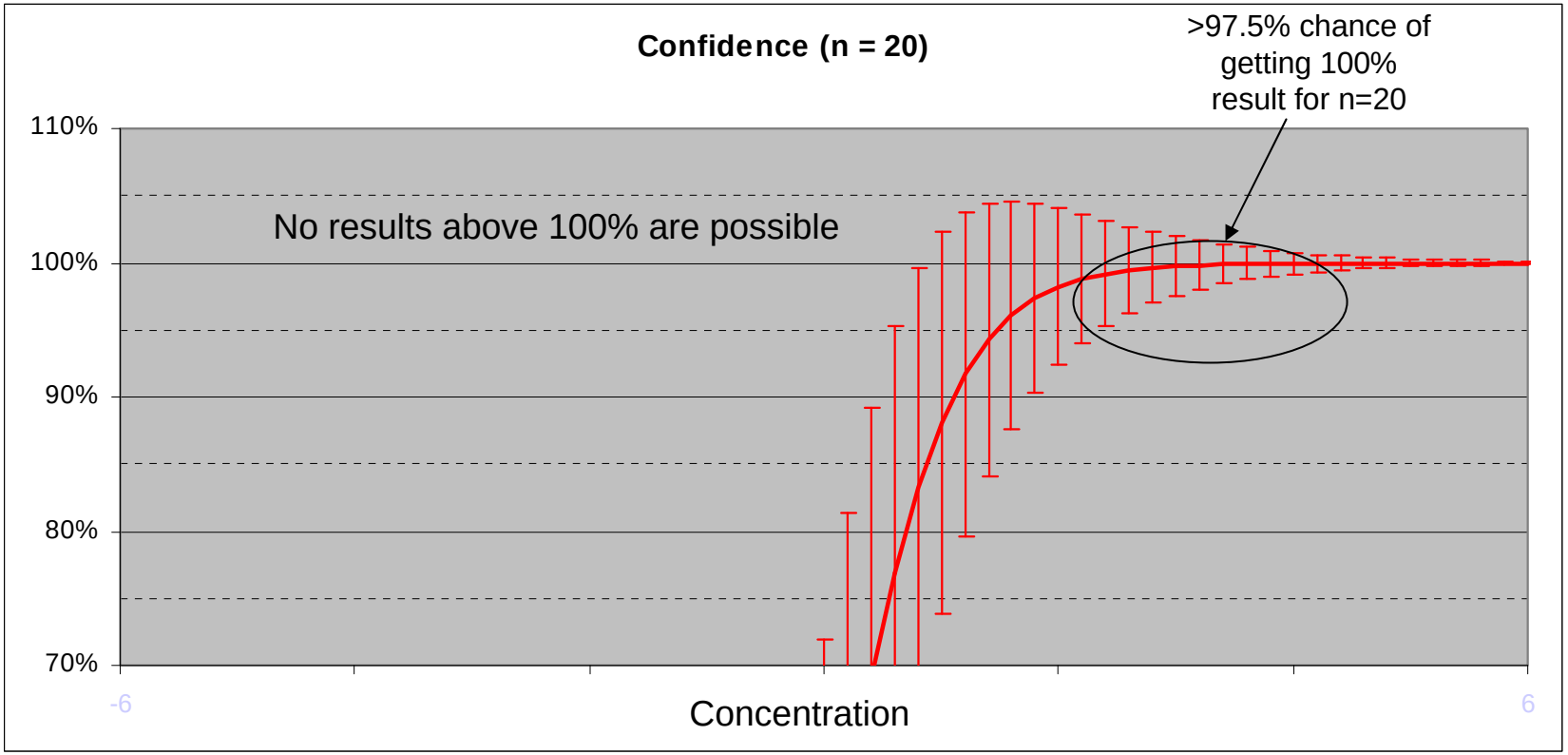
A sharp curve means there is a smaller indeterminate area for the assay but a small concentration change will make a bigger shift in performance therefore it is tougher to prepare the “95%” sample.

Error is Dependent on Probability



Both curves have the same error at the same probability, but a small change in concentration has a bigger effect on the sharper curve.

Finding 100% Levels



We may choose a level where the probability of a positive result appears to be 100%, but there may still be chance error. Even at a 99.9% positive level there is a 2% chance that 1 in 20 replicates will be negative. By 100 replicates the chance is 9.5% of at least 1 negative result. Yet at 1000 replicates, the chance of getting 1000 positive results is 36.8%.– probability works against perfection.

Simulation 1: Chance Performance for One Set of 20 Samples

Series1	Series 2	Series 3
48	50	65
31	40	3
52	27	86
54	71	46
31	16	84
34	96	5
57	13	6
22	16	75
24	7	89
15	62	77
30	38	83
77	49	6
56	75	63
68	90	30
87	40	98
57	66	73
20	23	91
72	61	92
79	13	68
62	69	18

- Twenty sets of numbers ranging from 1-100 were randomly generated
- Numbers were coded based on the response at a 95% probability (1-95 positive green, negative 96-100 red.)
- In this case, one site gave 100% positive responses, the other two gave 95% positive.

Statistically, results ranging from 85-100% positive would not be unexpected for n=20.

Simulation 2: CLIA Study Performance

- Twenty random numbers ranging from 1-100 were generated in three series.
- The number was compared to the likelihood of positive responses for 20 replicates of a 95% positive level:
  - 36% chance of 100% positive responses
  - 38% chance of 95% positive responses
  - 19% chance of 90% positive responses
  - 6% chance of 85% positive responses
  - 1% chance of 80% positive responses
- This gave likely CLIA outcomes for 20 studies where 20 replicates are run at 3 sites (series).
- Results were color coded to make comparison easier as shown.

Simulation 2: CLIA Study Performance Results

	Series 1	Series 2	Series 3
Set 1	95%	95%	95%
Set 2	100%	95%	100%
Set 3	95%	100%	90%
Set 4	95%	95%	95%
Set 5	100%	100%	90%
Set 6	100%	85%	100%
Set 7	95%	100%	100%
Set 8	100%	100%	90%
Set 9	100%	100%	90%
Set 10	100%	95%	90%
Set 11	100%	95%	90%
Set 12	90%	95%	100%
Set 13	95%	90%	95%
Set 14	95%	90%	100%
Set 15	90%	95%	85%
Set 16	95%	95%	95%
Set 17	100%	100%	90%
Set 18	95%	95%	90%
Set 19	90%	100%	95%
Set 20	95%	95%	100%

- CLIA Cutoff recommends 20 replicates at 3 sites.
- Each set of results represents a CLIA study when read across; and shows the variability that can randomly occur in a three site CLIA study.
- Each series is equally likely, and is not related to either device or operator performance.

Simulation 2: Analysis of the Performance

- For 20 replicates at a performance of 95% positive:
  - **100%** expect 36% / observed 35%
  - **95%** expect 38% / observed 40%
  - **90%** expect 19% / observed 22%
  - **85%** expect 6% / observed 3%
  - **80%** expect 1% / observed 0%
- When the values are looked at in groupings of 3 series (CLIA 3 x 20 replicates)
  - The probability of getting 3 sets at 100% is 4.7% ($36\%^3$). None were observed
  - The probability of getting 3 sets at 95% is 5.5% ($38\%^3$). Three were observed (15% occurrence on a 5.5% probability)
- **There *appear to be* differences in the results that don't really exist, due to random distribution of observations.**

Concluding Remarks

- The sharper the transition area for the visual device, the greater the error from variation in the concentration.
- Variability of the reading changes with position on the curve.
  - Greatest at 50%
- The fatigue factor from running large numbers of replicates by hand will add more variability in a cutoff study.
- Statistical models are inappropriate for small replicates of qualitative devices
- CAUTION: It is easy to overanalyze data and see patterns in random distributions that can lead to ‘useless’ root cause analysis.