

# Tackling Statistical Uncertainty in Method Validation

**Steven Walfish**  
**President, Statistical Outsourcing Services**  
steven@statisticaloutsourcingservices.com  
301-325-3129

# About the Speaker

- Mr. Steven Walfish is the President of Statistical Outsourcing Services, a consulting company that provides statistical analysis and training to variety of industries. Prior to starting Statistical Outsourcing Services, Mr. Walfish was the Senior Manager Biostatistics, Non-clinical at Human Genome Sciences in Rockville MD. Prior to joining HGS, Mr. Walfish was a Senior Associate at PricewaterhouseCoopers specializing in the pharmaceutical industry.
- Mr. Walfish brings over 18 years of industrial expertise in the development and application of statistical methods for solving complex business issues including data collection, analysis and reporting. Mr. Walfish has held positions with Johnson & Johnson and Chiron Diagnostics where he worked with large data sets for monitoring process data.
- Mr. Walfish holds a Bachelors of Arts in Statistics from the University of Buffalo, Masters of Science in Statistics from Rutgers University and an Executive MBA from Boston University.

# Learning Objectives

- Dealing with data that cannot be easily analyzed
- Dealing with failure in validation
- Outliers in your data

# Agenda

- Accuracy when a standard does not exist.
- What happens when you have too big or too small of a sample size.
- What happens when linearity fails.
- Quantification Limit issues in a high sensitivity assay.
- Outliers in the data

# Accuracy When No Standard Exists

- The easiest validation occurs when a WHO or other standard exists for the analyte tested.
- Usual ways to deal with a lack of a standard:
  - Spike in analyte to buffer.
  - Dilution of a sample.
  - Use standard curve material
  - Use another approved method to assess the analyte level.

# Issues With Each Approach

- Spike in analyte to buffer.
  - Is the sample homogenous?
  - Is the analyte spiked the same affinity as naïve analyte in a sample?
  - Calculation errors in the preparation.
- Dilution of a sample.
  - Dilution error carries through the samples.
  - Does the analyte concentration dilute proportionally?
  - Sample carryover.
- Use standard curve material
  - Not all methods use standard curve material.
  - Sometimes standard curve material is a different matrix than the samples being tested.
- Use another approved method to assess the analyte level.
  - If another methods exists, why would we validate this method? ☺
  - Is the other method as specific as the method under consideration?
  - Any bias in the other method carries over to this method validation.

# Dealing with a Lack of a Standard

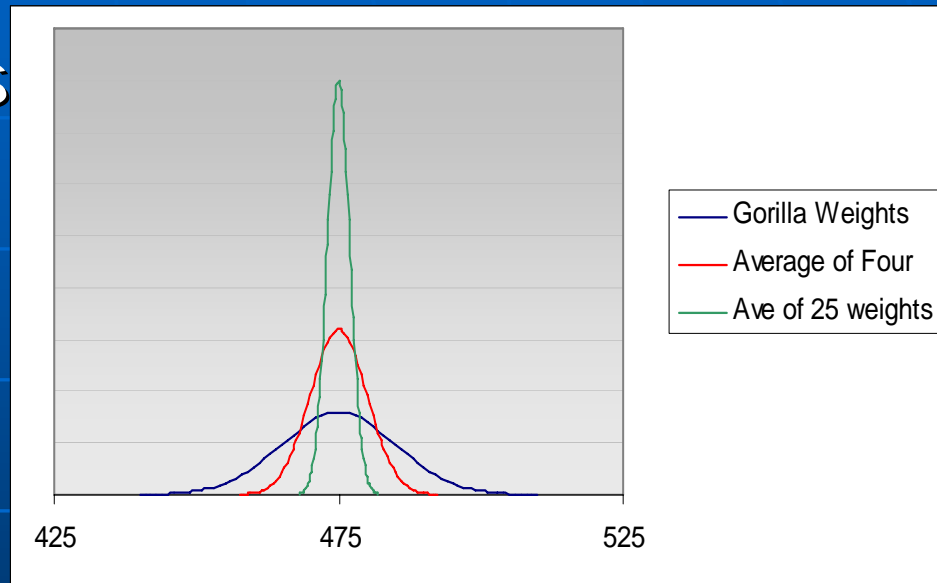
- Use confidence intervals around estimates.
- Increase sample size to increase precision of the estimate.
- Do a method comparison study with another method to assess accuracy and bias.
- Do specificity testing prior to accuracy testing to show test method is specific to the matrix and spiked in analyte.
- Set an overall CV specification to show that the measurements are not too variable.

# Sample Size, Statistical Precision, and Statistical Power

- Statistical power is defined as the ability to detect effects when the effect is present.
- It is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

# Sample Size, Statistical Precision, and Statistical Power

- Increasing the sample size increases the precision of the sample estimate
- If we take a large sample then the sample mean is closer (in distribution) to the population mean



# Sample Size, Statistical Precision, and Statistical Power

- Hypothesis Testing and Types of

Errors		REALITY	
		$H_0$ is True	$H_0$ is False & $H_A$ is True
DECISION	Accept $H_0$	Correct Decision	Type II error with Probability $\beta$ (Depends on true value of $\mu$ )
	Reject $H_0$	Type I error with Probability $\alpha$ (we get to specify $\alpha$ )	Correct Decision with Probability $1-\beta$ ( $1-\beta$ is called Power)

# Sample Size, Statistical Precision, and Statistical Power

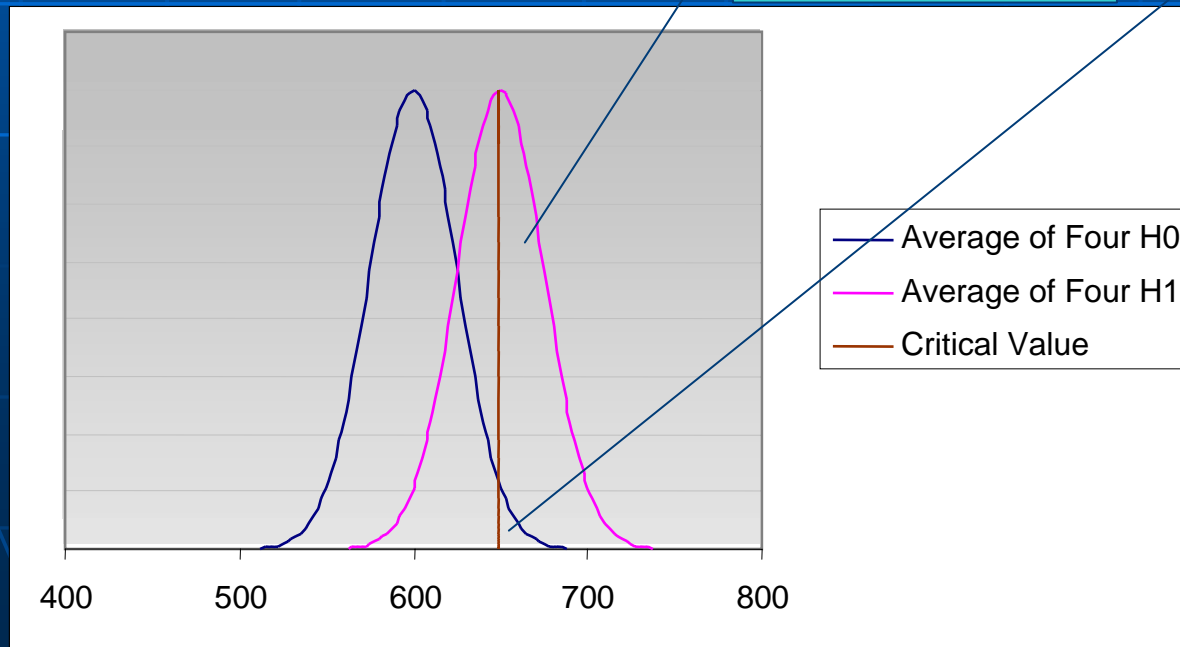
- Example: Test scores follow a normal distribution with population average test score of 600 and standard deviation of 50. There is a claim that a new study method will increase the population average test score.
- $H_0 : \mu = 600$   
(Null hypothesis claims method has the same average as historical average)
- $H_A : \mu > 600$   
(Alternative hypothesis claims method has increased the average. Note that the alternative hypothesis has many possible values for  $\mu$ .)

# Sample Size, Statistical Precision, and Statistical Power

- $H_0: \mu=600$ , with  $n = 4$  and  $\alpha=0.025$  we have power of 0.5160 (51.60%) for  $H_1: \mu= 650$
- In other words, based on a sample size of 4, we have a 51.60% chance of detecting if the samples came from a population with  $\mu=650$ .
- For  $n=25$  the power is 99.88%

Power = 0.5160

$\alpha=0.025$



# Selecting a Sample Size

- What is the risk of rejecting the null hypothesis when it is true ( $\alpha$ )?
- What is the risk of accepting the null hypothesis when it is false ( $\beta$ )? Power is  $1-\beta$ .
- How big a difference do we need to detect?
- Setting your sample size too small leads to the inability to detect a clinically meaningful difference.
- Setting your sample size too large leads to detecting differences that are not clinically meaningful.

# What Happens When Linearity Fails

- Failure of a method to show linearity can be caused by several different components.
  - Mis-specified acceptance criteria
    - $R^2$
    - Slope
    - Residuals
  - Outliers
  - Increasing bias, but within accuracy limits.

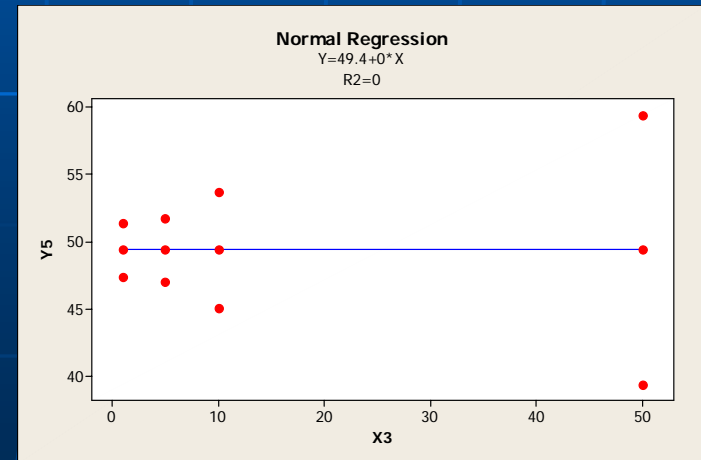
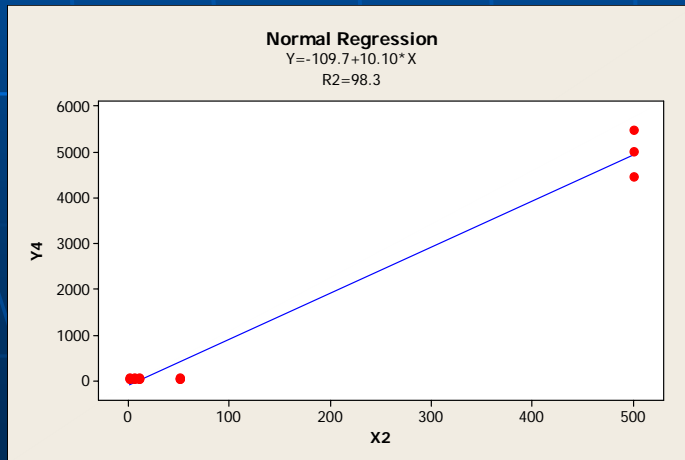
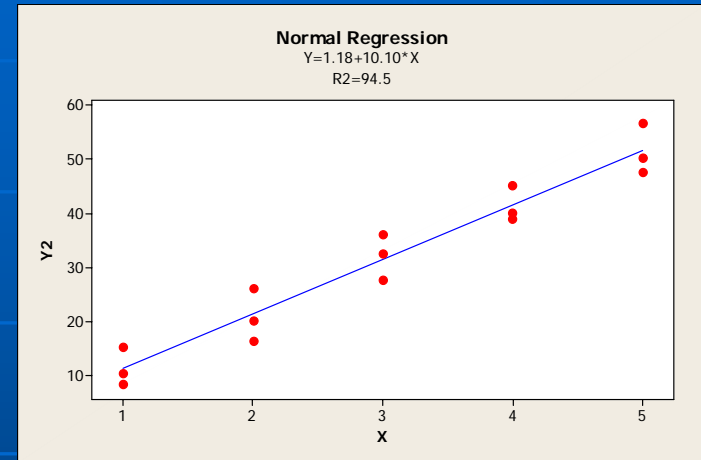
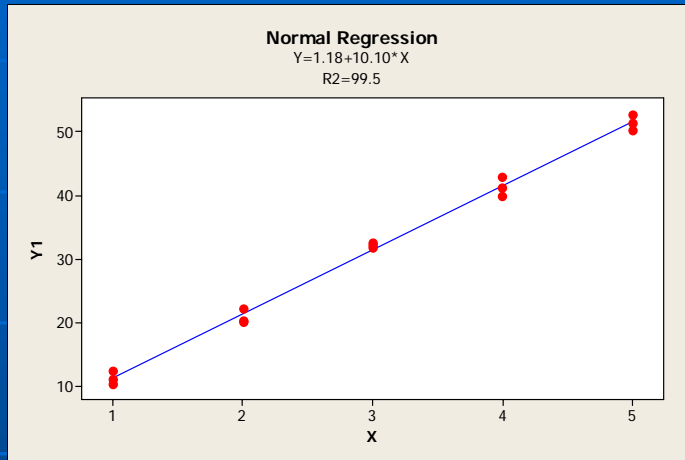
# Mis-specified Acceptance Criteria

- The standard acceptance criteria for linearity are  $R^2$ , slope and intercept values.
- Residuals and Studentized residuals are better acceptance criteria.
- Usually like a slope close to 1, though it is NOT required to be equal to 1. The response needs to be proportional to the spiked amount.
- Having a intercept close to 0 might not be realistic for assays that are several log scales away from zero.

# $R^2$

- $R^2$  can be defined as the percent of the variation explained by the model.
- Unfortunately,  $R^2$  is very sensitive to data distribution and spread.
- A high  $R^2$  does not necessarily mean a good fit.

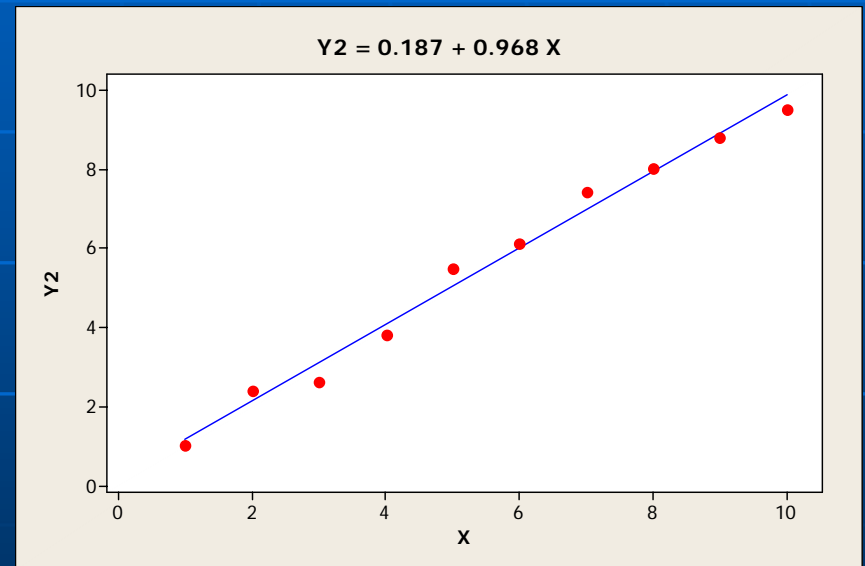
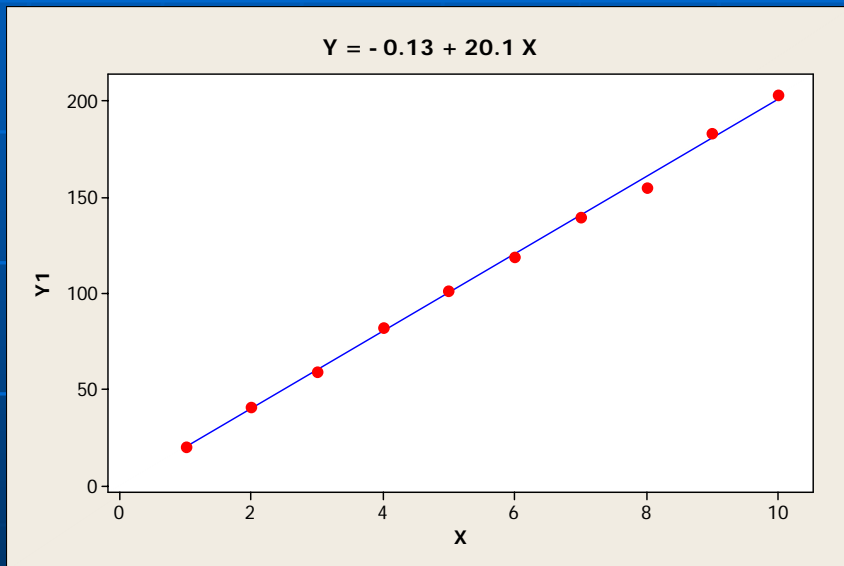
# R<sup>2</sup> Graphically



# Slope

- Typically we desire a slope that is equal to 1 or not statistically different than 1.
- The guidance document states “to obtain test results which are directly proportional to the concentration (amount) of analyte in the sample.”
- Which assay would you want.....?

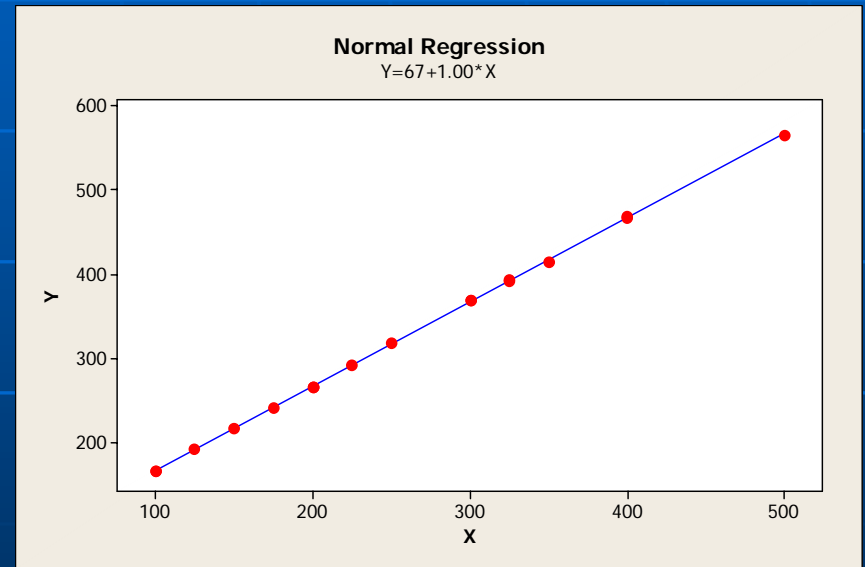
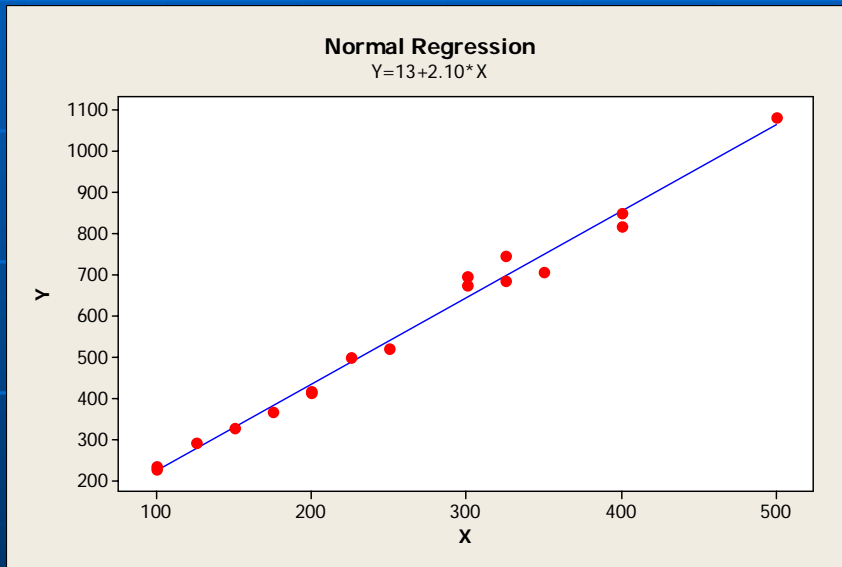
# Slope Graphically



# Intercept

- Typically we desire an intercept that is equal to 0 or not statistically different than 0.
- What does an intercept of 0 mean? It means when  $x=0$  then  $y=0$ , but is that important?
- If the  $x$  values tested are 2 log scales away from 0 (more than 100) and samples will never have values close to 0, the why should it matter?
- Which assay would you want.....?

# Intercept Graphically



# Residuals

- We always consider the validity of the assumptions to be doubtful and conduct analysis to examine the adequacy of the model.
- We cannot detect violations of the assumptions by examining summary statistics.
- A residual is defined as the difference between the observed value and the predicted value from the model.

$$e_i = y_i - \hat{y}_i$$

# Standardized Residuals

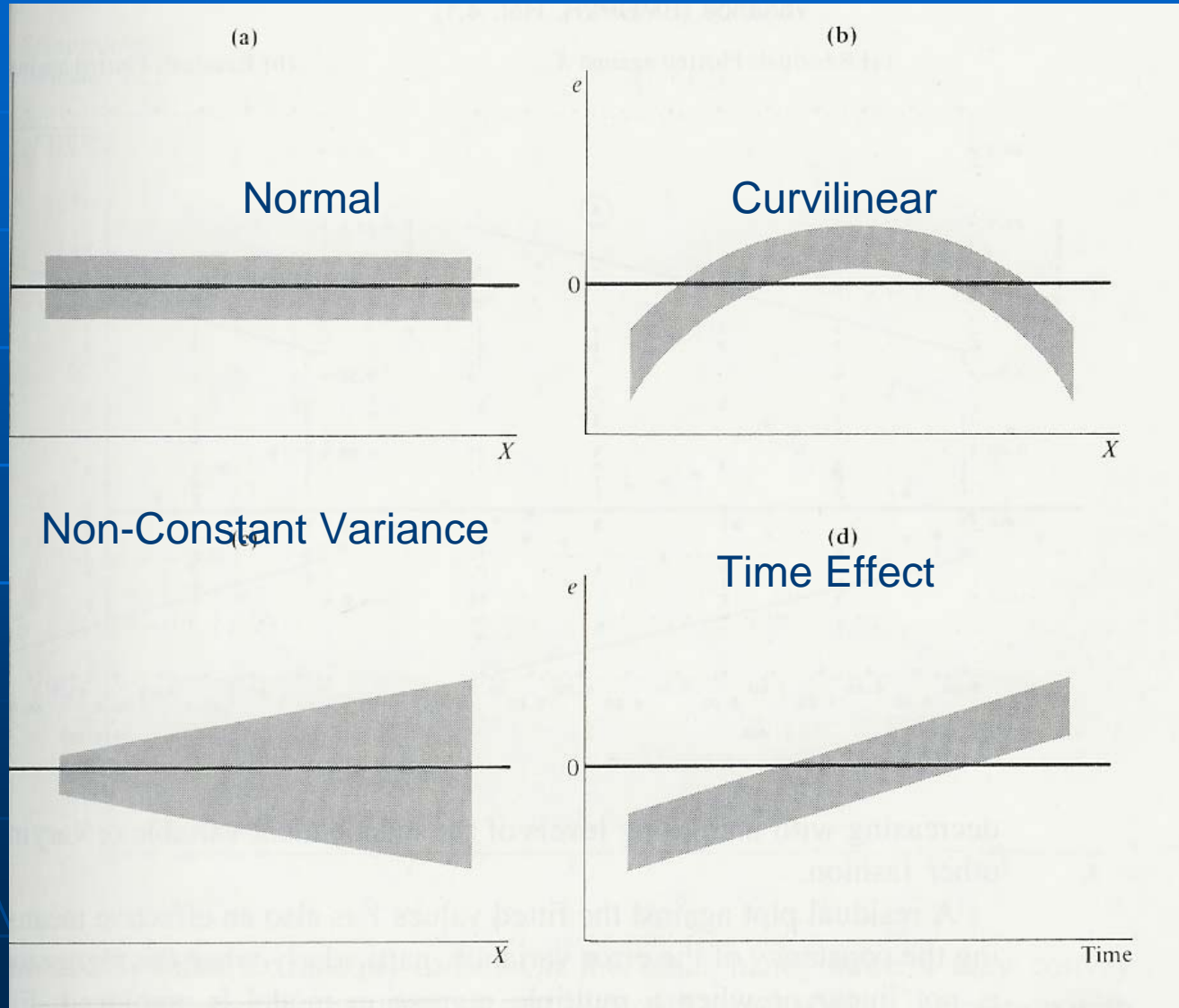
- The standardized residuals have zero mean and unit variance (like a standard normal variable).
- A standardized residual outside of  $\pm 3$  can be considered an outlier.

$$\frac{e_i}{\sqrt{MS_E}}$$

# Residual Plots

- Residual plots are used to assess model fit.
- The following are some uses for residual plots.
  - Regression function is not linear.
  - Error terms do not have constant variance.
  - Error terms are not independent.
  - Model fits all but one or a few outliers.
  - Error terms are not normally distributed.
- Plot residuals versus the X variable.
- For multiple regression plot residuals versus the predicted value.

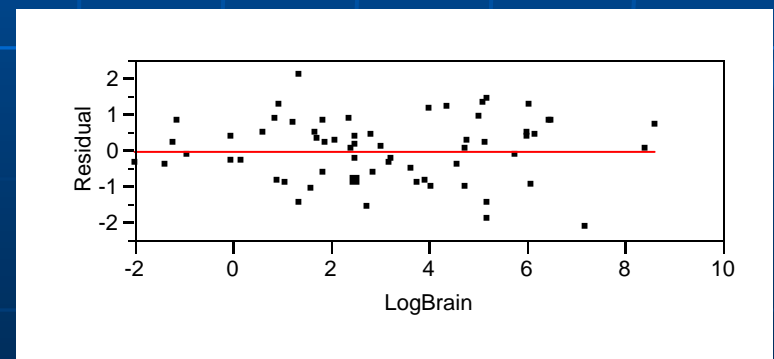
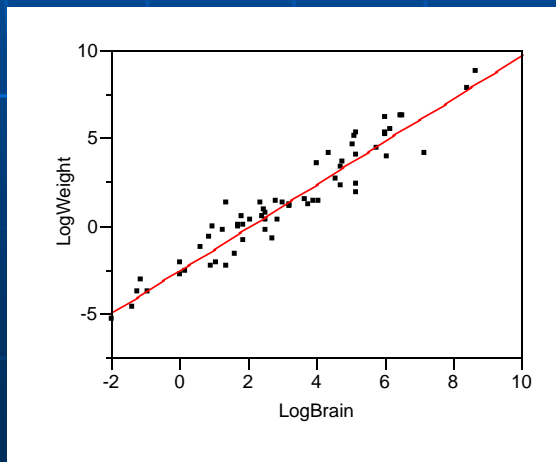
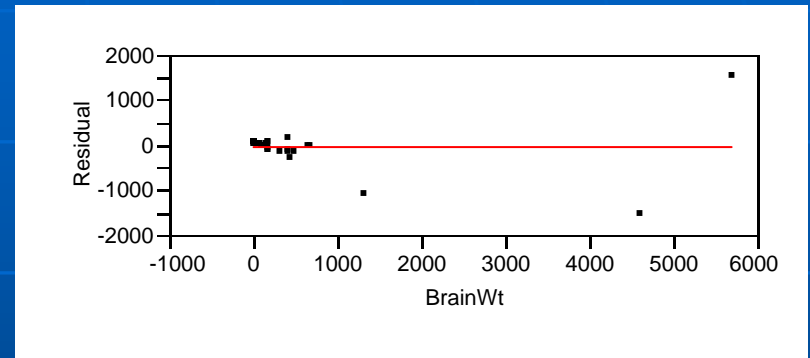
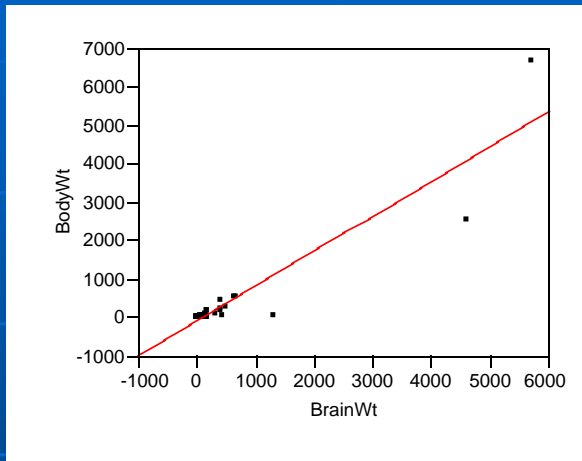
# Prototypical Residual Plots



# Transformations

- Transformations such as log, square root and inverse are normalize data.
- Transformations can also be used to stabilize non-constant variance.
- Transformations help when data is "bunched" together.....

# Log Transformation



# Issues with High Sensitivity Assays

- High sensitivity assays can give false signals.
- Precise assays can be problematic especially in method comparisons.
- As precision increases (variance decreases) statistical hypothesis testing will find smaller differences statistically significant.
- A combination of confidence intervals with acceptance ranges for equivalence.

# What is an outlier?

- An outlier is an observation that “appears” to be inconsistent with other observations in the data set.
- An observation that has a low probability of being a member of the target distribution.
- A discordant observations that is brought to the attention of the investigator.

# How to Handle Outliers

- How to record the data?
  - Always keep and record the outlier observation. They provide valuable information for future study.
  - Attempt to identify an “assignable” cause.
  - Revisit outlier observations as more data becomes available.
- How to handle the data analysis?
  - Use robust statistical methods when possible (i.e. medians).
  - Present the results with and without the suspected outlier for completeness.

# Extreme Studentized Deviate Test

- The extreme studentized deviate (ESD) test is quite good at identifying a single outlier.
- We declare  $x_j$  to be an outlier when

$$T_s = \max_i \{|x_i - \bar{x}| / s\} > \text{tabled value}$$

- Where  $x_j$  is the observation leading to the  $\max T_s$
- If no observation exceeds the tabled value, no observation is a subject as an outlier.

# Critical Values for ESD

N	$\alpha=0.05$	$\alpha=0.01$
10	2.29	2.48
11	2.35	2.56
12	2.41	2.64
13	2.46	2.70
14	2.51	2.76
15	2.55	2.81

# Masking

- The extreme studentized deviate method is susceptible to “masking”.
- Masking occurs when discordant observations cancel the effect of more extreme observations
- Given the following observations:  
5.4; 5.5; 5.6; 5.9; 6.1; 6.3; 6.6; 7.6; 15.4; 15.8
- $T_s$  is 1.92 ( $x_j=15.8$ )
- Critical value is 2.29

# Generalized ESD

- A method for dealing with masking is the idea of testing a pre-specified number of outliers called the generalized ESD.
- This method is quite robust when  $n$  is large ( $>25$ ) and masking effects are possible.
- Compute  $T_s$  for all observations.
- Remove the observation with the largest value of  $T_s$  and recompute with the  $n-1$  observations.
- Continue this process until all possible  $T_s$  are calculated.

# Critical Values

N	$\alpha=0.05$			$\alpha=0.01$		
	# Suspect Outliers					
	1	2	3	1	2	3
10	2.29	2.22	2.13	2.48	2.39	2.27
11	2.35	2.29	2.22	2.56	2.48	2.39
12	2.41	2.35	2.29	2.64	2.56	2.48
13	2.46	2.41	2.35	2.70	2.64	2.56
14	2.51	2.46	2.41	2.76	2.70	2.64
15	2.55	2.51	2.46	2.81	2.76	2.70

# Dixon Type Tests

- Methods based on ordered statistics.
- Flexibility of testing a specific observation as an outlier.
- Excellent for small sample size.
- No need to assume a distribution such as the normal distribution.

# Dixon Tests

- Testing the largest observation as an outlier

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}$$

- Testing the smallest observation as an outlier.

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1}$$

# Dixon Tests

- Single outlier, avoiding  $x_n$

$$r_{11} = \frac{x_n - x_{n-1}}{x_{n-1} - x_1}$$

- Single outlier, avoiding  $x_1$

$$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$$

# Critical Values for Dixon Tests ( $\alpha=0.05$ )

N	$r_{10}$	$r_{11}$
3	0.941	
4	0.765	0.955
5	0.642	0.807
6	0.560	0.689
7	0.507	0.610
8	0.468	0.554
9	0.437	0.512
10	0.412	0.477

# Final Thoughts

- Accuracy when a standard does not exist requires using spiked in amounts that might or might not reflect true levels in a sample (i.e. matrix effect)
- Statistical power can lead to detecting differences that are not clinical relevant.
- Failure of linearity might require a transformation of the data.
- Outliers in the data can create apparent unacceptable performance in the assay.