

# How Far is too Far? Statistical Outlier Detection



---

**Steven Walfish**  
**President, Statistical Outsourcing  
Services**

steven@statisticaloutsourcingservices.com  
301-325-3129



# Outline

---

- **What is an Outlier, and Why are they Important?**

- Why Study Outliers?
- How to Handle Outliers?
- Boxplots
- Trimmed Means
- Confidence Intervals

- **II. Outlier Identification**

- The Extreme Studentized Deviate Test (ESD)
- Generalized ESD
- Dixon-Type Tests

- • Shapiro-Wilks Test

- **III. Regression**

- Influential Observations
- Effect of Outliers on Slope, Intercept and R-Squared
- Residual Analysis

- **IV. Comparing the Methods**

- Which test to use when

- **VI. Interactive Exercise** – Interactive exercise to review outlier methods. Determine how to handle outliers. Deciding when to retest, when to reject and when to do nothing.



# 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.



# Why Study Outliers?

---

- Outliers provide useful information about the data.
- Outliers can come from a location shift or a scale shift in the data.
- Outliers can come from a gross recording error or measurement error.



# Why Study Outliers?

---

- Sample time or sample frame errors (i.e. morning versus afternoon).
- Sometimes the value is plausible, but unexpected.



# 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



# Outlier Labeling

---

- Prior to doing any statistical analysis, data should be reviewed and checked for assumptions.
- Graphical methods can be used to visually accept assumptions such as normality and the lack of outliers.
- Some methods include box plots, trimmed means and confidence intervals.

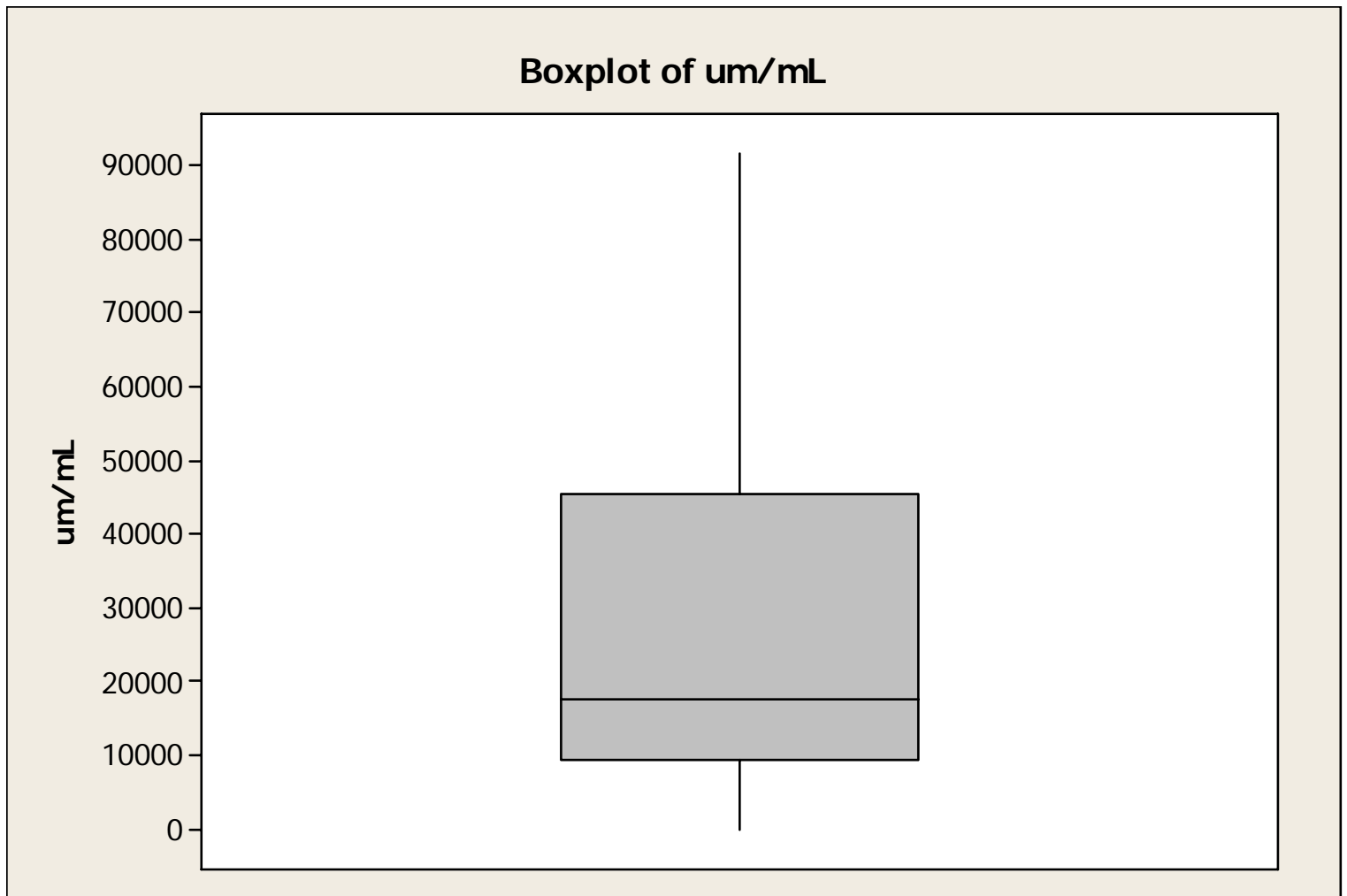


# Box Plots

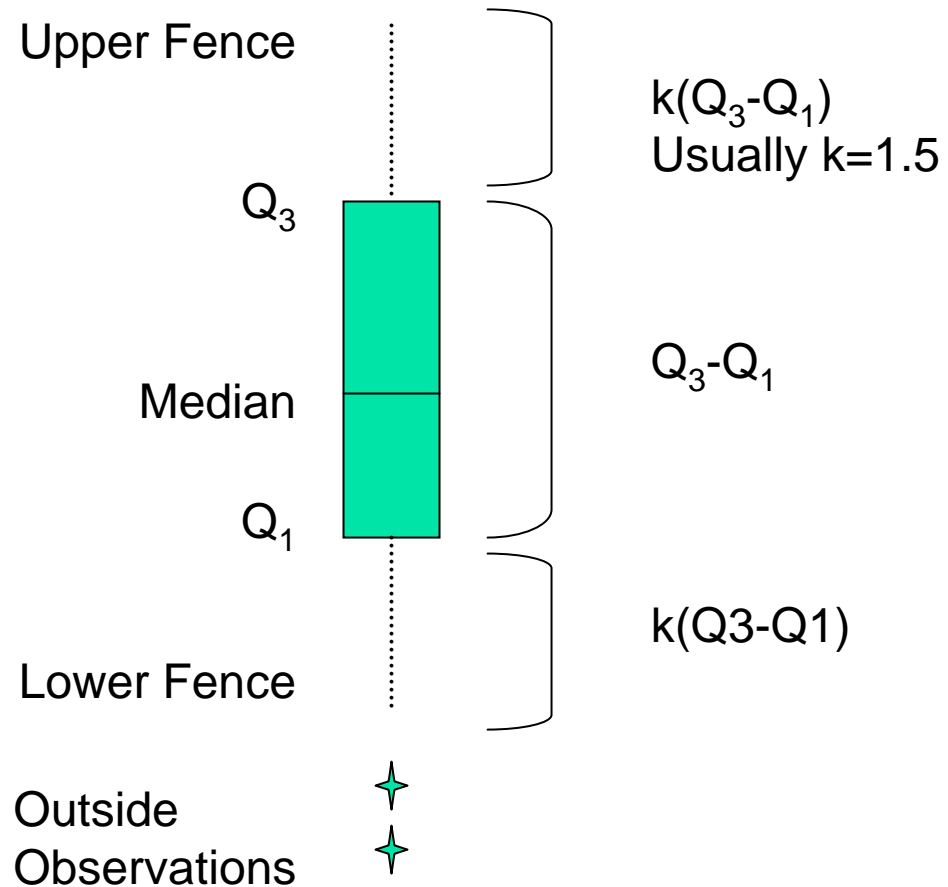
---

- A box plot is a graphical representation of dispersion of the data.
- The graphic represents the lower quartile ( $Q_1$ ), upper quartile ( $Q_3$ ), and median.
- The box includes the range of scores falling into the middle 50% of the distribution.

# Example of Box Plots



# Decomposition of a Box Plot





# Trimmed Means

---

- Assume we want to estimate the population average for purity. Clearly, the sample mean would be the logical choice. The mean of the 10 lots is 96.85.
- Though we know from the previous box plot, that the 90.1 purity value is flagged as “far out” when compared to the other observations.
- Since the smallest value might unduly influence the mean, we might want to drop it from the analysis. To preserve symmetry, we would also drop the largest value.
- The trimmed mean of the remaining 8 observations would be 97.36.



# Using Ordered Observations

---

- Instead of removing a single observation, we might be interested in having a  $100\alpha\%$  trimmed mean.
- Assume we want a 15% trimmed mean:
  - Calculate  $r$  as  $\sum_{i=r+1}^{n-r} \alpha(n+1)$ , rounded to the closest integer.

# The Steps Using the Purity Data

- $r=0.15(10+1)=1.65$  which rounds to 2.

- If we order the observations

90.1 95.4 96.1 96.3 97.6 97.9 98.1 98.3 99.2 99.5

- Calculating T by eliminating the smallest 2 and largest 2 observations yields a trimmed mean of 97.38
- Mean of all 10 observations = 96.85
- Mean of trimmed (n=8) = 97.36
- Mean of trimmed (n=6) = 97.38



# Confidence Intervals

---

- The standard method for labeling potential outliers is the  $\pm 3s$  or  $\pm 2s$  limits on the mean.
- The standard formula would be:

$$\bar{X} \pm t_{(1-\alpha/2;n-1)} * \frac{s}{\sqrt{n}}$$

- Potential outliers influence the estimate of the standard deviation ( $s$ ) and the sample mean
- Winsorized variances in conjunction with the trimmed mean can be used for robust methods.



# Outlier Identification

---

- Most outlier tests rely on the normal distribution.
- If the data is heavy tailed there is a likelihood for false positives.
- Later we will discuss which tests to use in a comparative workshop.

# 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  $\max T_s$



# 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



# Assuming 2 Outliers

Observation	$x_i$	$T_s$	$x_i$	$T_s$
1	5.4	0.647	5.4	0.555
2	5.5	0.623	5.5	0.523
3	5.6	0.598	5.6	0.491
4	5.9	0.524	5.9	0.397
5	6.1	0.474	6.1	0.333
6	6.3	0.425	6.3	0.270
7	6.6	0.351	6.6	0.176
8	7.6	0.104	7.6	0.140
9	15.4	1.824	15.4	2.605
10	15.8	1.922		
Mean	8.02		7.16	
Std Deviation	4.05		3.17	
Maximum $T_s$	1.92		2.60	
Critical Value	2.29		2.22	



# 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.

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

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

N	$r_{10}$
3	0.941
4	0.765
5	0.642
6	0.560
7	0.507
8	0.468
9	0.437
10	0.412



# Example

---

4.5
4.9
5.5
6.1
6.7
7.2
10.2
14.3
$r_{10}=0.418$



# Outliers in Regression Analysis

---

- Two different types of observations have an impact on regression analysis
  - Influential observations
  - Outliers
- Residual analysis is a methodology used to identify outliers.

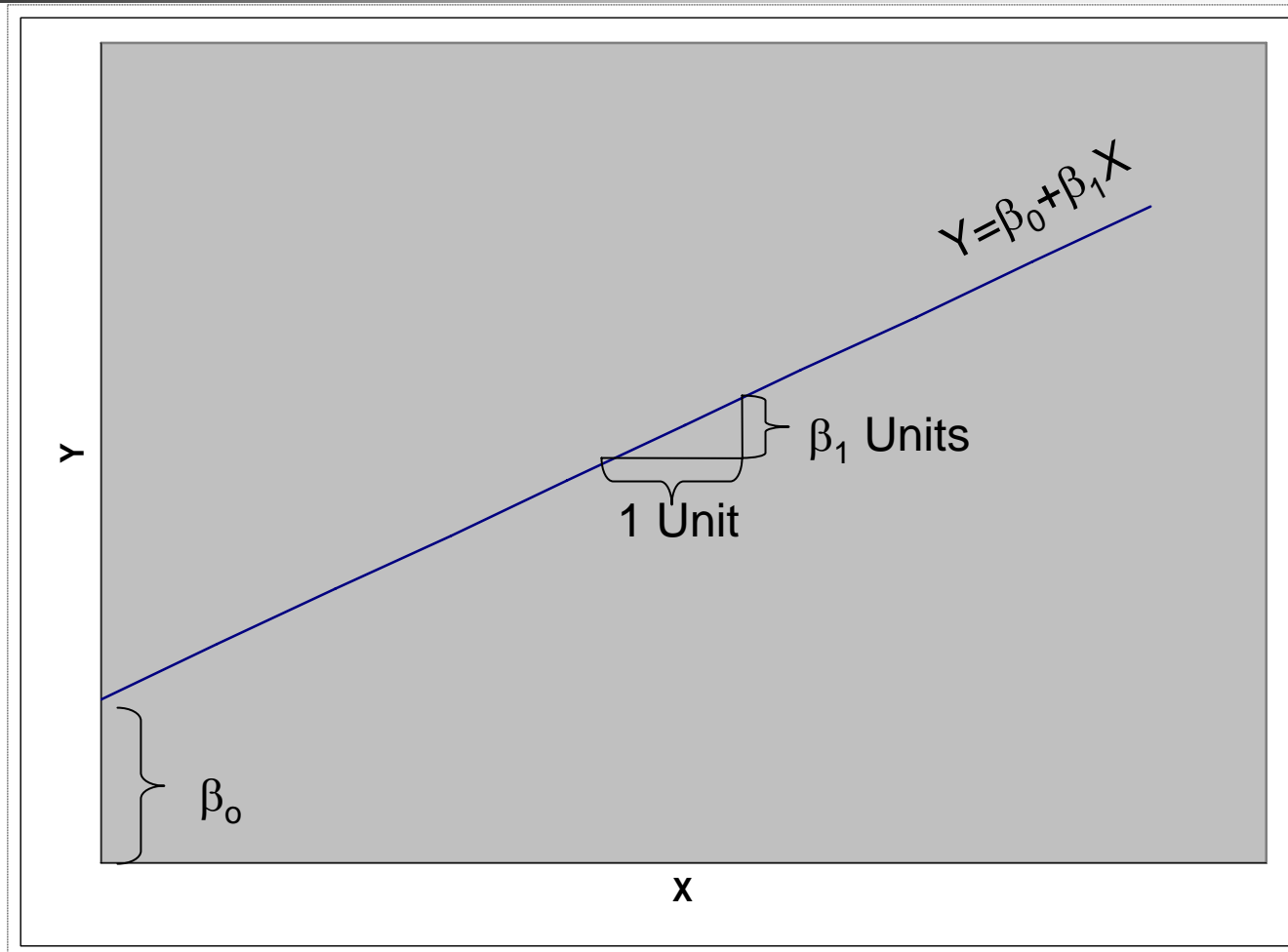


# Regression Analysis

---

- Regression analysis or least squares estimation is a statistical technique to establish a linear relationship between two variables.
- These techniques are highly sensitive to outliers and influential observations.
- Two parameters are estimated; intercept and slope.

# Simple Regression Model





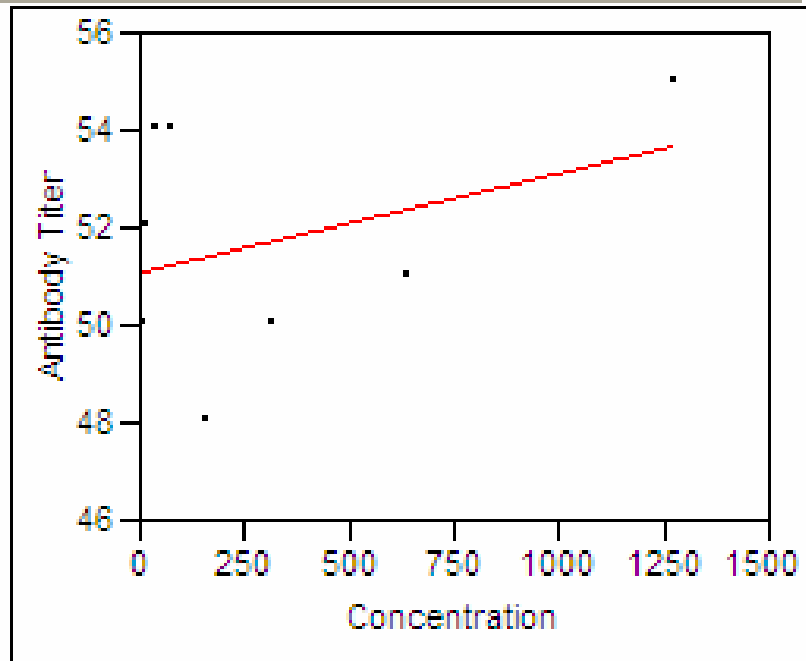
# Influential and Outlier Observations

---

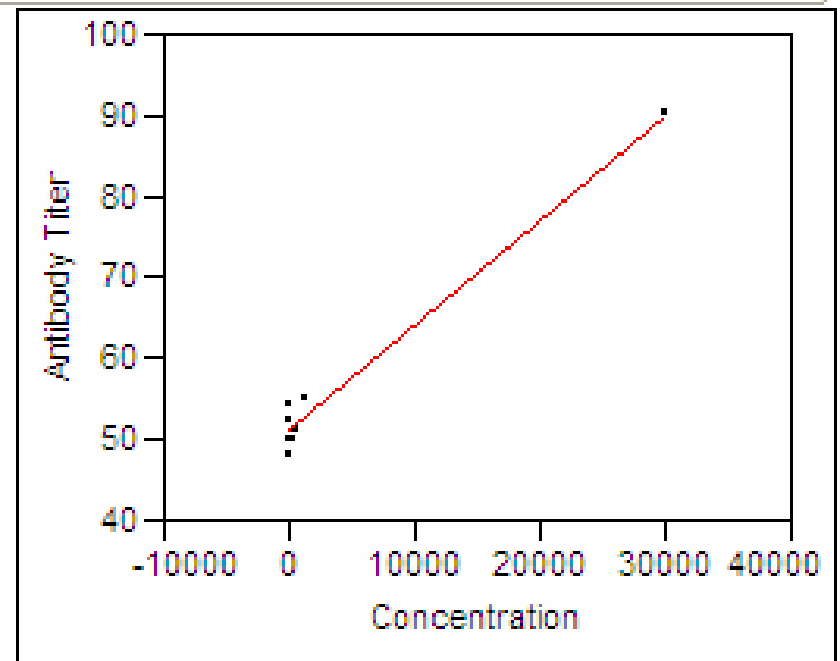
- Outliers are usually a single observation that usually do not influence the overall regression model.
- Influential observations are usually a single observation at the beginning or end of the  $x$  range that highly influences the model.

# Influential Observations

Regression Plot



Regression Plot





# Effect of Outliers on Slope, Intercept and $R^2$

---

- An outlier impacts the slope intercept and  $R^2$  in different ways.
- The slope can be pulled up or down based on the direction of the outlier.
- The intercept is more robust to outliers, but can be impacted by influential observations.



# Comparing the Outlier Detection Methods

---

- So which method is best?
- Outlier methods should be select PRIOR to data collection.
- Should be specified in a procedure.



# Comparison Grid

<b>Method</b>	<b>Sample Size</b>	<b>Number of Outliers</b>	<b>Ease of Use</b>
Box plot	Any size	Multiple	Most statistical software
ESD	Large	Multiple	Moderately easy
Dixon	Small	One or Two	Easily calculated



# Using Outlier Detection in OOS

---

- The cGMP regulations require that statistically value quality control criteria include appropriate acceptance and/or rejection levels (§211.165(d)).
- Outlier tests have no applicability in cases where the variability in the product is what is being assessed, such as content uniformity, dissolution, or release rate determinations.



# Handling Data in Retest Procedures

---

- Three possible ways to handle data:
  - Remove the data; using only the retest procedure.
  - Keep the data; retest and combine data.
  - Retest and compare both sets of data to determine if you can pool the data.

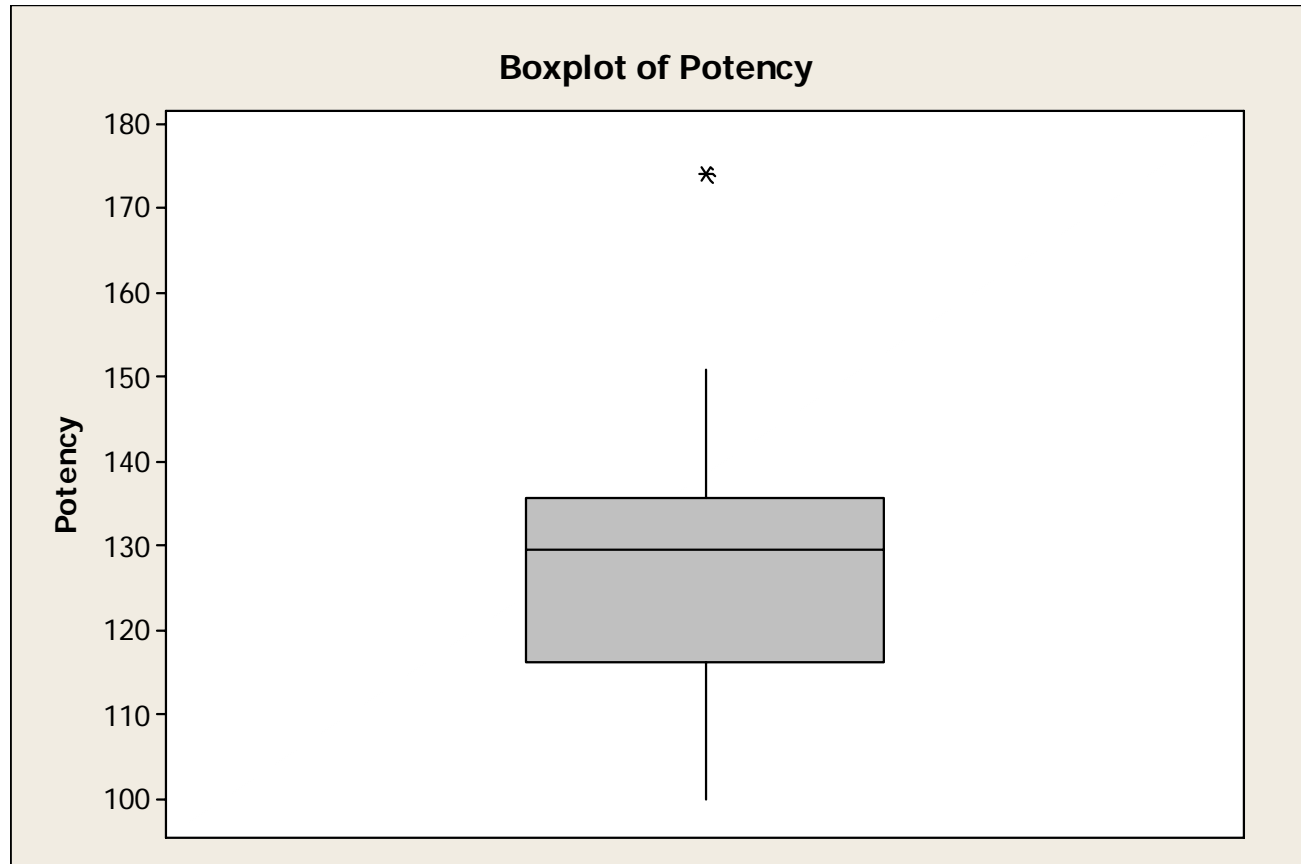


# Interactive Workshop

---

Day	Potency	Day	Potency	Day	Potency
1	119	2	141	3	132
1	106	2	117	3	104
1	107	2	138	3	131
1	120	2	128	3	145
1	104	2	113	3	121
1	131	2	131	3	146
1	100	2	105	3	123
1	132	2	134	3	136
1	151	2	132	3	135
1	106	2	116	3	119
1	127	2	144	3	123
1	<b>174</b>	2	131	3	144

# Boxplot





# Extreme Studentized Deviate

- Looking at all 36 data points, critical value=2.99

Day	Potency	$T_i$	Day	Potency	$T_i$	Day	Potency	$T_i$
1	127	0.010	1	120	0.428	3	144	1.076
2	128	0.073	2	134	0.449	3	145	1.138
3	123	0.240	1	119	0.491	3	146	1.201
3	123	0.240	3	119	0.491	1	107	1.243
1	131	0.261	3	135	0.512	1	106	1.306
2	131	0.261	3	136	0.574	1	106	1.306
2	131	0.261	2	117	0.616	2	105	1.368
3	131	0.261	2	116	0.679	1	104	1.431
1	132	0.324	2	138	0.700	3	104	1.431
2	132	0.324	2	113	0.867	1	151	1.514
3	132	0.324	2	141	0.888	1	100	1.681
3	121	0.366	2	144	1.076	1	174	2.956



# Extreme Studentized Deviate

- Looking at Day 1, 12 data points, critical value=2.41

Day	Potency	$T_i$
1	120	0.141
1	127	0.179
1	119	0.186
1	131	0.361
1	132	0.406
1	107	0.733
1	106	0.779
1	106	0.779
1	104	0.870
1	100	1.052
1	151	1.272
1	174	2.321



# Final Thoughts

---

- Statistical methods for identifying potential outliers in a data set are powerful methods to assist in any data analysis project.
- Pre-specify the method prior to data collection.
- Pre-specify the handling of outliers and retest procedures.