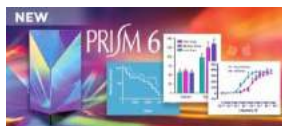


# Comonnly Used Methods in Applied Statistics

Daniel R. Jeske  
Professor and Chair  
Director of Statistical Consulting Collaboratory  
Department of Statistics  
University of California – Riverside

May 14, 2015

Goal for Today: Appreciation of possibilities, not mastery of them.



# Outline

1. Classification Methods
2. Clustering Methods
3. Nonlinear Least Squares
4. Logistic Regression
5. Poisson Regression
6. Sample Size for Two Group Comparison
7. Multiple Group Comparison
8. Sequential Designs
9. Two-Way ANOVA
10. Longitudinal Data (Repeated Measures)
11. Prospective vs. Retrospective Studies
12. Odds Ratios
13. Survival Analysis



# Classification - Trees

Fisher's Famous  
Iris Data



Species 1: Setosa



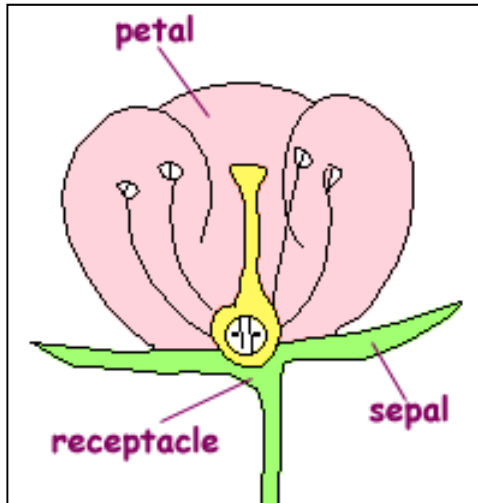
Species 2: Versicolour



Species 3: Virginica

# Classification - Trees

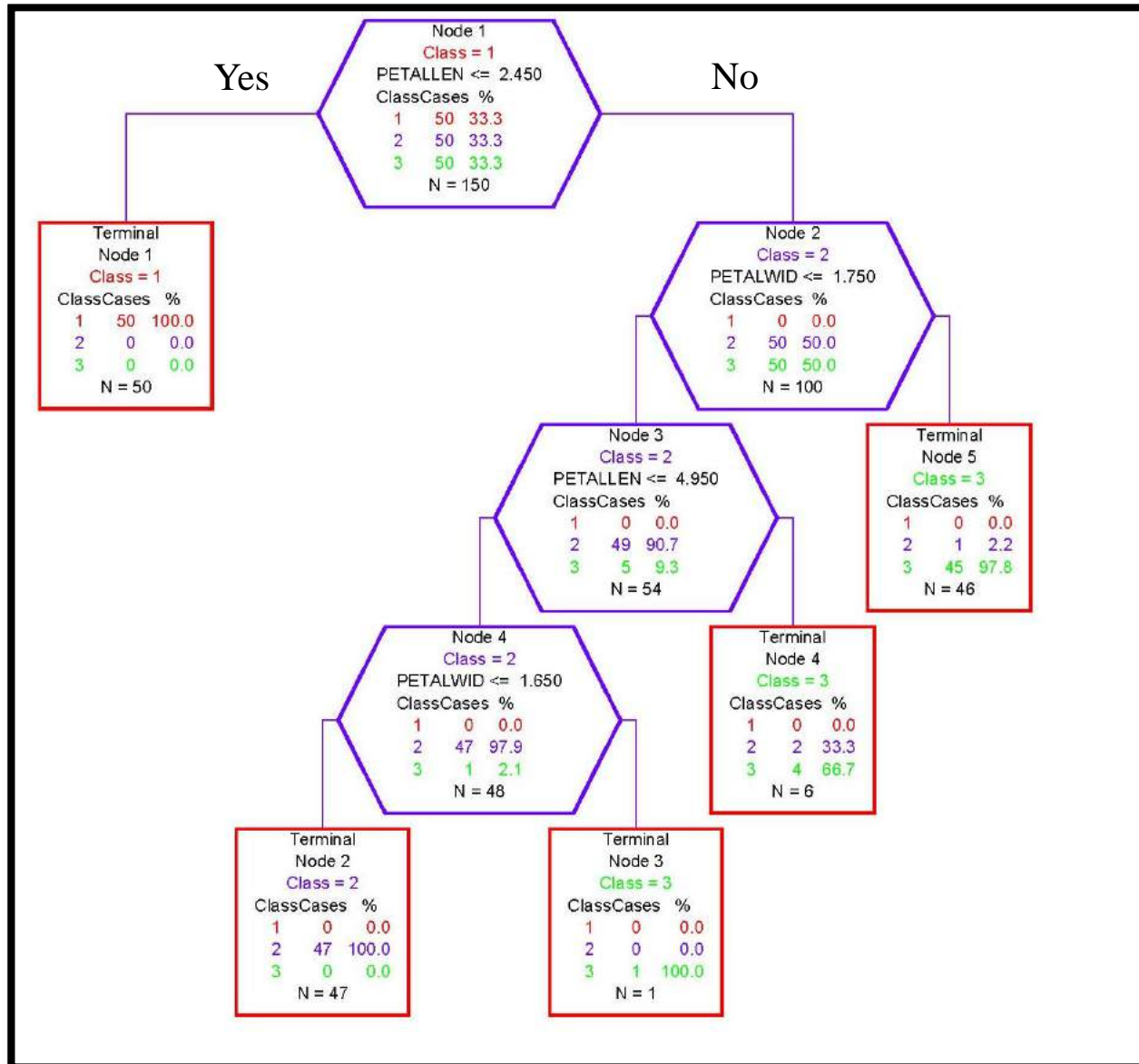
Fisher's Famous Iris Data



ID	SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	1	5.1	3.5	1.4	0.2
2	1	4.9	3	1.4	0.2
.	.	.	.	.	.
.	.	.	.	.	.
49	1	5.3	3.7	1.5	0.2
50	1	5	3.3	1.4	0.2
51	2	7	3.2	4.7	1.4
52	2	6.4	3.2	4.5	1.5
.	.	.	.	.	.
.	.	.	.	.	.
99	2	5.1	2.5	3	1.1
100	2	5.7	2.8	4.1	1.3
101	3	6.3	3.3	6	2.5
102	3	5.8	2.7	5.1	1.9
.	.	.	.	.	.
.	.	.	.	.	.
149	3	6.2	3.4	5.4	2.3
150	3	5.9	3	5.1	1.8

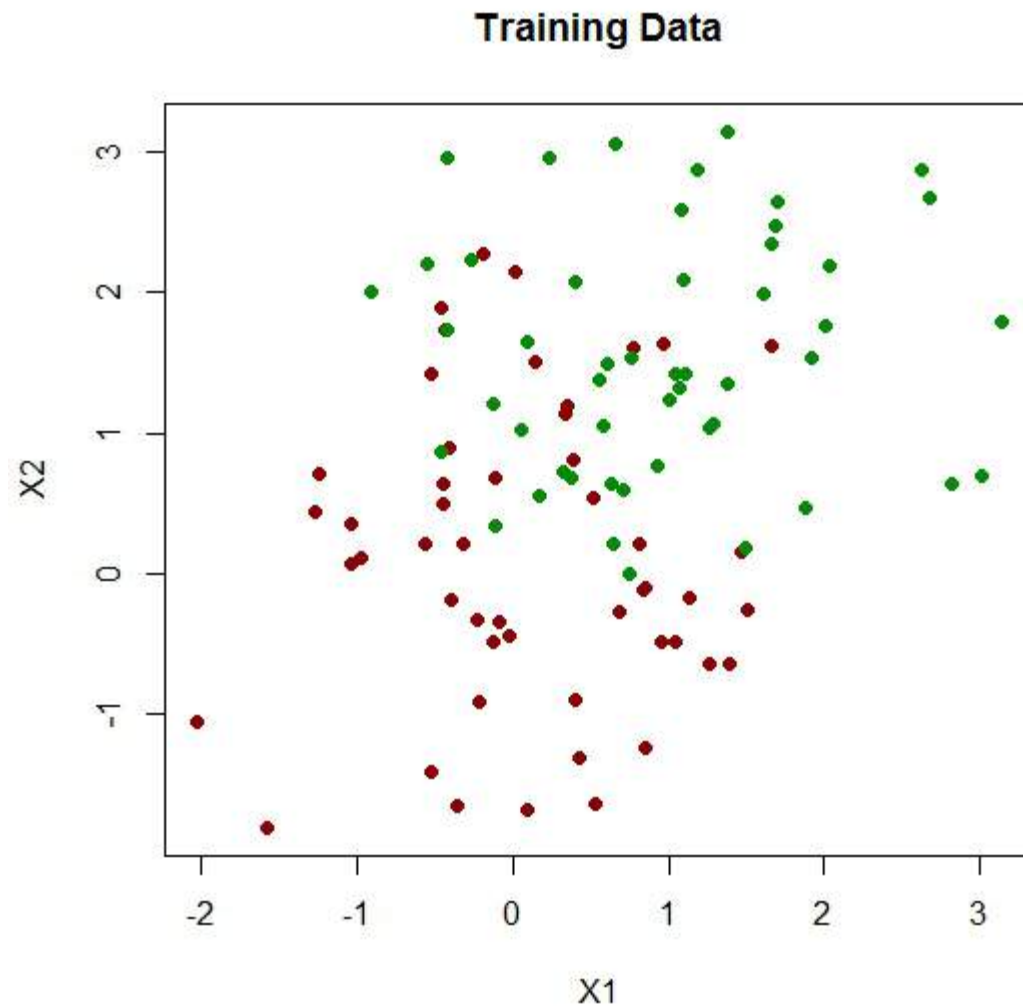
Species: 1 = Setosa , 2 = Versicolour , 3 = Virginica

# Classification - Trees

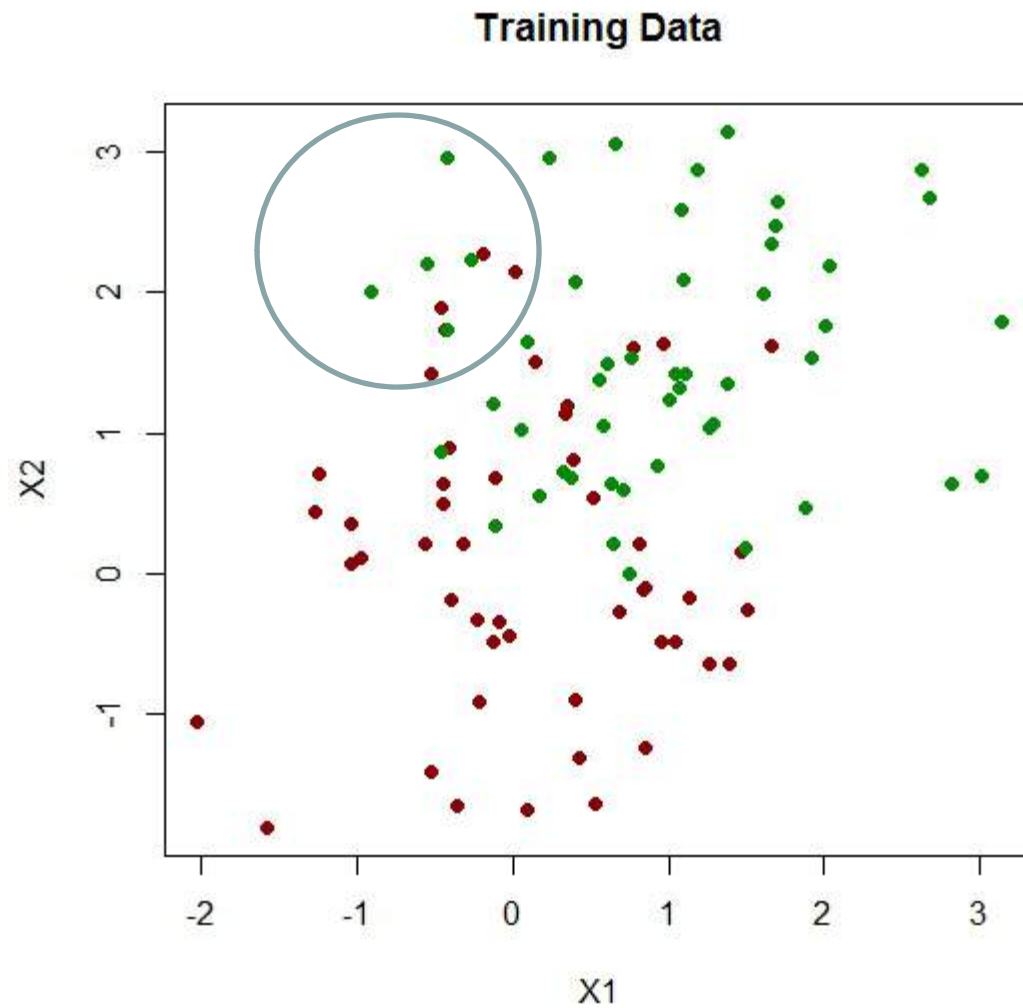




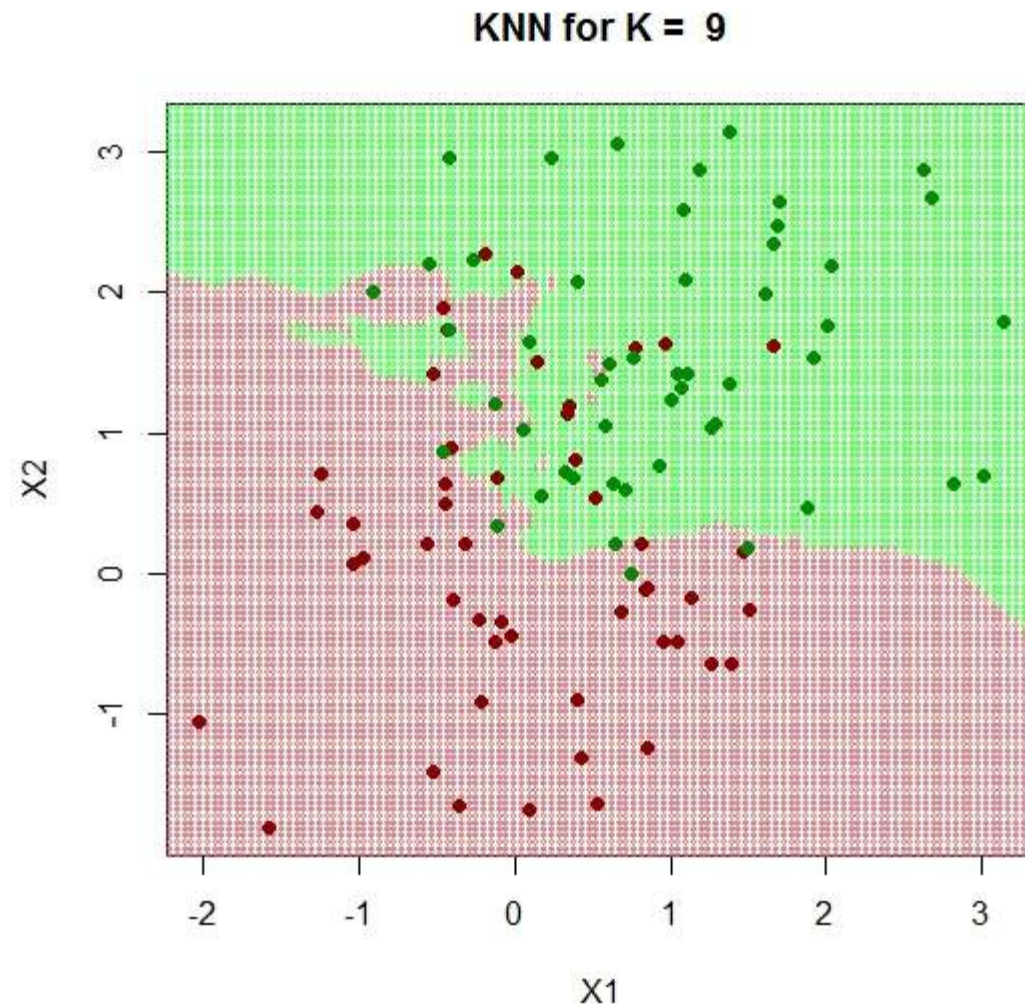
## Classification – k Nearest Neighbors



# Classification – k Nearest Neighbors

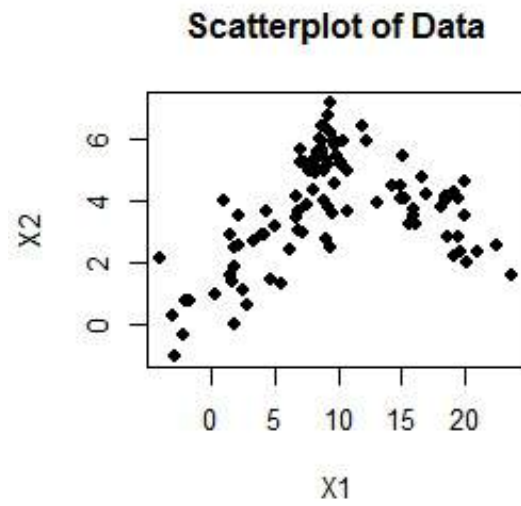


# Classification – k Nearest Neighbors

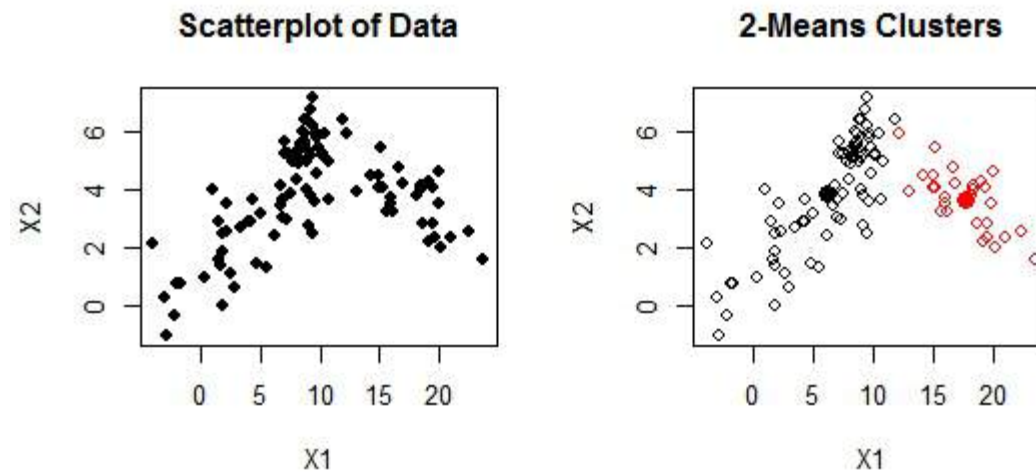




# Clustering – kMeans

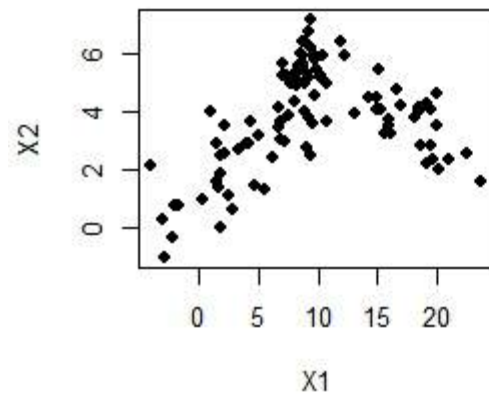


# Clustering – kMeans

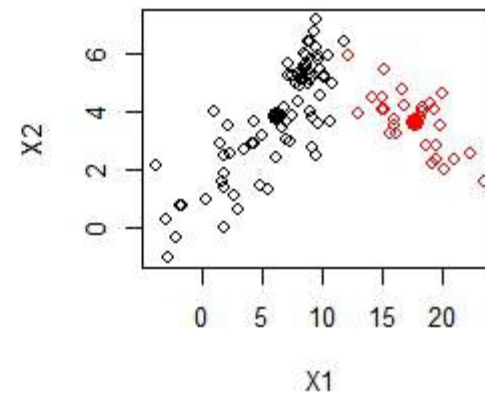


# Clustering – kMeans

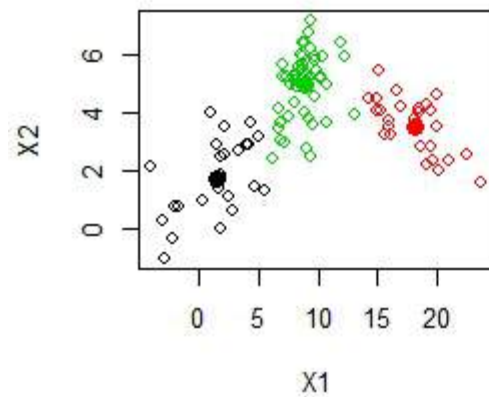
Scatterplot of Data



2-Means Clusters

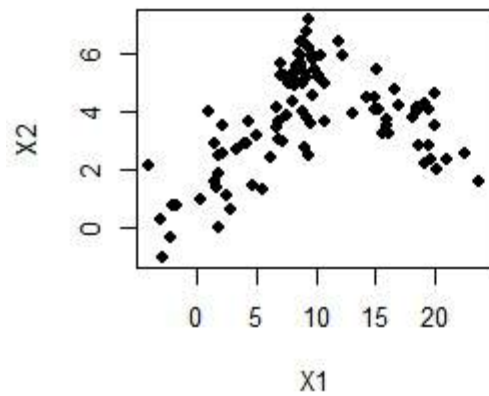


3-Means Clusters

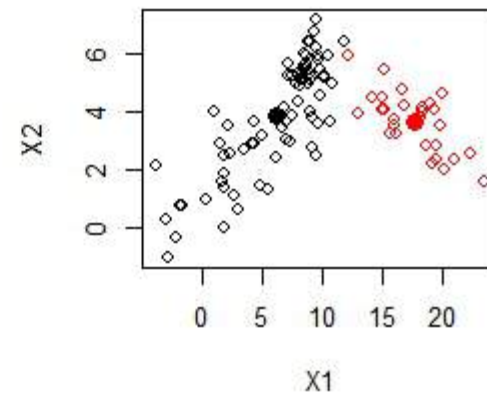


# Clustering – kMeans

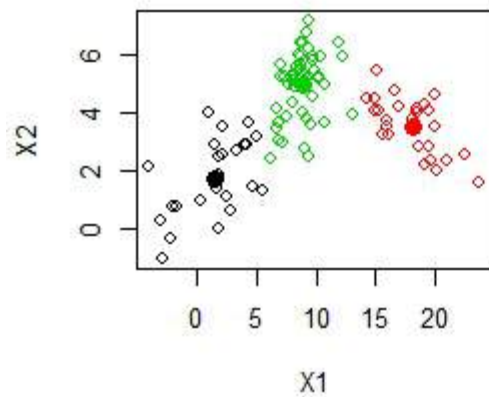
Scatterplot of Data



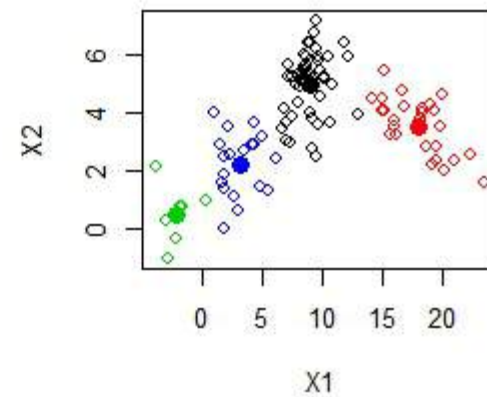
2-Means Clusters



3-Means Clusters



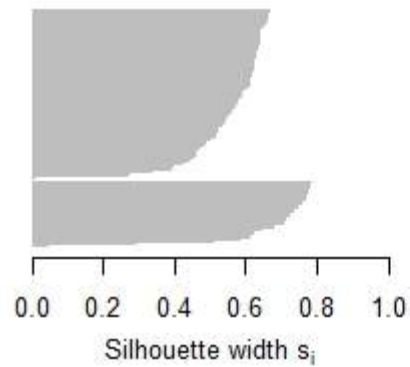
4-Means Clusters



# Clustering – kMeans

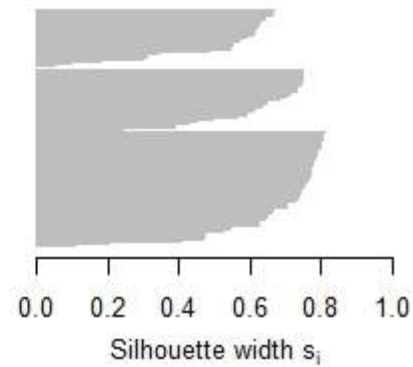
## Silhouette Plots

**2 clusters**



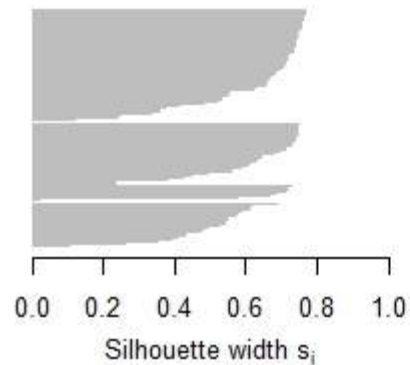
Average silhouette width : 0.59

**3 clusters**



Average silhouette width : 0.63

**4 clusters**



Average silhouette width : 0.61



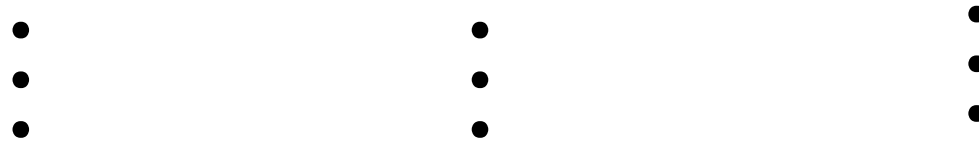


# Clustering - Dendrograms

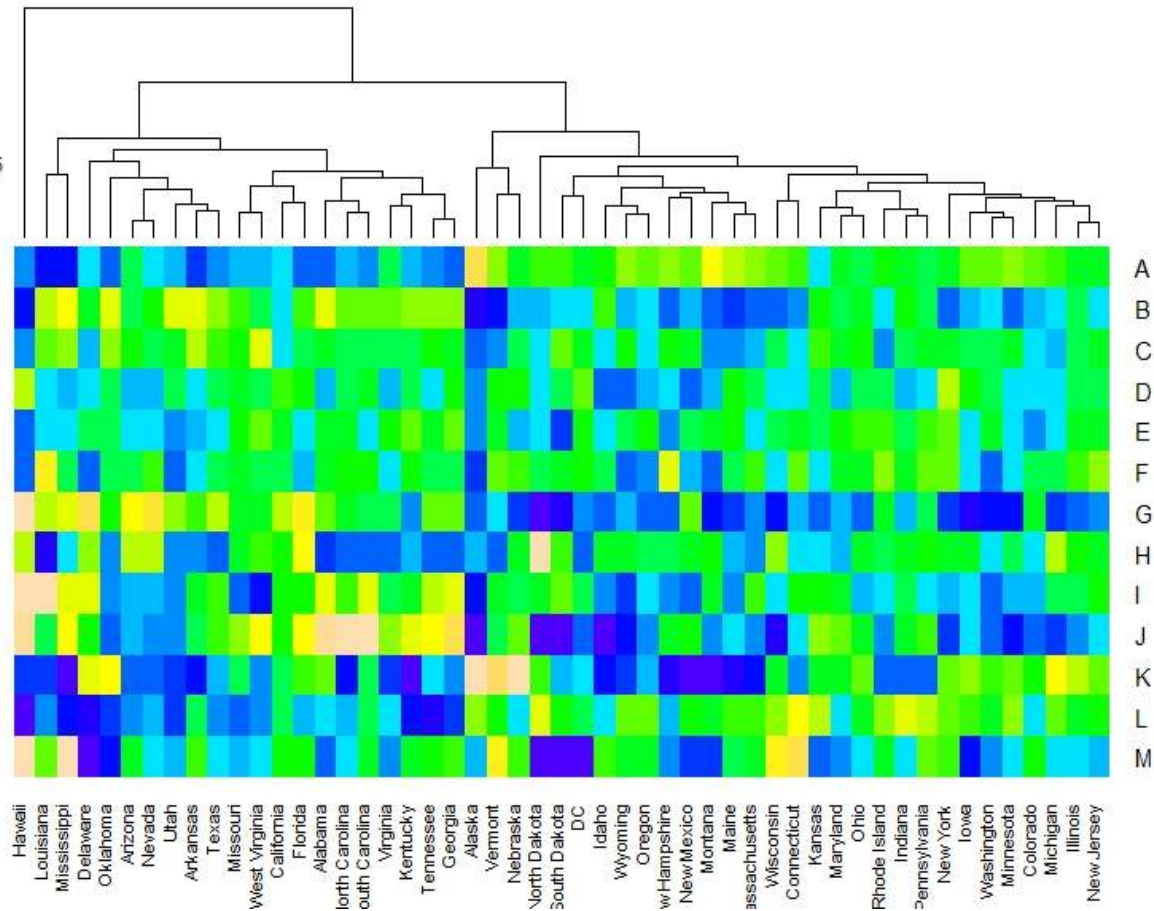
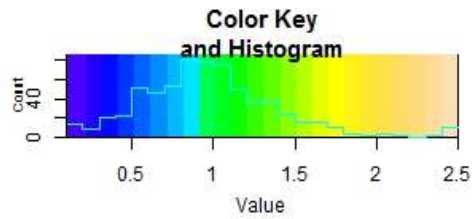
Company Data on sales of products by different states

Data entries are the ratio of state-wide sales of the product relative to Nation-wide sales of the product.

State	A	B	C	D	E	F	G	H	I	J	K	L	M
California	0.9	0.88	0.87	1.29	1.1	0.97	1.53	1.16	1.15	1.11	0.95	0.94	1.17
Texas	0.62	1.44	1.25	0.98	0.81	0.99	1.53	0.59	1.3	1.23	0.78	0.68	0.88
New York	1.04	0.59	1.03	1.58	1.36	1.4	0.48	1.03	0.74	0.45	1.32	1.31	1.22
Washington	1.39	0.83	0.96	0.92	1.04	0.55	0.39	0.88	0.56	0.55	1.24	1.08	0.62
Colorado	1.39	0.79	0.82	0.83	0.62	0.96	1.07	0.87	0.79	0.57	1.2	0.9	1.12
Florida	0.6	1.21	0.97	1.14	0.81	0.75	1.88	1.74	1.12	1.77	1.3	0.8	1.17
Illinois	1.09	0.91	0.95	0.94	1.08	1.21	0.56	1.16	0.91	0.63	1.59	1.09	0.83
Pennsylvania	0.91	0.95	1.1	0.9	1.22	1.33	0.91	1.2	0.89	1.26	0.52	1.51	1.36
Ohio	0.94	1.05	1.11	0.99	1.27	1.02	0.57	1.08	0.72	1.04	1.34	1.04	0.84
Michigan	1.25	0.82	0.76	0.9	0.88	0.91	0.5	1.58	0.94	0.48	1.79	1.35	0.9
Minnesota	1.49	0.6	1.06	0.82	0.81	0.85	0.35	0.92	0.71	0.37	1.4	1.47	0.82



# Clustering States

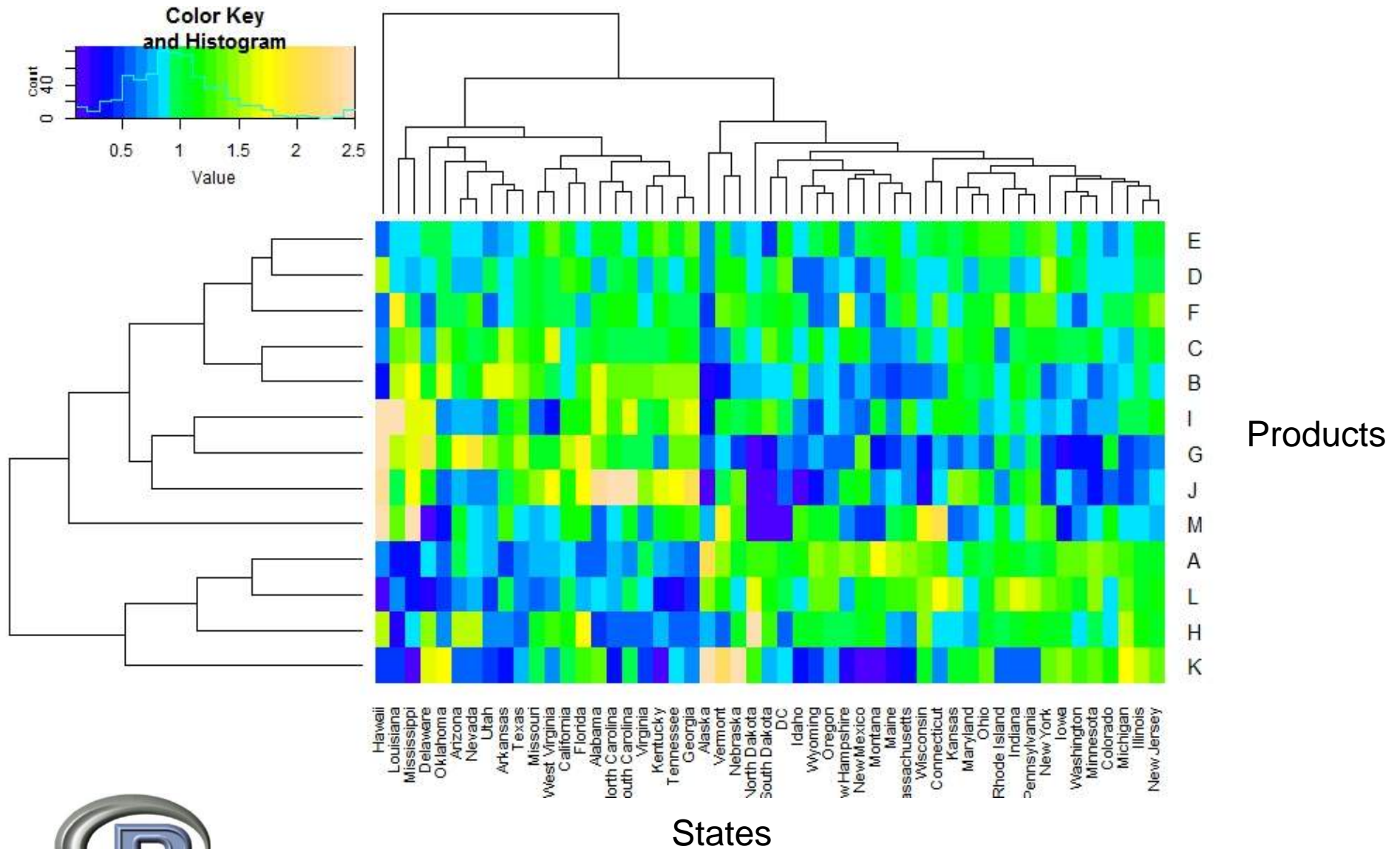


Products

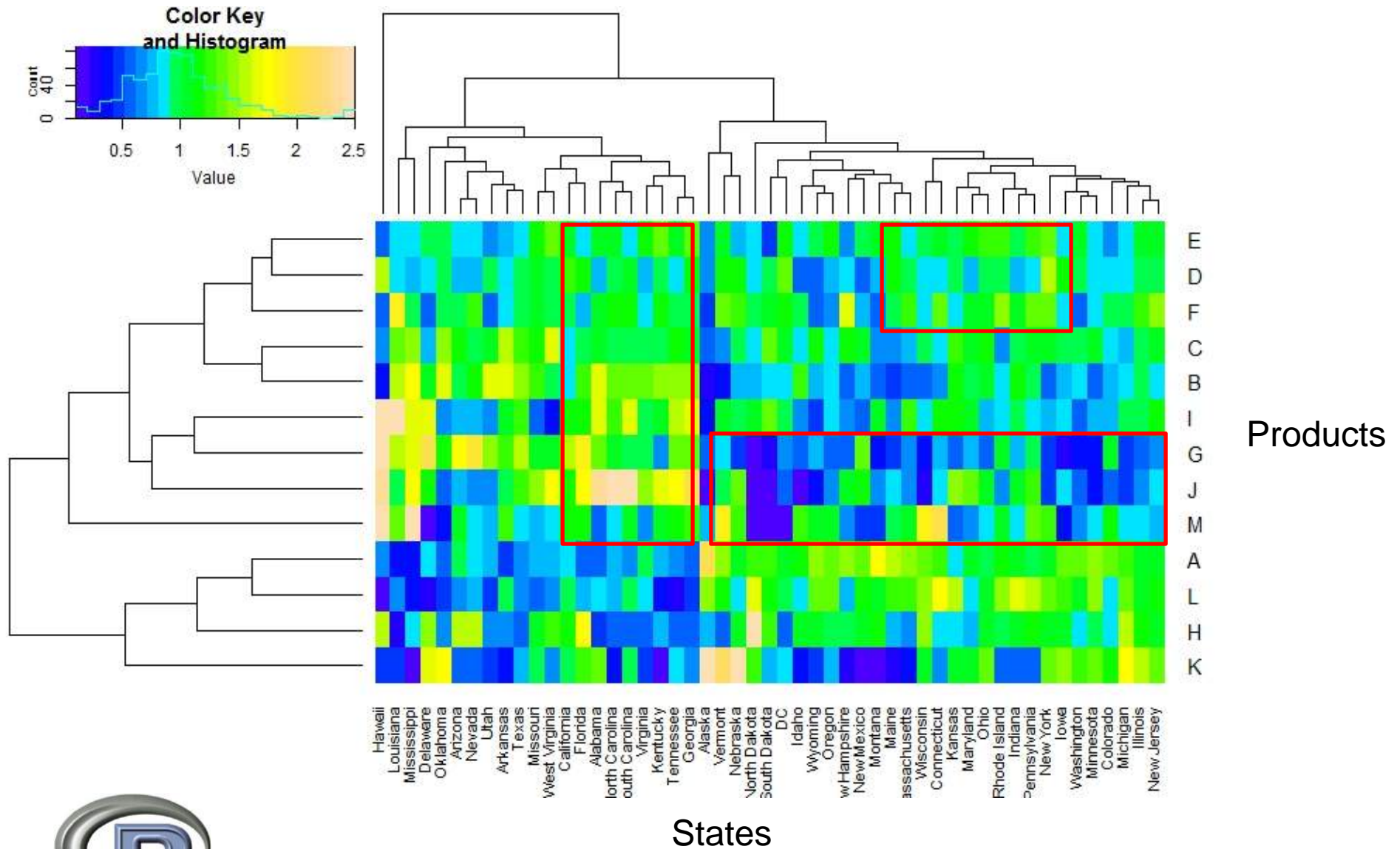
States



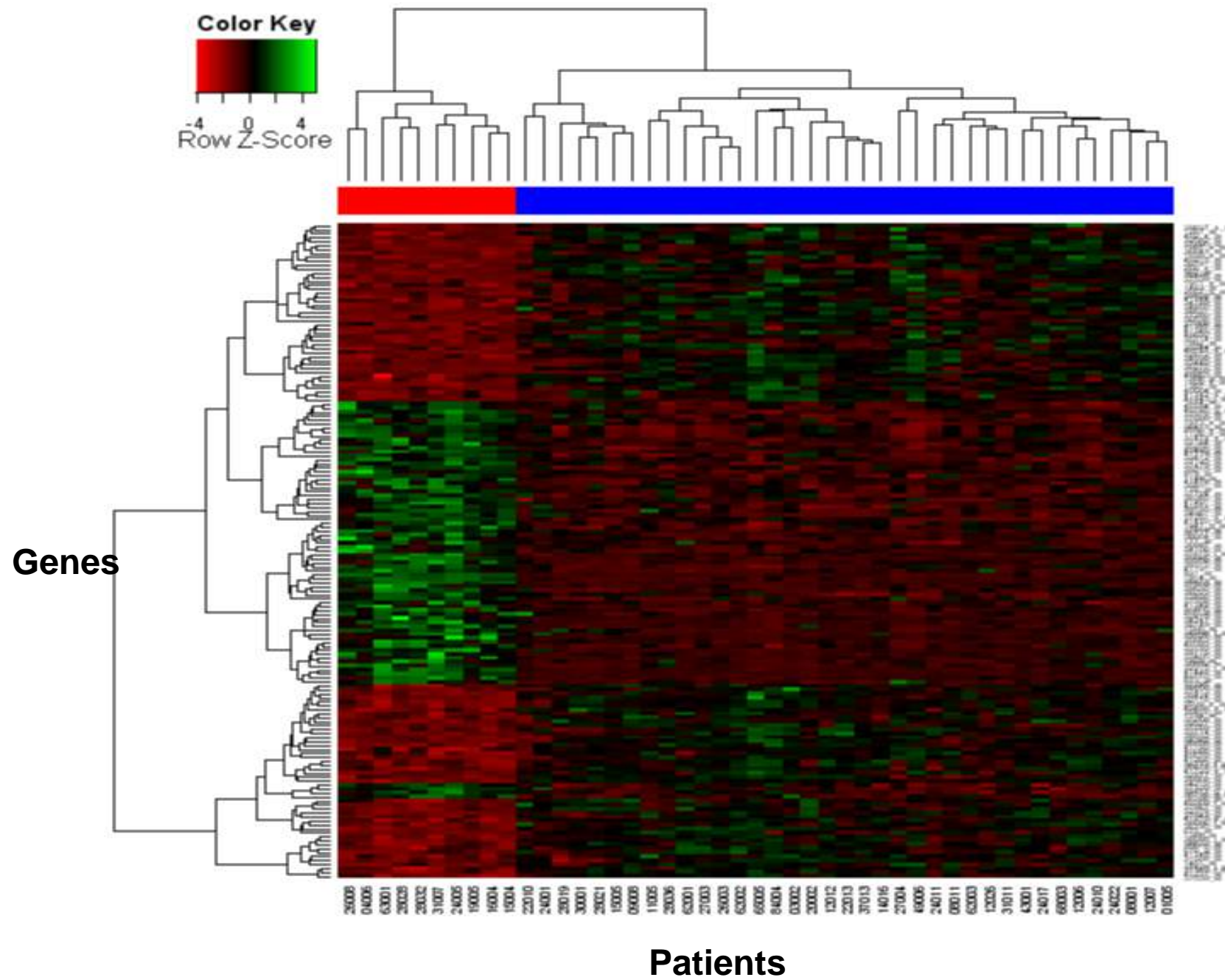
# Bi-clustering States and Products



# Bi-clustering States and Products









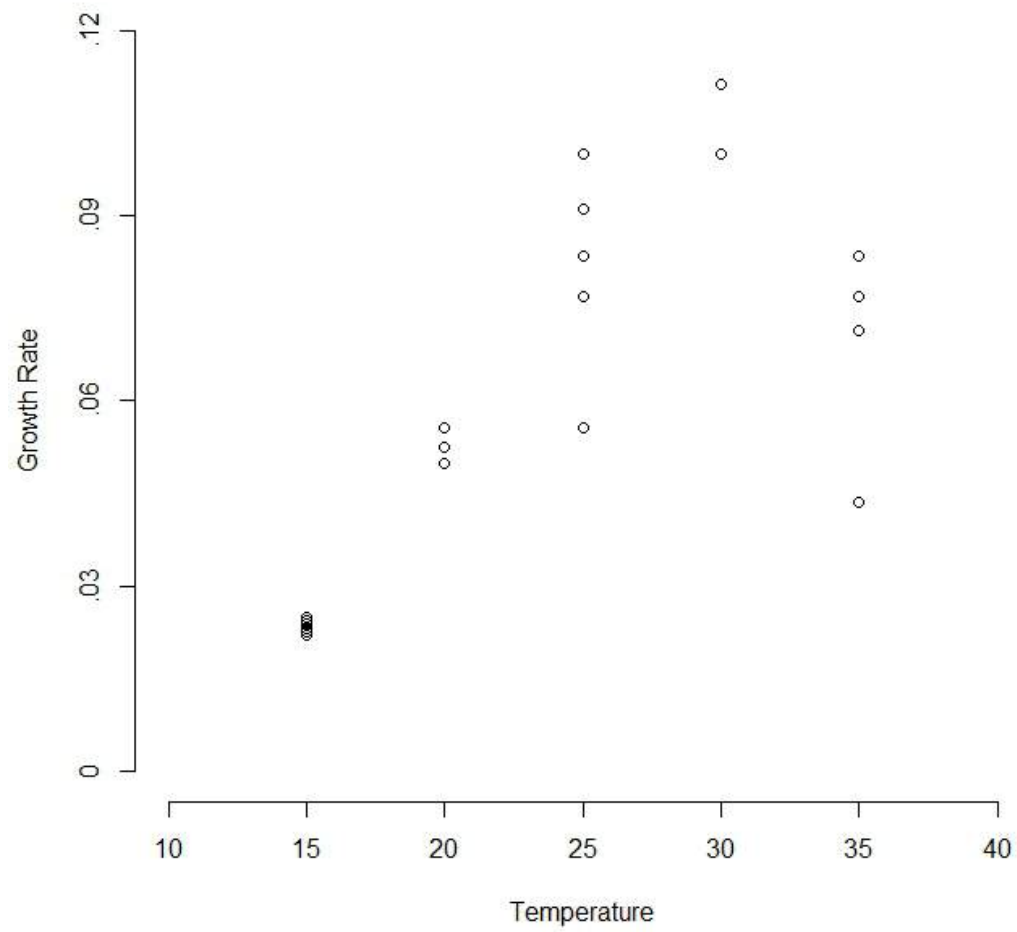
# Non-linear Least Squares

Development time of insects (glassy-winged sharp shooters) from eggs to adults depends on the temperature. Understanding this relationship can help predict outbreaks.

## Experiment

1. Select a few different temperatures
2. Allocate a certain number of eggs at each temperature
3. Record the number of days until the insect reaches adult stage
4. Plot the growth rates (reciprocal of number of days) at each temperature

# Non-linear Least Squares



# Non-linear Least Squares

## Scientific Theory

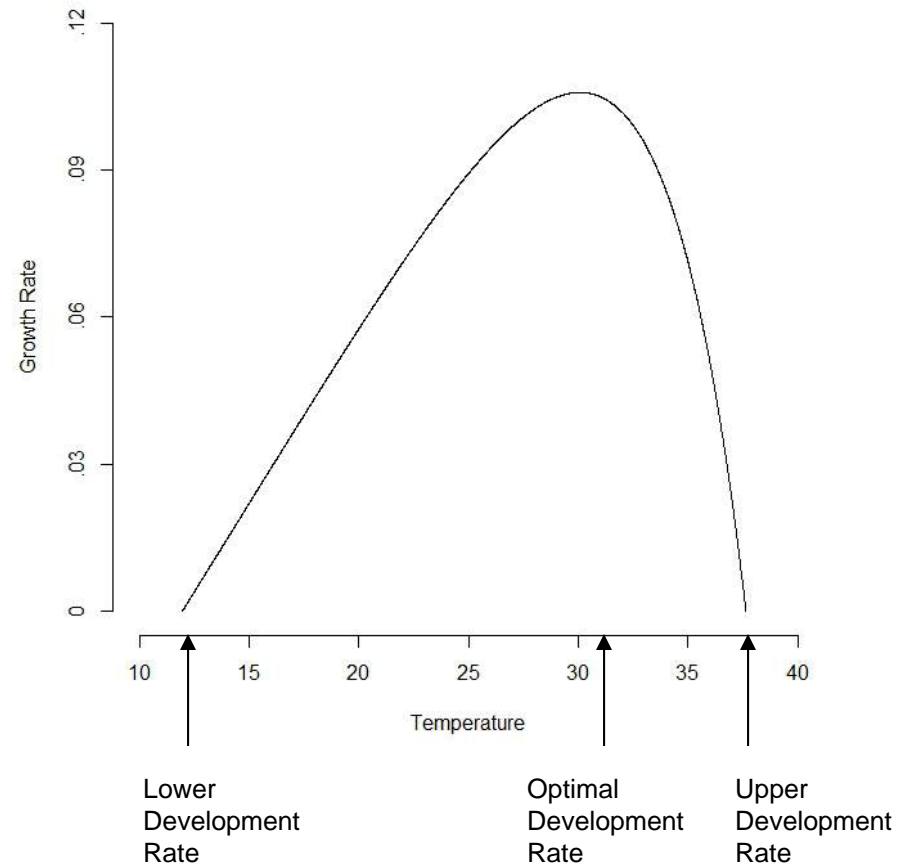
Smooth curves for insect development rates have been modeled by Lactin and Logan (1976, 1995)

$r(T)$  = rate of growth at temperature  $T$

$$= \exp(\rho T) - \exp(\rho T_m - (T_m - T) / \Delta) + \lambda$$

Parameters of the model:

$\rho$ ,  $T_m$ ,  $\Delta$ , and  $\lambda$



# Non-linear Least Squares

## Scientific Theory

Smooth curves for insect development rates have been modeled by Lactin and Logan (1976, 1995)

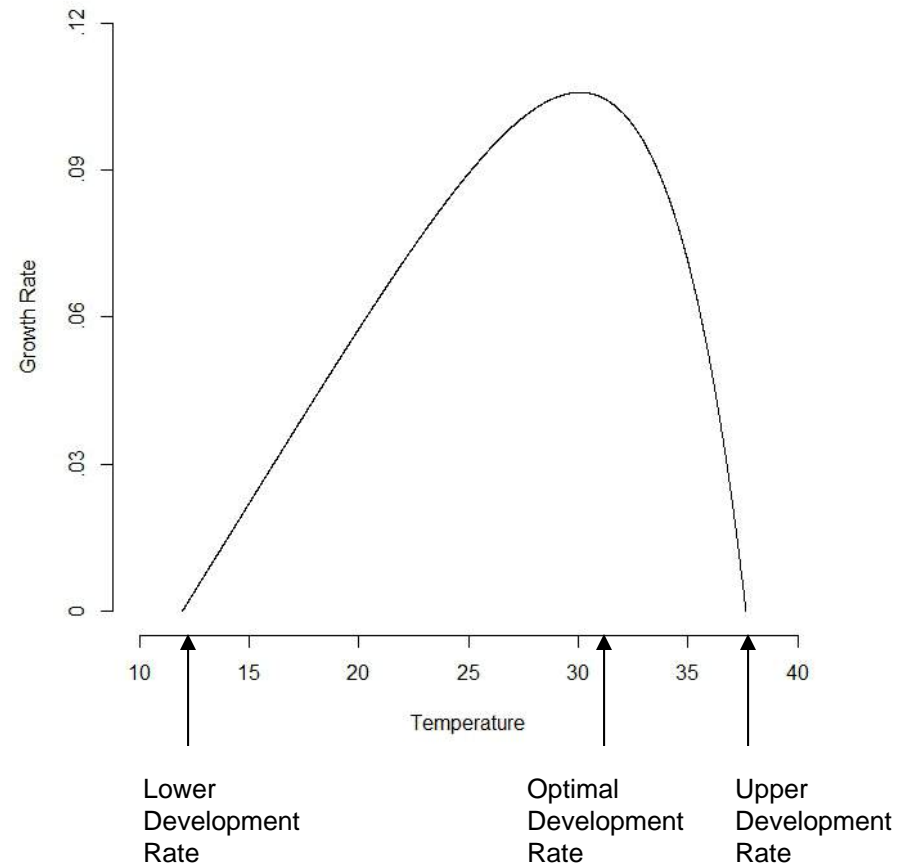
$r(T)$  = rate of growth at temperature  $T$

$$= \exp(\rho T) - \exp(\rho T_m - (T_m - T) / \Delta) + \lambda$$

Parameters of the model:

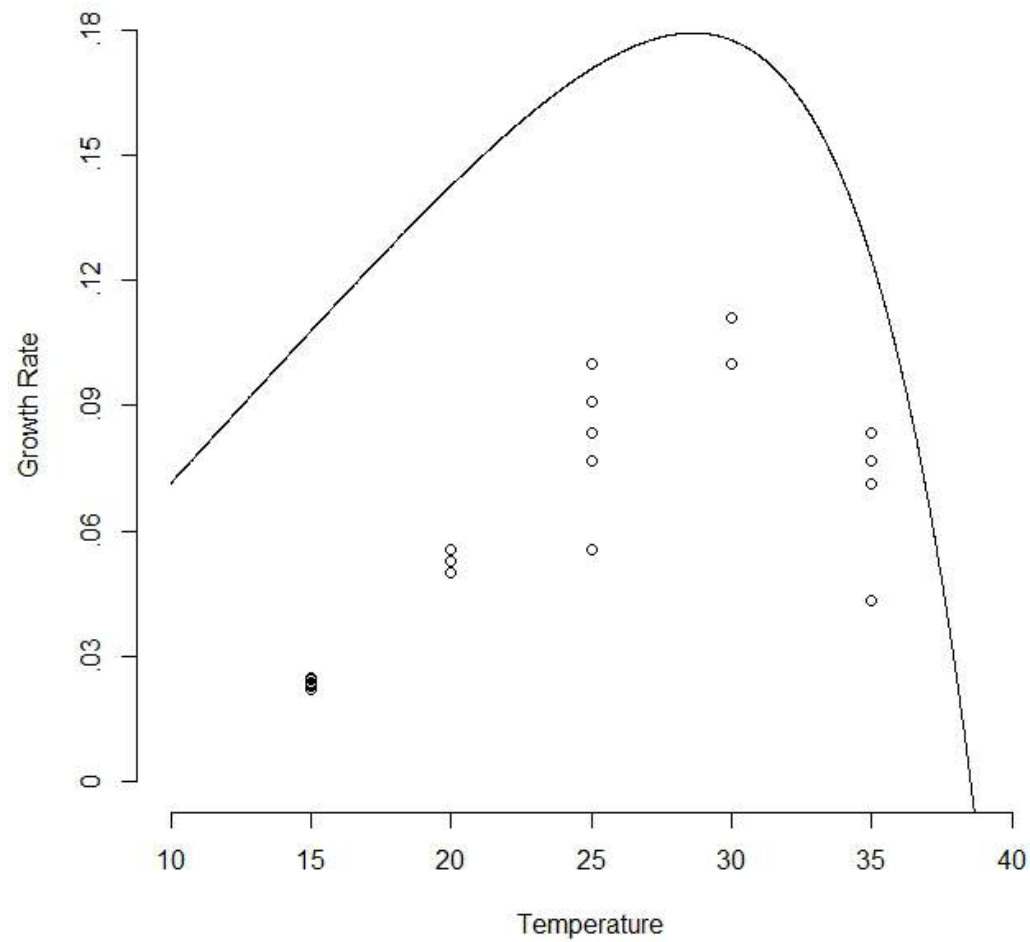
$\rho$ ,  $T_m$ ,  $\Delta$ , and  $\lambda$

Find the values of the parameters that minimizes the sum of squared deviations between observed and predicted growth rates



# Non-linear Least Squares

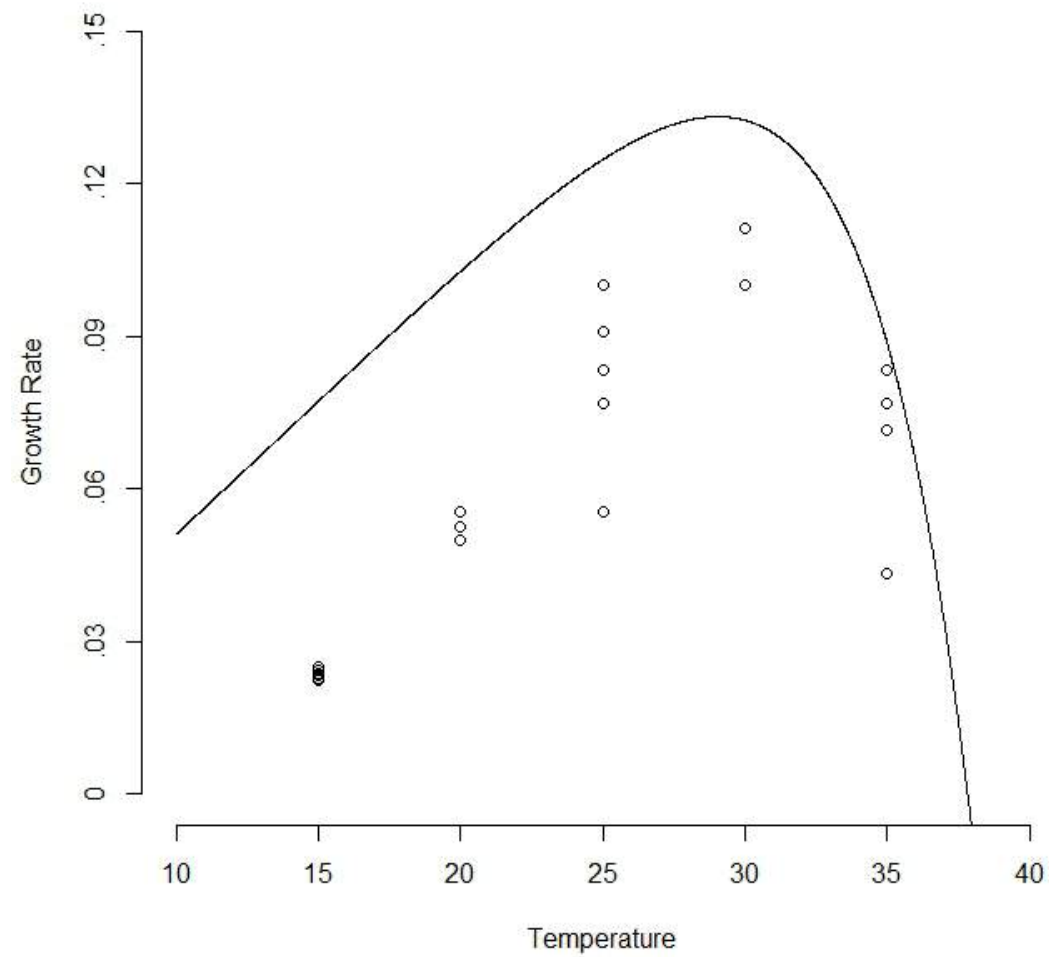
**First Guess** :  $(T_m, \rho, \Delta, \lambda) = (46, .007, 5, 1.0)$





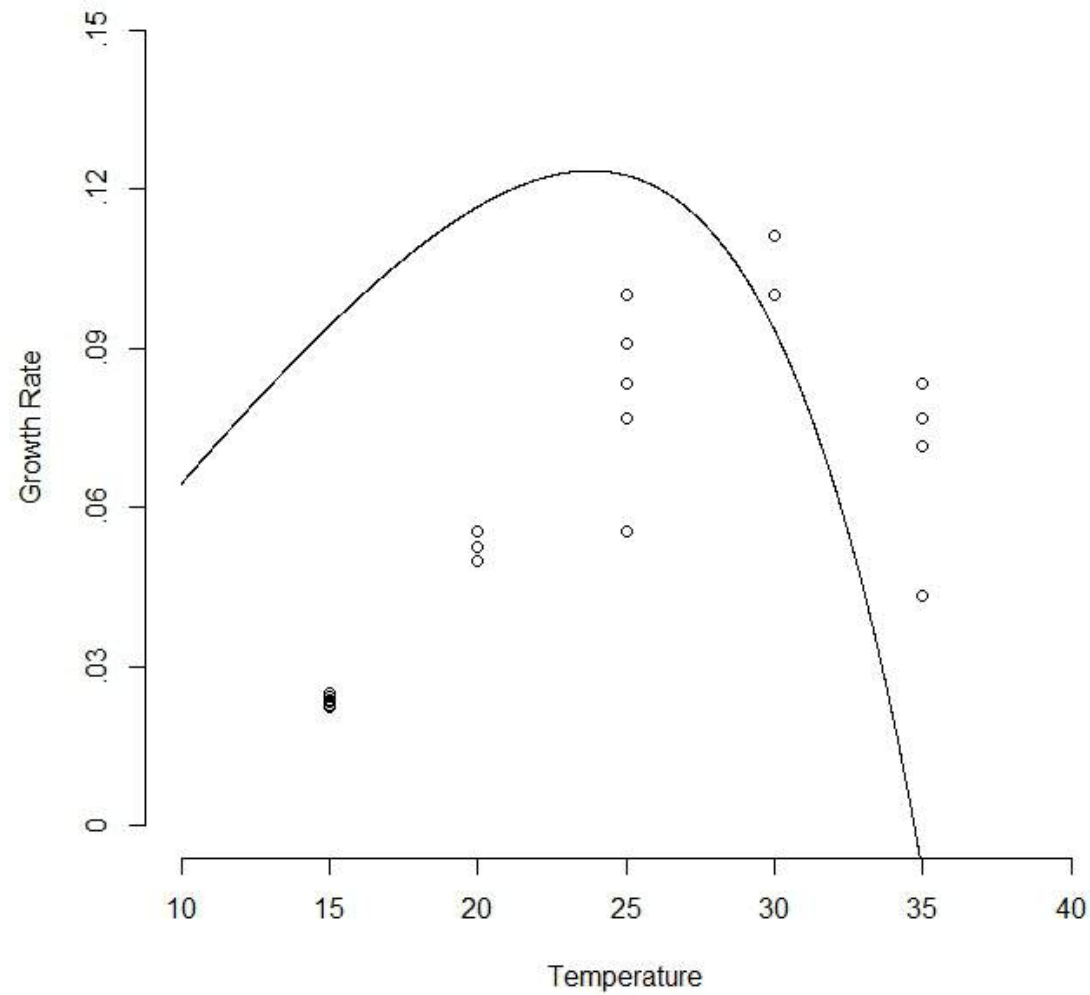
# Non-linear Least Squares

**Second Guess** :  $(T_m, \rho, \Delta, \lambda) = (45, .005, 4, 1.0)$



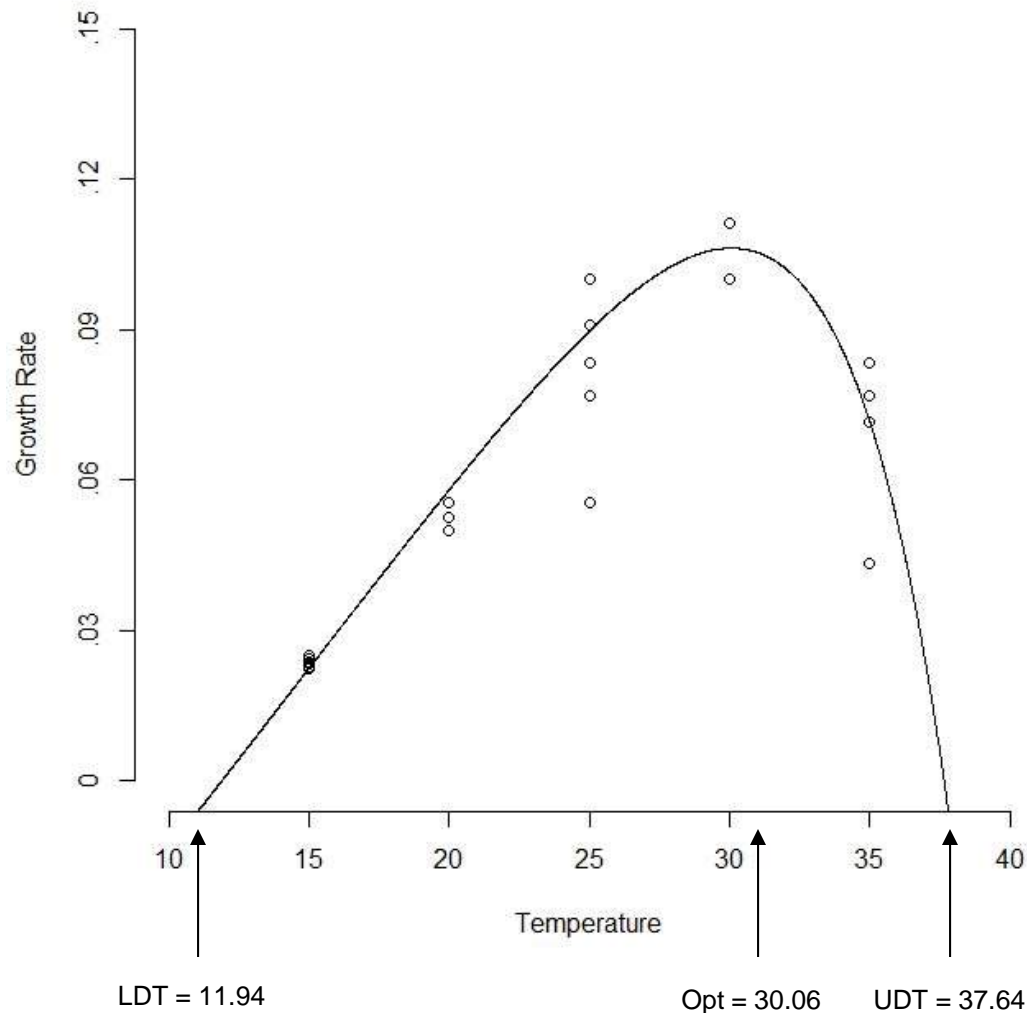
# Non-linear Least Squares

**Third Guess** :  $(T_m, \rho, \Delta, \lambda) = (46, .007, 7, 1.0)$



# Non-linear Least Squares

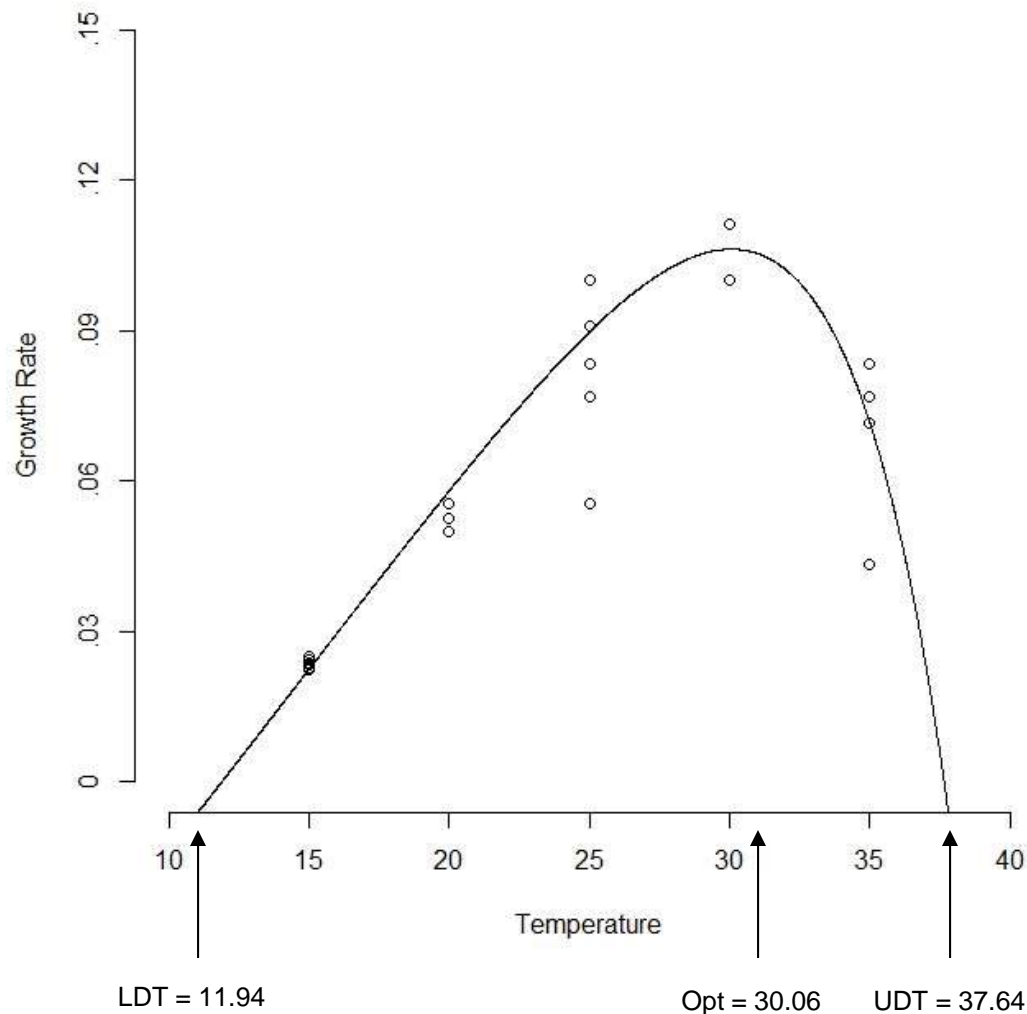
**Optimal Solution** :  $(T_m, \rho, \Delta, \lambda) = (45.83, .00675, 4.33, 1.08)$



```
proc nlin method=marquardt; by type;
model rate=exp(rho*T)-exp(rho*Tm-(Tm-T)/delta)+lambda;
parameters rho = .01 to .05 by .01
              tempM = 20 to 40 by 2
              delta = .5 to 6 by .5
              lambda = -2 to -0.5 by 0.5;
bounds rho T, delta > 0;
run;
```

# Non-linear Least Squares

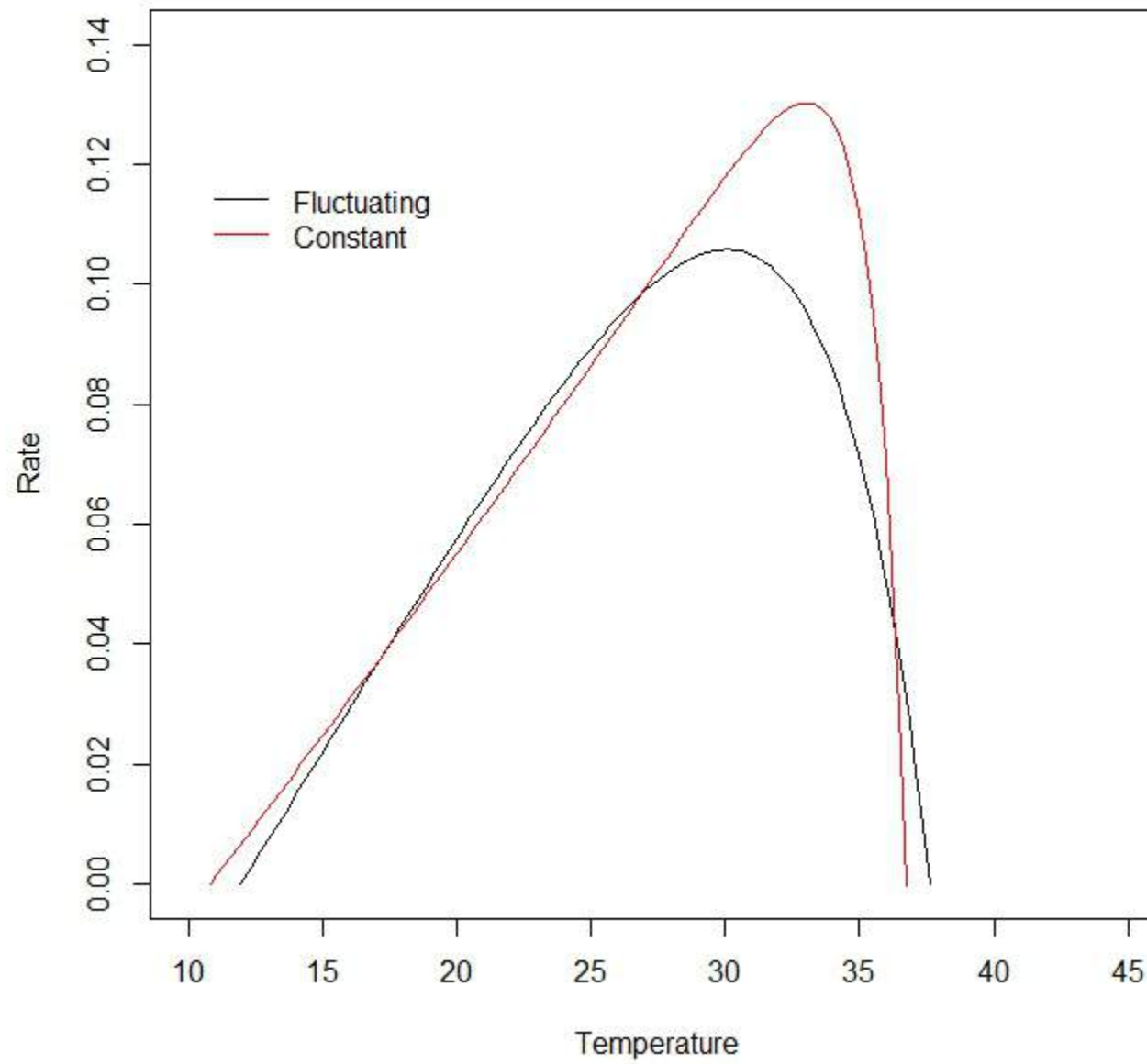
**Optimal Solution** :  $(T_m, \rho, \Delta, \lambda) = (45.83, .00675, 4.33, 1.08)$



```
proc nlin method=marquardt; by type;
model rate=exp(rho*T)-exp(rho*Tm-(Tm-T)/delta)+lambda;
parameters rho = .01 to .05 by .01
              tempM = 20 to 40 by 2
              delta = .5 to 6 by .5
              lambda = -2 to -0.5 by 0.5;
bounds rho T, delta > 0;
run;
```

Formulas for standard errors and confidence intervals for LDT, Opt and UDT are available and should be used in a complete analysis.

### Fluctuating vs. Constant Temperature





# Logistic Regression

## Motivating Example

Nearly 16,000 student athletes were surveyed on a variety of questions.

The following question was used to divide the student athletes into two groups:

*Would you have gone to a 4 year college somewhere even if you could not have been a student athlete?*

*No: 'Athlete Student' group*

*Yes: 'Student Athlete' group*

Interest is in studying the perceptions of academic experiences between these two groups.

# Logistic Regression

Four additional survey questions, each viewed as dependent variables, are used to illustrate the use of logistic regression.

(Responses are YES/NO)

1. Have you been involved, or do you plan to be involved, with work on a research project with a faculty member during your college experience?
2. Have you, or do you plan to, work with a faculty member on independent study during your college experience?
3. Have you been involved, or do you plan to be involved, with an internship during your college experience?
4. Have you, or do you plan to, study abroad during your college experience?

## Simple Descriptive Analysis

Calculate the % of YES responses for each group. Is the difference statistically significant?

# Logistic Regression

## Linear Regression Model

$$E(Y) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

$$x_1 = \begin{cases} 1 & \text{if athlete student group} \\ 0 & \text{if student athlete group} \end{cases}$$

$\beta_1$  measures the change in the mean value of  $Y$  for athlete students

## Logistic Regression Model

$$P(\text{YES}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$$

$\beta_1$  reflects the change in the probability of YES for athlete students

# Logistic Regression

## Logistic Regression Model

$$P(\text{YES}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$$

$\beta_1$  reflects the change in the probability of YES for athlete students

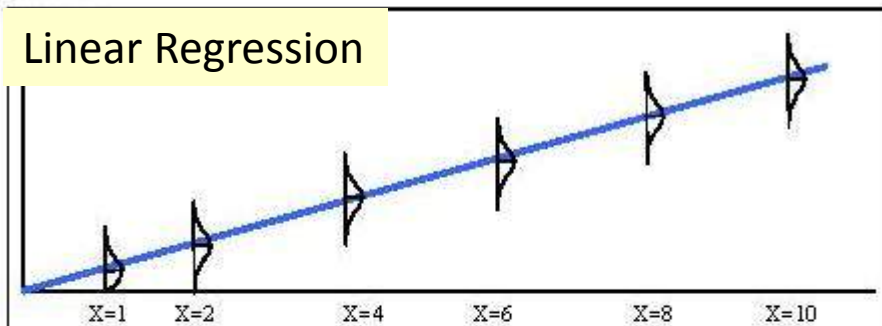
Activity	95% Confidence Interval for $\beta_1$	Interpretation
Research	(-.17 , .15)	No difference between the two groups
Internship	(-.25 , .035)	No difference between the two groups
Study Abroad	(.01 , .39)	Athlete students more likely to study abroad
Independent Study	(-.47 , -.02)	Student athletes more likely to engage in independent study

# Poisson Regression

Examples include number of cells, number of seizures, number of hospital visits, etc. Some of these variables can be variables measured in a clinical trial.

Poisson Regression is an example of using 'Generalized Linear Models.'

Linear Regression



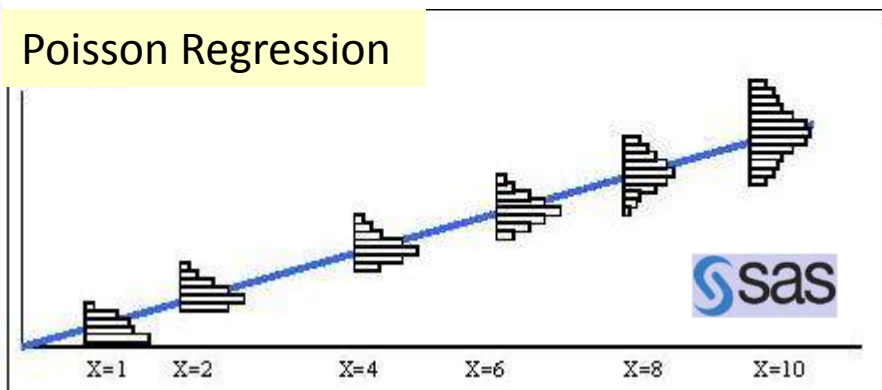
$$Y = \beta_0 + \beta_1 x + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2)$$

Least squares fitting is OK

All statistics packages

Even Excel can do it!

Poisson Regression



$$Y \sim \text{Poisson}(e^{\beta_0 + \beta_1 x})$$

Least squares fitting not the best way

Maximum likelihood fitting is preferred

SAS software PROC GENMOD

GENMOD can do negative binomial

regression too

## Comparing Two Groups – How many samples?

$$H_0 : \mu_2 - \mu_1 \leq 0$$

$$H_1 : \mu_2 - \mu_1 > 0$$

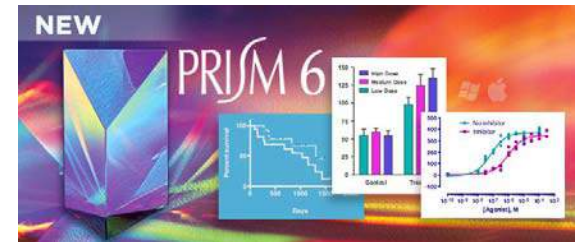
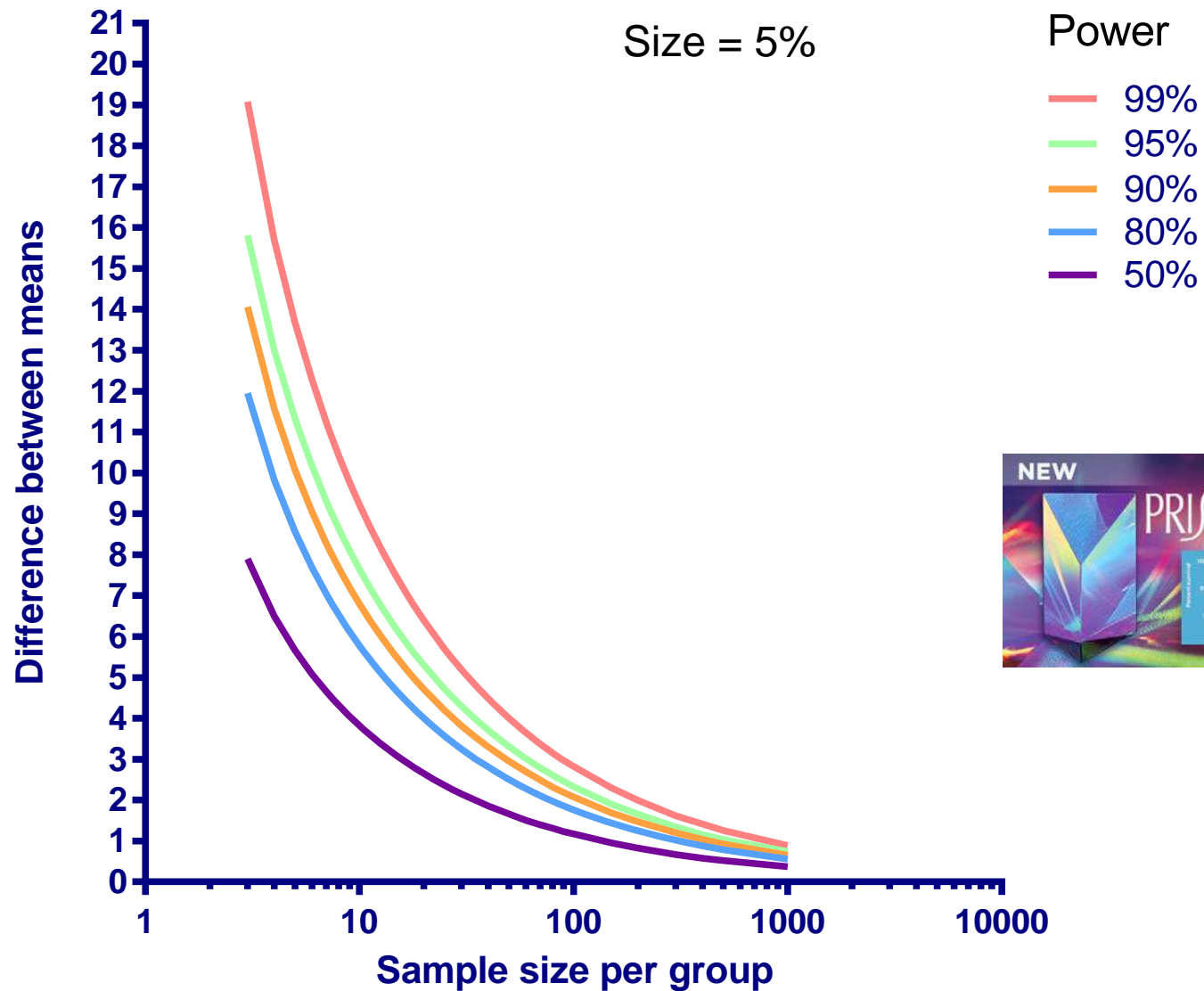
We should reject when  $\bar{X}_2 - \bar{X}_1$  is too large

How to choose the sample size  $n$  for each group?

Control Two Probabilities:

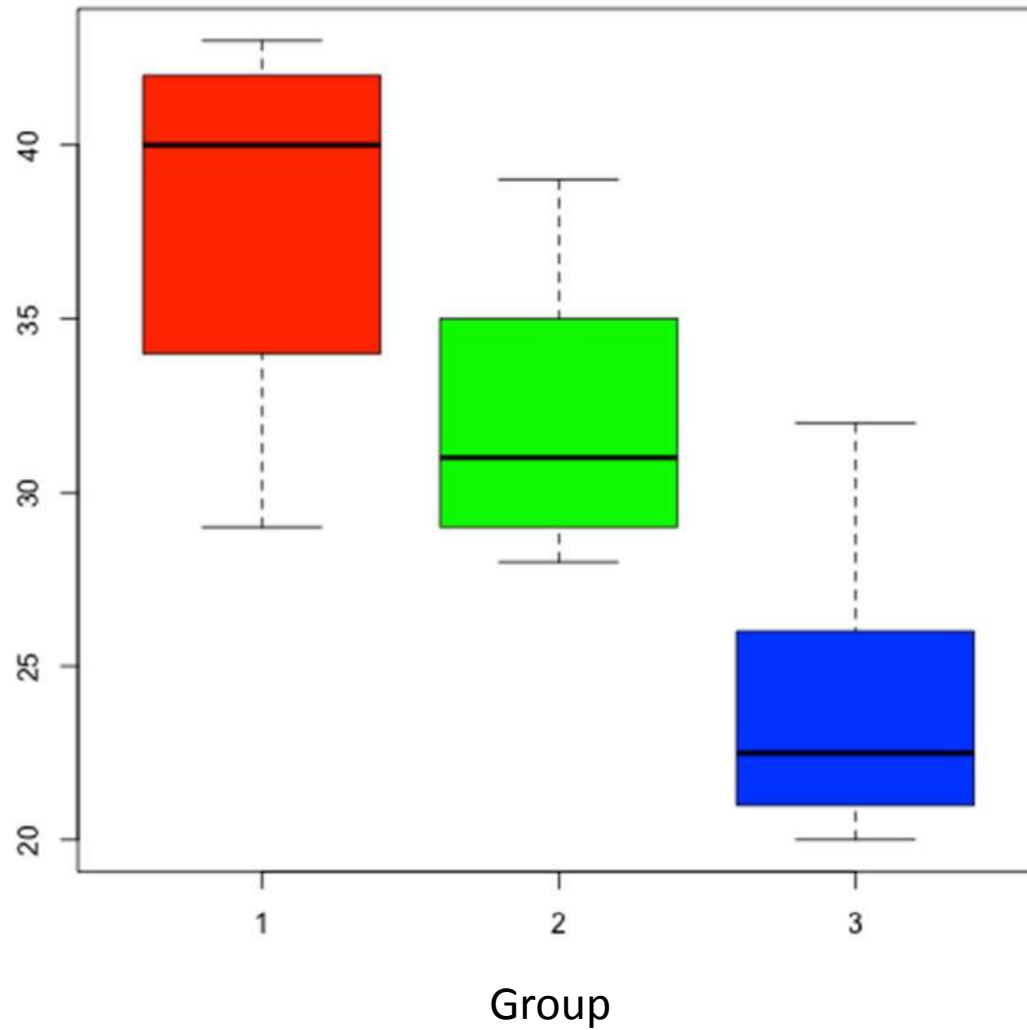
1. (Size)  $\text{Prob}(\text{Reject } H_0 \mid \mu_2 - \mu_1 = 0) = \alpha$
2. (Power)  $\text{Prob}(\text{Reject } H_0 \mid \mu_2 - \mu_1 = \delta) = 1 - \beta$

# Comparing Two Groups – How many samples?





# Comparing More Than Two Groups



Not a good idea to do all pairwise or t-tests

A better analysis is an analysis-of-variance table

# Comparing More Than Two Groups



	A	B	C	D	E	F	G	H	I	J	K
1	One-way ANOVA										
2											
3	Group 1	Group 2	Group 3		ANOVA: Single Factor						
4	13	12	7								
5	17	8	19		DESCRIPTION						
6	19	6	15		Groups	Count	Sum	Mean	Variance	SS	
7	11	16	14		Group 1	10	150	15	13.33333	120	
8	20	12	10		Group 2	10	111	11.1	18.76667	168.9	
9	15	14	16		Group 3	10	135	13.5	14.05556	126.5	
10	18	10	18								
11	9	18	11		ANOVA		Alpha		0.05		
12	12	4	14		Sources	SS	df	MS	F	P value	F crit
13	16	11	11		Between Groups	77.4	2	38.7	2.515407	0.099596	3.354131
14					Within Groups	415.4	27	15.38519			
15					Total	492.8	29	16.9931			

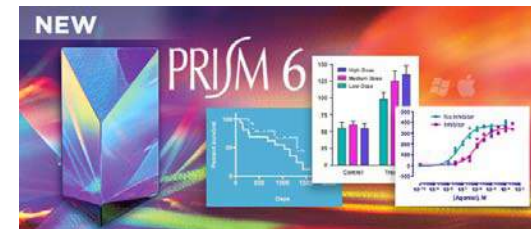
If the F-test is significant then sometimes follow-up pairwise t-tests are justified.

BUT, multiple comparison problem should usually be addressed.

# Comparing More Than Two Groups

ANOVA summary

F	2.515
P value	0.0996
P value summary	ns
Are differences among means statistically significant? ( $P < 0.05$ )	No
R square	0.1571



Bartlett's test	
Bartlett's statistic (corrected)	0.2989
P value	0.8612
P value summary	ns
Significantly different standard deviations? ( $P < 0.05$ )	No

ANOVA table	SS	DF	MS	F (2, 27)	P value
Treatment (between columns)	77.40	2	38.70	2.515	P = 0.0996
Residual (within columns)	415.4	27	15.39		
Total	492.8	29			

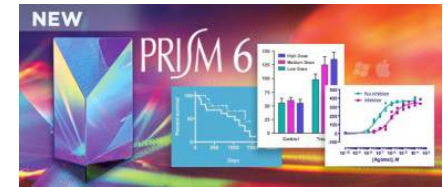
# Multiple Comparisons



1. K=5 groups implies 10 pairwise comparisons
2. Specialized methods for 'each versus one' situations (Dunnett)
3. Specialized methods for balanced designs (Tukey)
4. General procedures for controlling False Discovery Rate (e.g., Scheffee, Bonferonni, FDR)

# Multiple Comparisons

Group 1: Control  
Group 2 : Treatment  
Group 3: Treatment + Antagonist



Number of families	1				
Number of comparisons per family	2				
Alpha	0.05				
Dunnett's multiple comparisons test	Mean Diff.	95% CI of diff.	Significant?	Summary	
Control vs. Treated	-38.33	-53.61 to -23.06	Yes	****	
Control vs. Treated+Antagonist	-3.500	-18.07 to 11.07	No	ns	

Number of families	1				
Number of comparisons per family	3				
Alpha	0.05				
Tukey's multiple comparisons test	Mean Diff.	95% CI of diff.	Significant?	Summary	
Control vs. Treated	-38.33	-54.61 to -22.06	Yes	****	
Control vs. Treated+Antagonist	-3.500	-19.02 to 12.02	No	ns	
Treated vs. Treated+Antagonist	34.83	18.56 to 51.11	Yes	***	

Conclusions about the pairwise comparisons are based on adjusted t-tests.

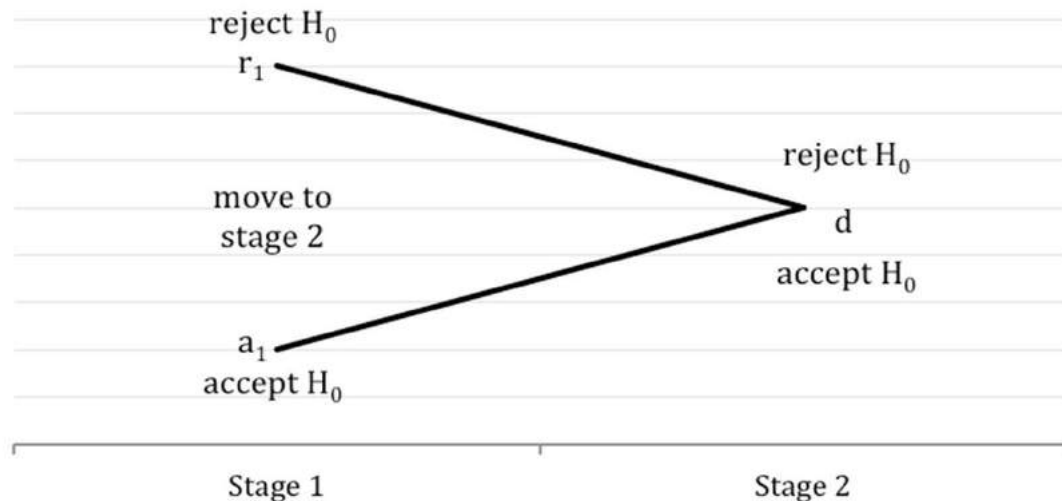
The adjustment depends on the number of pairwise comparisons that are being made.

# Sequential Clinical Trials

Multiple looks at the data, rather than just one look at the end of the study.

Potential for early stopping (i.e., fewer patients utilized in the study) when statistically credible evidence of either no difference or detectable difference comes early.

Ethical and cost advantages compared to fixed sample size trials.



$$H_0 : \mu_T \leq \mu_C$$

$$H_1 : \mu_T > \mu_C$$

Size  $\alpha$  test and

Power  $1 - \beta$  for  $\mu_T - \mu_C = \delta$

m patients allocated to each treatment in stage 1.

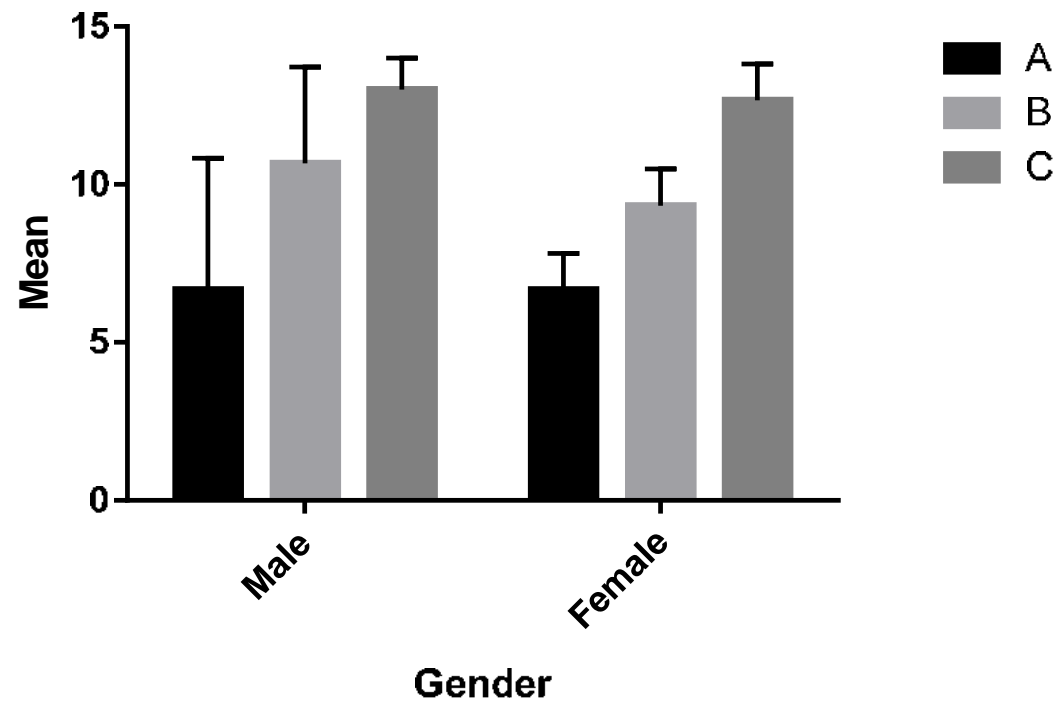
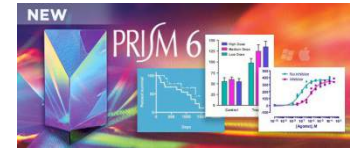
If necessary, m additional patients allocated to each treatment in stage 2.

## Two-Way ANOVA

There is often an additional factor(s) that could be of interest itself, but often is a “nuisance” factor that needs to be accounted for in order to make unbiased inference about the primary factor of interest

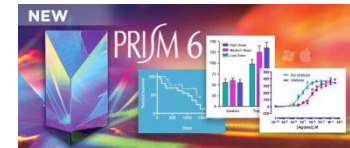
		Treatment		
		A	B	C
Gender	Male	10	14	12
		8	6	16
		2	10	14
	Female	6	8	12
		8	10	14
		6	10	12

# Two-Way ANOVA





# Two-Way ANOVA



Two-way ANOVA  
Alpha 0.05

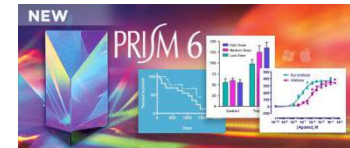
Source of Variation	% of total variation	P value	P value summary	Significant?
Interaction	0.8002	0.8734	ns	No
Gender	0.7695	0.6173	ns	No
Treatment	63.34	0.0021	**	Yes

ANOVA table	SS	DF	MS	F	P value
Interaction	1.444	2	0.7222	F (2, 12) = 0.1368	P = 0.8734
Gender	1.389	1	1.389	F (1, 12) = 0.2632	P = 0.6173
Treatment	114.3	2	57.17	F (2, 12) = 10.83	P = 0.0021
Residual	63.33	12	5.278		

# Two-Way ANOVA

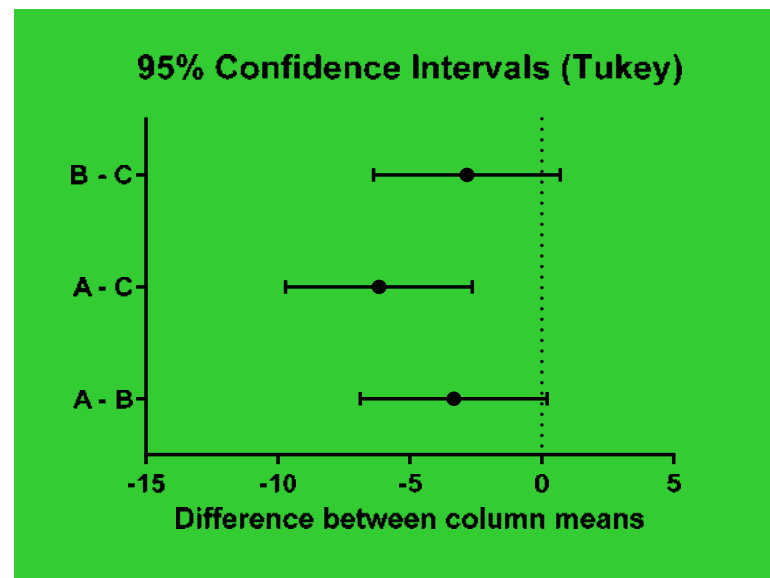
Number of families  
Number of comparisons per family  
Alpha

1  
3  
0.05

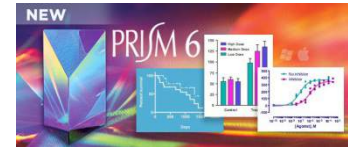


## Tukey's multiple comparisons test

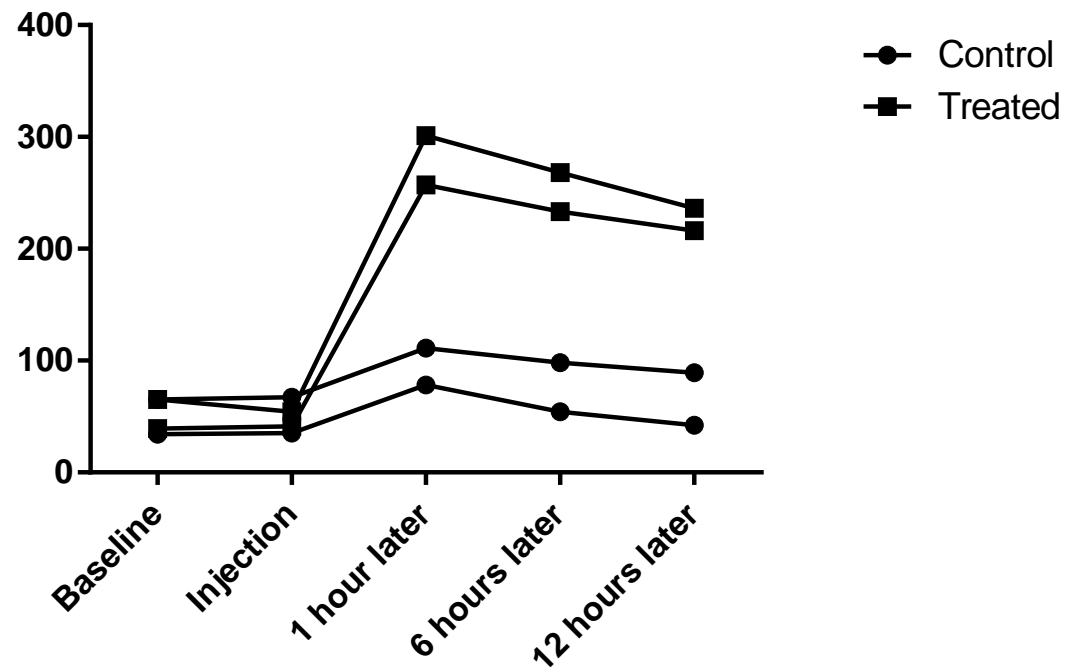
	Mean Diff.	95% CI of diff.	Significant?	Summary
A vs. B	-3.333	-6.872 to 0.2052	No	ns
A vs. C	-6.167	-9.705 to -2.628	Yes	**
B vs. C	-2.833	-6.372 to 0.7052	No	ns



# Longitudinal Data



**Animal Profiles**

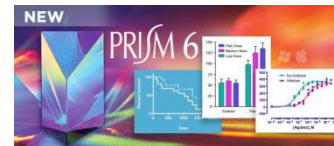


Repeated measurements on the same animal are correlated. Analyses must take this into consideration.

# Longitudinal Data

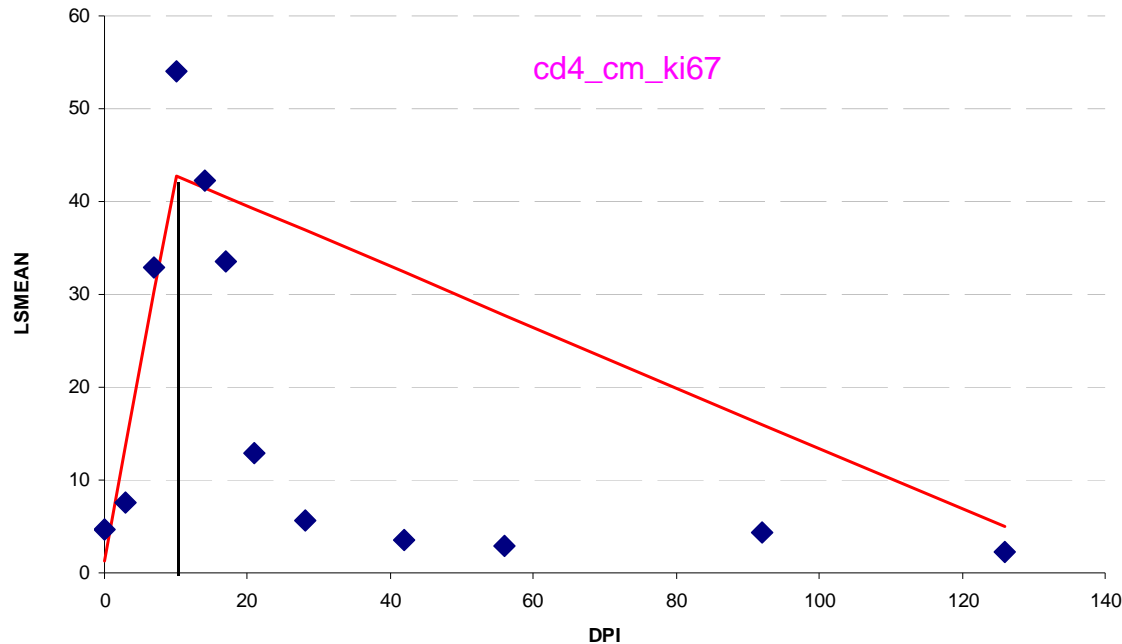
Two-way RM ANOVA  
Alpha

Matching: Stacked  
0.05



ANOVA table	SS	DF	MS	F	P value
Interaction	36501	4	9125	$F(4, 8) = 177.5$	$P < 0.0001$
Time	67147	4	16787	$F(4, 8) = 326.6$	$P < 0.0001$
Group	53768	1	53768	$F(1, 2) = 19.91$	$P = 0.0467$
Subjects (matching)	5401	2	2701	$F(2, 8) = 52.54$	$P < 0.0001$
Residual	411.2	8	51.40		

# Longitudinal Data – Peak Analysis



Peak Estimate: 7.00

90% Confidence Interval: (0.33 , 13.65)

```
proc nlin method=marquardt maxiter=100 converge=1e-3 data=cd4cmki67 alpha=.1;
parameters CP = 5 to 10 by 1
            beta0=-2 to 2 by .5
            beta1=2 to 6 by 1
            beta2=-10 to 0 by 1;
bounds CP>0;
if (dpi <= CP) then
mean = beta0 + beta1*dpi;
else mean = beta0 + beta1*dpi +beta2*(dpi-CP);
model estimate = mean;
run;
```



# Prospective Study

Two groups of subjects with differential treatment

Each patient followed over some period of time (length of study)

Outcome is the realization, or not, of a certain event (e.g., illness, death)

Example (New England Journal of Medicine, 1988, p. 262-264)

	Myocaridal Infarction		No Myocaridal Infarction	Total
Placebo	189		10845	11034
Aspirin	104		10933	11037

$$\text{Relative Risk} = \frac{\text{Pr ( MI| Placebo )}}{\text{Pr ( MI| Aspirin )}}$$

Strength of association

Relative Risk

95% confidence interval



1.818

1.433 to 2.306

# Retrospective Study

Two groups of subjects with presence and absence of an outcome

Each patient researched backwards in time for presence or absence of an exposure

Outcome is the realization, or not, of being exposed

Example (British Medical Journal, 1950, p. 739-748)

		Lung Cancer		No Lung Cancer
	Smoked	688		650
	Never Smoked	21		59
Total		709		709

Design does not allow for estimation of Relative Risk because the wrong set of marginal totals is fixed.

But we CAN estimate the odds ratio.

# Odds Ratio

		Lung Cancer		No Lung Cancer
	Smoked	688		650
	Never Smoked	21		59
Total		709		709

$$\text{Odds of Smoking for Lung Cancer Patients} = \frac{688/709}{21/709} = 32.76$$

$$\text{Odds of Smoking for NO Lung Cancer Patients} = \frac{650/709}{59/709} = 11.02$$

$$\text{Odds Ratio: } 32.76/11.02 = 2.97$$

Strength of association

Relative Risk

1.959

95% confidence interval

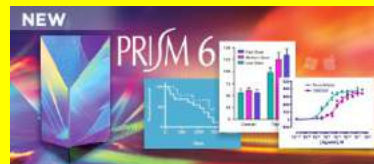
1.352 to 2.839

Odds ratio

2.974

95% confidence interval

1.787 to 4.950





# Odds Ratio

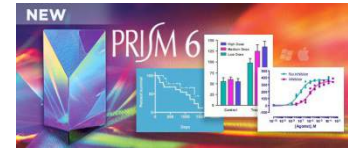
		Lung Cancer		No Lung Cancer
	Smoked	688		650
	Never Smoked	21		59
Total		709		709

Interpretation: The odds that a lung cancer patient smoked is 3 times higher than the odds a NO lung cancer patient smoked.

**Equivalent:** The odds that a smoker developed lung cancer is 3 times higher than the odds a non-smoker developed lung cancer.

In general RR is not recoverable from Odds Ratio. But if the prevalence of the outcome event (Lung Cancer in this example) is small, the two will be close to each other.

# Survival Analysis



Response variable is time to a specific event.

- Time recovery
- Time to relapse
- Time to death

Exact times to the event may not be measured completely

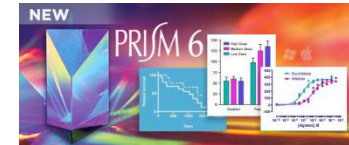
- Patients are 'lost to follow-up'
- Study is terminated before all patients encounter the event

Example: (Days to Relapse)

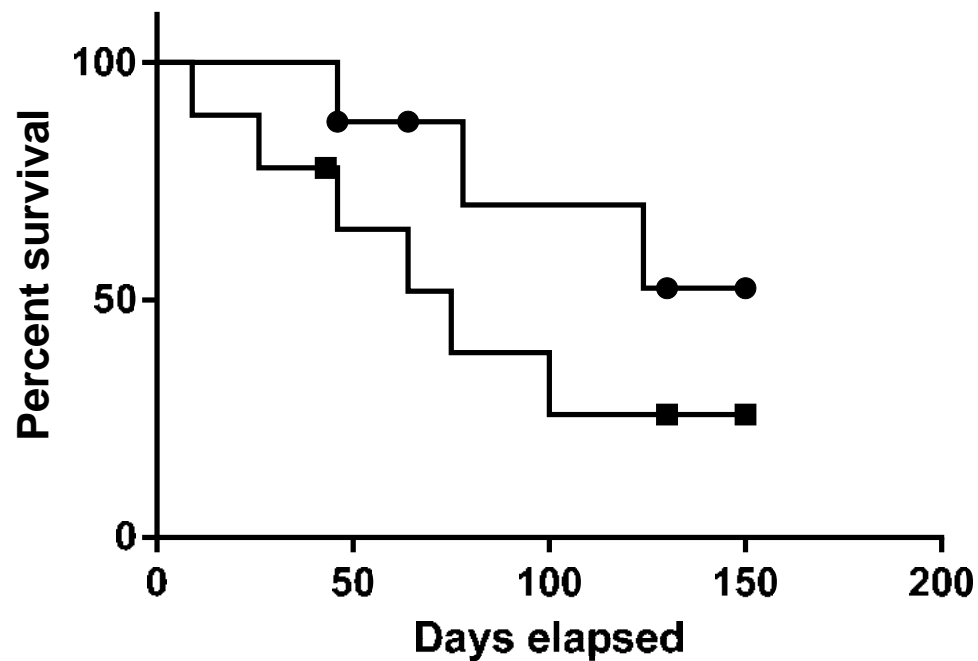
Control Group: 46, 46+, 64+, 78, 124, 130+, 150+, 150+

Treated Group: 9, 26, 43+, 46, 64, 75, 100, 130+, ,150+

# Survival Analysis



## Survival proportions: Survival of Two groups



● Treated  
■ Control

### Comparison of Survival Curves

Log-rank (Mantel-Cox) test

Chi square 2.010

df 1

P value 0.1563

P value summary ns

Are the survival curves sig different? No

Kaplan-Meier Survival Curves

**The End**