Cluster Analysis of Himachal Tomato

Girja Sharan, Professor
T Madhavan, Professor
Kishor Rawale, Academic Associate

Centre for Management in Agriculture
Indian Institute of Management, Ahmedabad

Abstract

A crate of Himachal tomato was obtained from Azad Mandi, Delhi. It contained 252 fruits. Each fruit was weighed and its axial dimension measured. Data of all 252 fruits was then subjected to cluster analysis, using weight and axial dimension separately as basis. The tool of *Cluster Analysis* enables us to divide the sample in groups that are relatively homogeneous in size on the basis of weight or axial dimensions, whichever is desired. Analysis also yields mass proportion of tomato contained in each group as also the number. The utility of cluster analysis lies in the fact that it can indicate how many homogenous groups can be made of a lot of ungraded produce in advance and what will be the physical characteristics of produce in each group. This will be useful to those designing size graders for tomato, other fruits and vegetables.

Introduction

Consumers in large cities have begun to show preference for clean, fresh and well-graded produce. Our attention was drawn to this while developing packaging boxes for the tomato growers of Himachal who send produce to *Azadpur Mandi* in Delhi for sale.**[1]** Design of size grader required data on the physical dimensions and weight of tomato. Such data was not readily found in literature. Accordingly, we made measurements of weight and physical dimensions of tomato grown in the region. Distribution underlying the weight and axial length were determined. Subsequently, cluster-analysis was used to determine the number of size-grades that may be made based on weight or axial dimension. Initially, the results of analysis based on weight are presented. It is then followed by analysis based on axial dimension.

Sample

One crate of Himachal tomato was purchased from Azadpur Mandi, Delhi on July 12 2002. The crate contained 252 tomatoes. Axial dimensions of each tomato were measured using digital Vernier Caliper. This included the longitudinal axis and two horizontal axes. Each tomato was weighed on digital balance. Shape of the tomato can be called `**elliptical regular**.' The cultivar was `**Safal-99**.'

Distribution underlying Weight

Weight varied from 26 g to 124 g. Weight data was transformed by subtracting the minimum (26 g) from actual for all 252 observations.

W actual weight

X = W - 26

Table-1 shows the frequency of classes based on X. **Figure-1** shows the data graphically. Shape of the graph suggested the possibility of Weibulll being the underlying distribution **[1]**. Weibull density function is

$$f(x;\alpha,\beta) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} \qquad x \geq 0 \qquad (1)$$

α, β parameters

Mean (μ) = $\beta\, \Gamma\, (1 + 1/\, \alpha)$

Variance (σ^2) = $\beta^2\, \{ \Gamma\, (1 + 2/\, \alpha)$ -[$\Gamma\, (1 + 1/\, \alpha)]^2\}$

Estimates of mean and variance of, X, from sample data are

$\overline{x} = 29$

$S_x^2 = 310$

Using the method of moments, α and β were estimated

$\alpha = 1.708$

$\beta = 32.7$

The particular distribution thus is

$$f(x) = 0.00442\ x^{0.708}\ e - (x/32.7)^{1.708} \qquad (2)$$

Cumulative density function of Weibull

$$F(x; \alpha, \beta) = 1 - e^{-(x/\beta)^{\alpha}} \qquad x \geq 0 \qquad (3)$$

Table-1 and **Figure-1** also show the expected relative frequency using (3). Goodness of fit was tested by Chi square. The computed χ^2 (4.462) is less than the tabulated $\chi^2_{0.05,3}$ (12.837), which indicates Weibull provides a good description of the weight data.

Cluster Analysis using Weight as a basis

The goal in cluster analysis **[2]** is to divide a set of given objects in to a desired number of clusters such that the objects in a cluster are relatively more homogeneous. In other words, the division is so made that within-cluster variation is less than between-cluster variation. The two popular clustering techniques are Hierarchical and Partition technique. Partition technique allows reallocation of objects if their initial allocation was inaccurate. The use of partitioning techniques usually assumes that the number of the final clusters is known and specified in advance. Partitioning technique includes, ' **K-Means**' clustering, in which one of the similarity measures used is Euclidean distance between individuals.

Computationally, this method can be called as analysis of variance (ANOVA) "in reverse." The program will start with k random clusters, and then move objects between those clusters with the objective so as to

(1) minimize variability within clusters and

(2) maximize variability between clusters.

This is analogous to "ANOVA in reverse" in the sense that the significance test in ANOVA evaluates the between group variability against the within-group variability when computing the significance test for the hypothesis that the means in the groups are different from each other. In k-means clustering, the program tries to move objects (e.g., cases) in and out of groups (clusters) to get the most significant ANOVA results. Usually, as the result of a k-means clustering analysis, we would examine the means for each cluster on each dimension to assess how distinct our k clusters are. Ideally, we would obtain very different means for most, if not all dimensions, used in the analysis. The magnitude of the F values from the analysis of variance performed on each dimension is another indication of how well the respective dimension discriminates between clusters.
For our data, the analysis has been carried out for two, three and four clusters, using SYSTAT package.

Two Clusters

We use the weight for the purpose of clustering. The procedure divides the whole lot in two clusters as specified (**Table-2**). It puts 184 cases (out of 252) in the **first cluster**. The mean weight of tomato in this cluster is 47 gm. Tomato of this size makes up 62 per cent of the ungraded lot by weight. **Second cluster** contains 68 cases. The mean weight of tomato in this cluster is 78 gm. Tomato of this size makes up 38 per cent of the ungraded lot by weight. Table-2 also shows the mean axes length of tomato in each cluster. **Table-3** shows the results of analysis of variance. Since the computed 'F' is much higher than the tabulated value, we can conclude that the two clusters are different from each other.

Three Clusters

The procedure divides the whole lot in three clusters (**Table-4**). It puts 126 cases in the **first** cluster. The mean weight of tomato in this cluster is 42 gm. Tomato of this size makes up 38 per cent of the ungraded lot by weight. **Second** cluster contains 105 cases. The mean weight of tomato in this cluster is 63 gm. Tomatoes of this size make up 48 per cent of the ungraded lot by weight. The remaining 21 cases are in the **third** cluster. The mean weight of tomato in this cluster is 97 gm. Tomatoes of this size make up 14 per cent of the ungraded lot by weight. The table also shows the mean axes length of tomato in each cluster. **Table-5** shows the results of analysis of variance. Note, the Computed '**F**' is now much greater than the tabulated. It indicates that the three clusters are more distinctly different from each other than was the case with first two clusters.

Four Clusters

The procedure divides the whole lot in four clusters (**Table-6**). It puts 96 cases in the **first** cluster. The mean weight of tomato in this cluster is 39 gm. Tomatoes of this size make up 27 per cent of the ungraded lot by weight. **Second** cluster contains 88 cases. The mean weight of tomato is 55 gm. Tomatoes of this size makes up 35 per cent of the ungraded lot by weight. **Third** cluster contains 47 cases. The mean weight of tomato is 69 gm. Tomatoes of this size makes up 23 per cent of the ungraded lot by weight. The remaining 21 cases are in **fourth** cluster. The mean weight of tomato is 97 gm. Tomatoes of this size makes up 15 per cent of the ungraded lot by weight. The table also shows the mean axes length of tomato in each cluster. **Table-7** shows the results of analysis of variance. Comparison of computed 'F' with the tabulated indicates that the four clusters are even more different from each other than was the case with three.

Table-1: Frequency Distribution of Variable *X*

Class Interval (gm)	Observed relative frequency (%)	Observed no. of cases	Expected relative frequency (%)	Expected no. of cases	$\frac{(o_i - e_i)^2}{e_i}$	χ^2	Table $\chi^2_{0.05,3}$
0-20	34	86	35.1	88	0.045		
21-40	45	113	40.5	102	1.186	4.462	12.837
41-60	14	35	18.4	46	2.63		
61-80	5	13	5	13	0.6		
81-100	2	5	1	2			
Total	100	252	100	252			

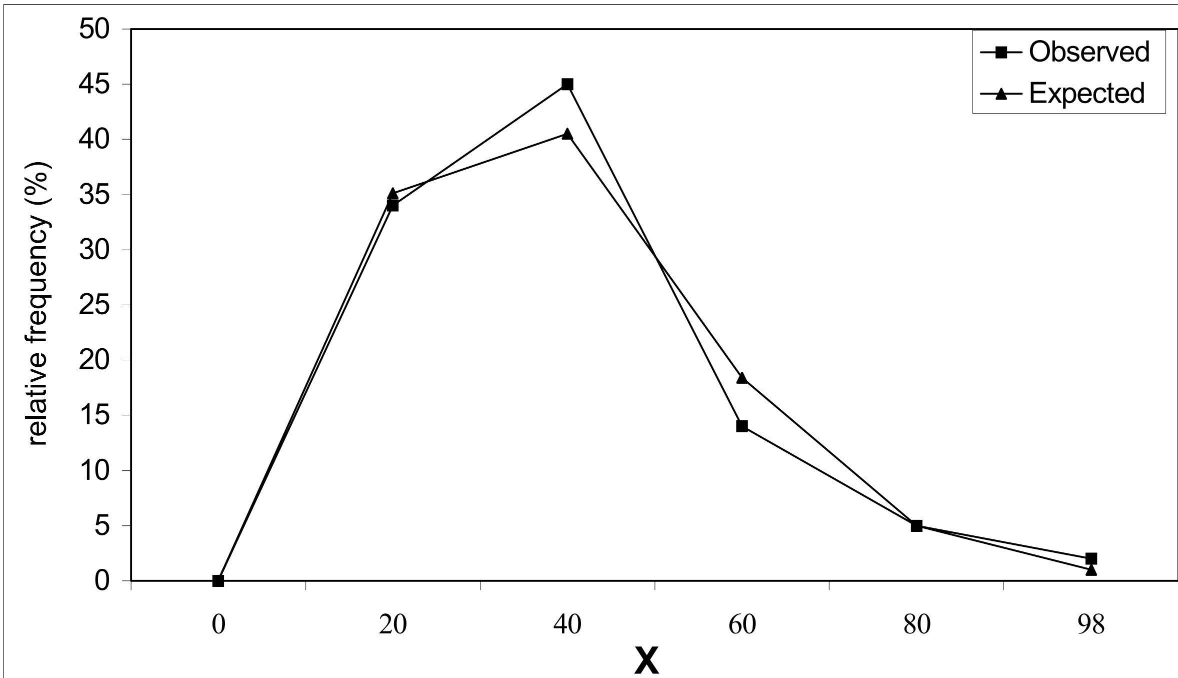

Figure-1: Frequency Distribution of Himachal Tomato

Table-2: Weight and Physical Dimensions of Tomato in Two Clusters						
Cluster	Number of pieces	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal – maximum (mm)	Horizontal - minimum (mm)
1	184	47	62	47	44	42
2	68	78	38	55	52	50
Total	252		100			

Table-3: Analysis of Variance of Two Clusters						
Variable	Sum of square between clusters	df	Sum of square within clusters	df	F ratio	Table $F(1,250)_{0.05}$
Weight (gm)	47173.699	1	30652.701	250	384.7	254
Total	47173.699	1	30652.701	250		

Table-4: Weight and Physical Dimensions of Tomato in Three Clusters						
Cluster	Number of pieces	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal -max (mm)	Horizontal -min (mm)
1	126	42	38	45	42	41
2	105	63	48	52	49	47
3	21	97	14	59	54	52
Total	252		100			

Table-5: Analysis of Variance of Three Clusters						
Variable	Sum of squares between clusters	df	Sum of squares within cluster	df	F-ratio	Table $F(2,249)_{0.05}$
Weight	65098.414	2	12727.988	249	636.766	19.5
Total	65098.414	2	12727.988	249	636.766	19.5

Table-6: Physical Dimensions and Weight of Tomato in Four Clusters						
Cluster	Number of pieces	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal –max (mm)	Horizontal -min (mm)
1	96	39	27	44	42	40
2	88	55	35	50	47	45
3	47	69	23	50	49	47
4	21	97	15	59	56	54
Total	252		100			

Table-7: Analysis of Variance of Four Clusters						
Variable	Sum of squares between clusters	df	Sum of squares within cluster	df	F-ratio	Table $F(3,248)_{0.05}$
Weight	70094.331	3	7732.072	248	749.406	8.54
Total	70094.331	3	7732.072	248		

Distribution underlying Longitudinal Axis

Longitudinal axis varied from 37 to 68mm. Data was transformed by subtracting the minimum (37 mm) from actual for all 252 observations.

L actual longitudinal axis

$Y = L - 37.$

Estimates of mean and variance of Y, from sample data are

$\bar{y} = 12$

$S_y^{\,2} = 36$

Using the method of moments α, β were estimated

$\alpha = 2.099$

$\beta = 13.55$

Using these parameters the particular distribution thus is

$$f(y) = 0.00883\ y^{1.099}\ e - (y/13.55)^{2.099} \qquad (4)$$

Table-8 and **Figure-2** show the expected frequency using (4). Goodness of fit was tested by Chi square test. The computed χ^2 (1.6) is less the tabulated $\chi^2_{0.05,3}$ (5.992), which indicates that the Weibull provides a good description of the data.

Table-8: Frequency Distribution of Variable y

Transformed longitudinal axis class (mm)	Observed relative frequency (%)	Observed number of cases	Expected cumulative relative frequency (%)	Expected number of cases	$\frac{(o_i - e_i)^2}{e_i}$	χ^2	Table $\chi^2_{0.05,2}$
0-10	43	108	41.1	104	0.189	1.60	5.992
11-20	49	124	48.5	122	0.026		
21-30	0.796	20	9.9	26	1.385		
31-40	0.004		0.5				

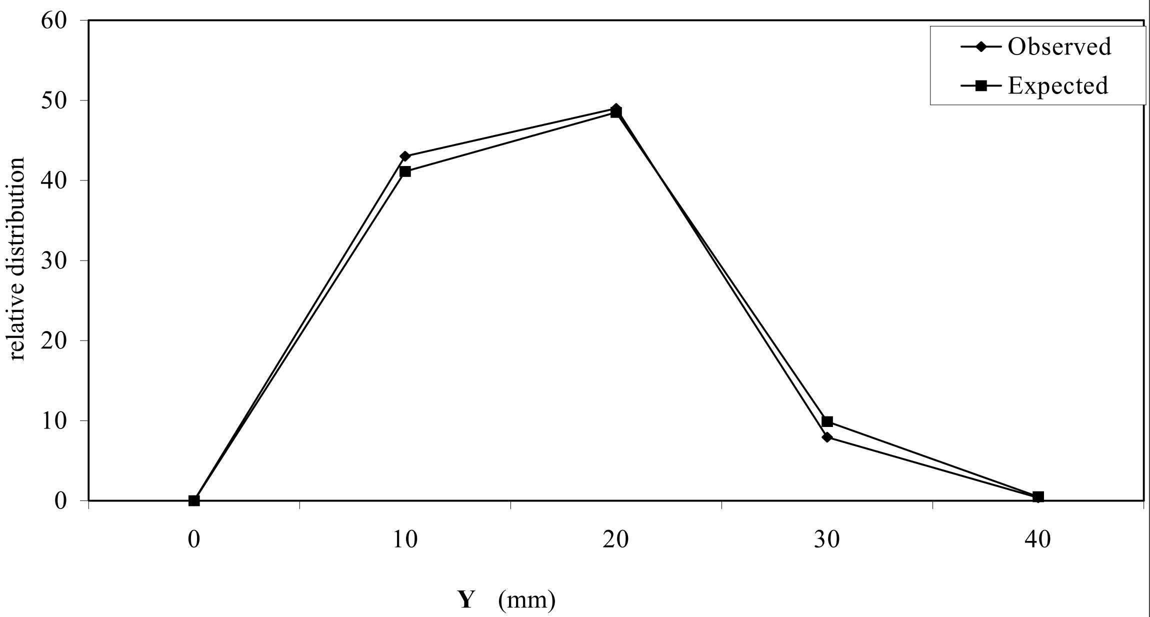

Figure-2: Observed and Expected Frequency Distribution of Longitudunial Axis of Himachal Toamto

Cluster Analysis using Axial Dimension as a Basis

Two Clusters

The procedure divides the whole lot in two clusters (**Table-9**). It puts 137 cases in the **first cluster**. The tomatoes contained in this cluster are smaller, with mean longitudinal axis of 45 mm. The mean of horizontal-max and horizontal-min axis are 43mm and 41 mm respectively. **Second cluster** contains 115 cases with mean longitudinal axis of 54mm. The mean of horizontal-max and horizontal-min axis are 50mm and 48 mm respectively. Table also shows the mean weight and mass proportion of tomato in each cluster.

Table-10 shows the results of analysis of variance. Comparison of computed 'F' with the tabulated indicates that the two clusters are different from each other only in horizontal axes. The longitudinal axes is not really different- computed F value is equal to that of tabulated. The analysis indicates that division of tomato in just two clusters is likely to be unsatisfactory. There will be considerable overlap in sizes of tomato in the two clusters. It would be desirable to increase the cluster number, which we will now try.

Three Clusters

The results of three clusters are given in **tables-11 and 12** and of four clusters in **tables- 13 and 14**.

Table-9: Physical Dimensions and Weight of Tomato in Two Clusters

Cluster	Number of pieces (% of total)	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal -max (mm)	Horizontal -min (mm)
1	137 (54)	43	42.5	45	43	41
2	115 (46)	70	57.5	54	50	48
Total	252		100			

Table-10: Analysis of Variance of Two Clusters

Axes	Sum of squares between clusters (mm^2)	df	Sum of squares within cluster (mm^2)	df	F-ratio	Table $F(1,250)_{0.05}$
Longitudinal	4561.786	1	4488.526	250	254.080	254
Horizontal -max	3739.666	1	2782.052	250	336.053	
Horizontal-min	3567.284	1	2586.781	250	344.761	
Total	11868.736	3	9857.359	750		

Table-11: Physical Dimensions and Weight of Tomato in Three Clusters						
Cluster	Number of pieces (% of total)	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal -max (mm)	Horizontal -min (mm)
1	87 (35)	39	24	43	41	40
2	111 (44)	55	44	50	47	45
3	54 (21)	81	32	57	53	51
Total	252		100			

Table-12: Analysis of Variance of Three Clusters						
Axes	Sum of squares between clusters (mm^2)	df	Sum of squares within cluster (mm^2)	df	F-ratio	Table $F(2,249)_{0.05}$
Longitudinal	6055.290	2	2995.012	249	251.713	19.5
Horizontal -max	4597.668	2	1924.039	249	297.504	19.5
Horizontal -min	4385.247	2	1768.844	249	308.655	19.5
Total	15038.204	6	6687.895	747		

Table-13: Physical Dimensions and Weight of Tomato in Four Clusters						
Cluster	Number of pieces (% of total)	Mean weight (gm)	Mass proportion (%)	Mean axes length		
				Longitudinal (mm)	Horizontal -max (mm)	Horizontal -min (mm)
1	66 (26)	38	18	43	41	39
2	74 (29)	49	26	48	45	43
3	88 (35)	64	40	53	49	47
4	24 (10)	94	16	59	56	54
Total	252		100			

Table-14: Analysis of Variance for all Four Clusters						
Axes	Sum of squares between clusters (mm^2)	df	Sum of squares within cluster (mm^2)	df	F-ratio	Table $F(3,248)_{0.05}$
Longitudinal	6363.421	3	2686.900	248	195.8	8.54
Horizontal- max	5182.647	3	1339.068	248	319.9	
Horizontal -min	4829.624	3	1324.463	248	301.4	
Total	16375.693	9	5350.430	744		

Table-15 shows the summary of analysis based on weight and **table-16** that based on axial dimension.

Final decision on the number of clusters to be made will depend on consumer preference in a particular market. This can be achieved by test-marketing. Once the number of clusters to be made is finalised, mass-proportion data can help determine the price schedule.

Conclusions

(1) Weight and longitudinal axis length of Himachal tomato are both described satisfactorily by Weibull distribution.

(2) Cluster Analysis divides the ungraded produce into any number of clusters desired, yielding useful information on physical sizes of tomato in each cluster and mass proportion.

Results can be used for determination of dimension of size grading mechanism such as slots in the endless bolt mechanism. Results can also be used for pricing and test marketing.

Table-15 : Clusters based on Weight				
If Produce is divided in	**Mean Weight of tomato (gm) in Cluster Number**			
	One	**Two**	**Three**	**Four**
Two Clusters	47 (Small 62%)	78 (Large 38%)	**-**	**-**
Three Clusters	42 (Small 38%)	63 (Medium 48%)	97 (Large 14%)	-
Four Clusters	39 (Small 27%)	55 (Medium 35%)	69 (Large 23%)	97 (Extra Large 15%)
Note : Number in parenthesis is the mass proportion.				

Table-16 : Clusters based on Axial Dimensions				
If produce is divided in	**Mean Length of Longitudinal Axis (mm) in Cluster Number**			
	One	**Two**	**Three**	**Four**
Two Clusters	45 (Small 42.5%)	54 (Large 57.5%)	**-**	**-**
Three Clusters	43 (Small 24%)	50 (Medium 44%)	57 (Large 32%)	-
Four Clusters	43 (Small 18%)	48 (Medium 26%)	53 (Large 40%)	59 (Extra Large 16%)
Note : Number in parenthesis is the mass proportion.				

References

1. Sharan G. and Rawale K. (2001). "New Packaging Options for Transporting tomatoes in India." Food Chain: *International Journal of Small Scale Food Processing*, No.29 (November), Intermediate Technology Development Group Limited, UK.

2. Dillion W.R. and Mathew Goldstein (1984). Multivariate Analysis: Methods and Applications. John Wiley and Sons.

3. Devore Jay. L. (2000). Probability and Statistics for Engineering and Sciences. Duxbury Thomson Learning, USA.