Winter school on "Current trends in packaging of fish and fishery products" organised by Central Institute of Fisheries Technology, Cochin, 17th november - 7th december, 2005

# Statistical Applications and Data Considerations in Fish Processing Studies

V. Geethalakshmi

By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other (definition by Prof. Horace Secrist).

In essence, Statistics is the science which deals with the collection, analysis and interpretation of numerical data. The main problems in statistical inference can be broadly classified into two areas namely estimation of population parameters along with confidence intervals and hypothesis testing. The article highlights the statistical applications in studies related to fish processing.

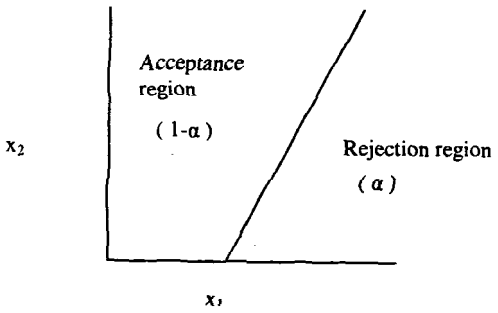
## 1. Hypothesis Testing

A statistical hypothesis is some statement or assertion about a population or equivalently about the probability distribution characterizing the population which we want to verify on the basis of information available from a sample. In industry, statistics is very widely used in 'Quality Control'. Krishna Rao(1971) has described the use of Control charts in studies related to shrimp canning. He has observed that application of control charts to filling performance has led to good deal of savings of material in many industries.

In production engineering, statistical tools are of extreme importance to find whether the product is conforming to specifications or not.

In hypothesis testing the null hypothesis should be framed in such a way that there is no bias from the statistician's side. If the testing of effectiveness of four types of antibiotics is to be carried out, the null hypothesis should be that 'there is no difference between the effects of four drugs'.

The sample observations together say 'n' in number can be taken as a point in the n-dimensional space and we specify some region of the space and see whether this point lies within this region or outside this region. The null hypothesis gets rejected if the sample point falls in this region. This region is called the rejection or critical region. The level of significance is called the size of the critical region. Depending on the objectives of the experiment and nature of controls applied the size can vary. Usually for experiments we set the level of significance at 1% or 5%. In quality control terminology, $\alpha$ is termed as producer's risk and $1-\alpha$ is termed as consumer's risk.




Two ways of hypothesis testing are parametric approach and non parametric approach. In both the approaches standard statistical tests are available for various

situations. The popular parametric tests are t-test, F-test, z-test and the ANOVA technique. The parametric tests are based on population assumptions and the results are valid only in the case when these assumptions are satisfied. The non-parametric tests are not based on any population assumptions and are easy to apply for qualitative type of data.

For studies on transportation of fresh fish from one centre to another the quality changes in the sample can be tested using the t-test by taking microbial counts from different samples before and after transportation.

To test if sample mean $\bar{x}$ differs significantly from the hypothetical value $\mu$, the population mean the t-test is used. The *Student's t is defined by the statistic*

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}} \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The test can also be applied to test the significance of an observed sample correlation coefficient The assumptions of the t-test are

(i) The parent population from which the sample is drawn is Normal

(ii) The sample observations are independent i.e. the sample is random

(iii) The population standard deviation s is unknown.

NOTE : The significant values of t at level of significance for a single tailed test can be obtained from those of two-tailed test by looking the values at level of significance '$2\alpha$'.

**Analysis of variance** is a statistical tool to express the total variation in the data into different components such as due to factors or causes (independent variables) and due to chance factor. The latter component is called the experimental error.

The ANOVA technique is used to test the equality of means of several populations. When the efficiency of more than two feeds have to be compared in a feeding experiment, ANOVA can be used. In some situations when treatments have to be compared there may be 2 levels of control. In such cases we use 2-way ANOVA. For example if the growth of phytoplankton has to be studied in different seasons and stations, one may use the 2-way ANOVA technique.

For the validity of the F-test in ANOVA, the following assumptions are made:

(i) the observations are independent.

(ii) Parent population from which observations are taken is normal

(iii) Various treatments and environmental effects are additive in nature

In an in vitro growth study of a particular species 4 different feeds are tested in different ponds. If the experiment is replicated in 3 seasons to test the efficacy of feeds, using the ANOVA technique one can find the difference between the feeds and variation in growth during various seasons. In testing the difference between the quality of fishery products packed in different packaging material such as cans and pouches two-way ANOVA can be applied. While using ANOVA technique it should be noted that there should be sufficient number of replications to allow necessary error d.f. for valid results.

*2. Non-parametric tests*

A non-parametric test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. However the sample observations should be independent and the variable under study should be continuous and the lower order moments should exist. They are useful to deal with data which are given in ranks or have

seemingly numerical scores based on the strength of ranks. In sensory evaluation of fishery products the data is mainly ranks given by the taste panel. These tests can be applied in such analysis. If two judges rank the acceptability of 10 products on a 10 point scale we may wish to know whether there is any degree of agreement among the rankings. The rank correlation coefficient due to Spearman usually denoted by $r_s$ is the ordinary correlation coefficient r between the ranked values. If d is difference in ranks, the correlation coefficient is given by

$$r_S = 1 - 6\frac{\sum d^2}{n(n^2-1)}$$

Another measure of degree of concordance closely related to $r_s$ Kendall's $\tau$. To compute this rearrange the two rankings so that one of them is in the order 1,2,3,...n. Taking each rank given by judge no.2 in turn count how many of the ranks to the right of it are smaller than it and add these counts. Find the total Q. The Kendall's is given by

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

It should be noted that Kendall's $\tau$ test is the non-parametric equivalent of one-way ANOVA. In almost all situations the values of Spearman's rank correlation and Kendall's $\tau$ are very close and would invariably lead to the same conclusions.

Another important statistical test frequently used is the Chi-square test. It is also a non-parametric test. Suppose products are locale based and the difference/impact of locality on product has to be ascertained. In such situations we can use the testing of attribute using $\chi^2$. Let us consider two attributes say A and B divided into r & s classes each say $A_i$ & $B_j$. Let $(A_i)$ i=1,r be the number of persons having attribute $A_i$ and $(B_j)$ j=1,s be the number of persons having attribute $B_j$. $(A_i B_j)$ be the no. of persons with both the attributes $A_i$ and $B_j$. Also $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j$ is the total frequency.

Then $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{(A_iB_j)-(A_iB_j)_0}{(A_iB_j)_0}\right]^2$ is distributed as a variate with (r-1)(s-1) d.f.

The problem is to test whether the two attributes are independent or not.

Here

$$\chi^2 = \frac{N(ad-bc)2}{(a+c)(b+d)(a+b)(c+d)}, N = a+b+c+d$$

with 1 d.f. where a,b,c,d are the cell frequencies

Another application of Chi-square test is the testing the goodness of fit between theory and experiment. If $O_i$ and $E_i$ are a set of observed and expected frequencies, then

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}, (\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i)$$ follows $c^2$ with (k-1) d.f.

For goodness of fit $\chi^2$-test to be valid the following conditions must be satisfied

(i) The sample observations must be independent

(ii) Constraints on cell frequencies, if any, should be linear i.e. $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$

(iii) No theoretical cell frequency should be less than 5. This is essential for maintaining the character of continuity. If any theoretical cell frequency isles than 5 then it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

### 3. Multivariate analysis of statistical data

Consumer acceptability depends on parameters like colour, odour, taste, flavour, firmness, fibrousness and succulence.

Sensory evaluation is one of the most reliable method for evaluation of freshness of fishery products. Quite a good number of tests are used in sensory evaluation of qualities of foods and beverages. They can be broadly classified under four heads as difference tests, preference tests, descriptive tests and sensitivity tests.

The data obtained from the sensory evaluation are multivariate data. For consumer analytics and market preference research multivariate statistical methods can be successfully applied. Suppose 3 new products are developed and one wishes to find out the discriminating causes which distinguishes these products from the standard one. Clustering can be done to the data. Consumer acceptance of a product is always due to a combination of causes. Say for example a product may be more preferred to another due to taste and nutrition though the other appears more acceptable. The unobservable latent causes lying behind the acceptability of a product can be found by performing a factor analysis and obtaining the factor loading for each combination of factors.

**Multi-dimensional Scaling**

When the data is rank based the best approach to use is the multi-dimensional scaling. Multidimensional scaling (MDS) can be considered to be an alternative to factor analysis. In general, the goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects. In factor analysis, the similarities between objects (e.g., variables) are expressed in the correlation matrix. With MDS one may analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices. MDS attempts to arrange "objects" (preference pattern in case of the consumer acceptance example) in a space with a particular number of dimensions so as to reproduce the observed distances. As a result, we can "explain" the distances in terms of underlying dimensions; in our example, we could explain the distances in terms of the preferences for a particular fishery product with respect to a set of parameters.

The "beauty" of MDS is that we can analyze any kind of distance or similarity matrix. These similarities can represent people's ratings of similarities between objects, the percent agreement between judges, the number of times a subjects fails to discriminate between stimuli, etc. For example, MDS methods used to be very popular in psychological research on person perception where similarities between trait descriptors were analyzed to uncover the underlying dimensionality of people's perceptions of traits. They are also very popular in marketing research, in order to detect the number and nature of dimensions underlying the perceptions of different brands or products.

In general, MDS methods allow the researcher to ask relatively unobtrusive questions ("how similar is brand A to brand B") and to derive from those questions underlying dimensions without the respondents ever knowing what is the researcher's real interest.

**4. Regression analysis**

When the cause-effect relationship between the responses and the variables or factors are to be worked out we go for regression analysis. The outcome of this analysis is a set of coefficient of causes. When the factors affecting the productivity of a certain species is to be determined multiple regression analysis can be done using different models. Basically models are of two forms namely Linear and the Non-linear models. A simple linear model is of the form

$$Y = a + bX$$

A statistical model assigns an element of error into the model and the form of the model becomes

$$Y\ a + bX + e$$

where e is the vector of errors and it follows a Normal distributions. The parameter b is called the regression coefficient and a is called the constant. The parameter b is tested using the t-test. When the responses are qualitative in nature we use the logistic regression. Non-linear models are employed while analyzing data from growth related studies.

**Logistic regression**

There are many important research topics for which the dependent variable is "limited" (discrete not continuous). Researchers often want to analyze whether some event occurred or not, such as voting, participation in a public program, business success or failure, morbidity, mortality, a hurricane and etc. Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1).

"Why shouldn't I just use ordinary least squares?" Good question.

Consider the linear probability (LP) model:

$$Y = \mathbf{a} + \mathbf{B}X + e$$

where

- Y is a dummy dependent variable, =1 if event happens, =0 if event doesn't happen,
- a is the coefficient on the constant term,
- B is the coefficient(s) on the independent variable(s),
- X is the independent variable(s), and
- e is the error term.

Use of the LP model generally gives you the correct answers in terms of the sign and significance level of the coefficients. The predicted probabilities from the model are usually where we run into trouble. There are 3 problems with using the LP model:

1. The error terms are heteroskedastic (heteroskedasticity occurs when the variance of the dependent variable is different with different values of the independent variables):

   var(e)= p(1-p), where p is the probability that EVENT=1. Since P depends on X the "classical regression assumption" that the error term does not depend on the Xs is violated.

2. e is not normally distributed because Y takes on only two values, violating another "classical regression assumption"

3. The predicted probabilities can be greater than 1 or less than 0 which can be a problem if the predicted values are used in a subsequent analysis. Some people try to solve this problem by setting probabilities that are greater than (less than) 1 (0) to be equal to 1 (0). This amounts to an interpretation that a high probability of the Event (Nonevent) occuring is considered a sure thing.

The "logit" model solves these problems:

$$\ln[p/(1-p)] = a + bX + e$$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"

ln[p/(1-p)] is the log odds ratio, or "logit".

The logistic distribution constrains the estimated probabilities to lie between 0 and 1. The estimated probability is:

$$p = 1/[1 + \exp(-a - b X)]$$

- if you let a + b X =0, then p = .50 as a + b X gets really big, p approaches 1 as a

+ b X gets really small, p approaches 0

Since:

ln[p/(1-p)] = a + bX + e

The slope coefficient (b) is interpreted as the rate of change in the "log odds" as X changes ... not very useful.

Since:

p = 1/[1 + exp(-a - b X)]

The marginal effect of a change in X on the probability is: Mp/MX = f(b X) b