Copy 11

UCRL-2190

UNCLASSIFIED

UNIVERSITY OF CALIFORNIA

Radiation Laboratory

Contract No. W-7405-eng-48

UCRL LECTURES ON NUMERICAL ANALYSIS AND APPLIED MATHEMATICS

Lecture XII

Edward S. Robbins

December 9, 1952

Berkeley, California

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

University of California
Radiation Laboratory
Berkeley, California

UCRL LECTURES ON NUMERICAL ANALYSIS AND APPLIED MATHEMATICS

Lecture XII

December 9, 1952

Edward S. Robbins

THE CHI-SQUARE DISTRIBUTION

1. Introduction

The chi-square distribution was encountered originally as a test to determine the goodness of fit of experimental data to frequency curves. It also occurs in consideration of the distribution of sample variances with respect to the true variance of the population being sampled. Using this function we are able to find a range of values within which we may reasonably expect the true value of the variance to lie.

First, we recall certain definitions from statistics. Statistics essentially considers data which are obtained by repeating certain operation, or which can be considered as obtained repetitively. Certain mathematical models are used in describing such data, usually one which can be used in predicting the frequency with which certain results can be expected to occur in the operation.

In the discussion here, all data will be treated as though it has been classified, that is, we arrange the data in a certain number of classes. Such an arrangement of data forms a frequency distribution. The theoretical

frequency distribution is the mathematical model or population of possible osbervations. The data are considered as a sample extracted from this population.

2. Definitions

When data are classified all members of a class are considered as having their measurements at the midpoint of the class interval. This value is called the class mark and is denoted by x_i. The number of measurements in the i'th class interval is f_i, and n denotes the total number of measurements taken. The arithmetic mean of the sample is denoted by $\bar{x}$.

The variance of the sample is

$$s^2 = \frac{1}{n} \sum_i (x_i - x)^2 f_i .$$

The chi-square function with v degrees of freedom is

$$f(\chi^2) = \frac{1}{2^{v/2} \Gamma(v/2)} (\chi^2)^{\frac{v-2}{2}} e^{-\frac{\chi^2}{2}} .$$

Unclassified data may be considered as the above, x_i being the observed value, $f_i = 1$.

A confidence limit for a parameter k is the range of values in which we may reasonably expect the true value of k to lie. That is, if 95% confidence limits for s^2 are $0.05 < s^2 < .65$, it is to be expected that, were the experiment repeated many times, s^2 should fall in the range indicated 95% of the time.

3. Theorems

We give without proof several theorems concerning the properties of the chi-square distribution. Consider x, normally distributed with $\bar{x} = 0$, and a random sample of n cases $x_1 \ldots x_n$. It is desired to consider the distribution of

$$\mu = \sum_1^n x_i^2 ,$$

where each of the x_i are selected independently and so have the same distribution function.

Theorem I: If x is distributed normally with zero mean and unit variance, μ has a chi-square distribution with n degrees of freedom.

Theorem II: If x is distributed normally with variance σ^2, and s^2 is a sample variance from a random sample of size n; then $\frac{n s^2}{\sigma^2}$ has a chi-square distribution with n-1 degrees of freedom.

Theorem III: If χ_1^2 and χ_2^2 possess independent χ^2 distributions with v_1 and v_2 degrees of freedom respectively, then $x_1^2 + x_2^2$ possesses a χ^2 distribution with $v_1 + v_2$ degrees of freedom.

4. Confidence Limits for σ^2.

Let x have a normal distribution with variance σ^2, and s^2 be a sample variance based on a random sample of size n. Determination of 95% confidence limits for σ^2 is made as follows. From a table of the chi-square function find a value of χ_1^2 such that the probability that $\chi^2 > \chi_1^2$ is 0.975 and a value χ_2^2 such that the probability is 0.025 that $\chi^2 > \chi_2^2$. Then the probability is 0.95 that

$$\chi_1^2 < \frac{n s^2}{\sigma^2} < \chi_2^2 \quad \text{or that} \quad \chi_2^2 < \frac{\sigma^2}{n s^2} < \chi_1^2 .$$

5. Uses of the Chi-Square Function.

As an introduction to the use of the chi-square function in frequency curve fitting consider the following example. The problem is to determine whether a single die is honest or not. We assume that if in 96 throws we get a distribution which could be expected less than 5% of the time, then the die should be rejected. The observed frequency is O_i, the expected or theoretical frequency is e_i. The results are:

Face	1	2	3	4	5	6
O_i	10	22	9	19	19	17
e_i	16	16	16	16	16	16

As a measure of the compatibility of the observed and theoretical data compute

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - e_i)^2}{e_i} \quad :$$

We get

$$\chi^2 = 8.75 \quad .$$

From a table of the chi-square function for 5 degrees of freedom we find that the probability that $\chi^2 > 11.1$ is .05 . Thus, if the experiment were repeated, more than 5% of the repeats would give a value greater than 8.75; in fact, more than 10% will do so. The value of the degrees of freedom in a case such as this is the number of independent cell frequencies.

In cases of this type, the data is discrete and cannot be approximated by a distribution curve. However, if the number of possible values of χ^2 is large, the histogram is quite regular and approximates the χ^2 case quite well. As a general rule, if $e_i \geq 5$ and $k \geq 5$, the approximation is good; if $k < 5$, then e_i must be larger than 5. Reclassification of the data may enable the application of the test to a case that does not at first appear to fulfill the requirements.

In general, when considering the fitting of a theoretical frequency distribution by an observed frequency distribution, the chi-square test is used as a measurement of the goodness of fit. The number of degrees of freedom is the number of classifications minus the number of parameters of the distribution determined by sample estimates. One must be careful not to count those arbitrarily determined from other considerations (i.e., as restrictions on the experiment). If in an experiment, the binomial distribution $N(q + p)^n$ is fitted to the data where N is the sample size, p the probability of success, $q = 1 - p$, and n the number of trials, then in general $v = k - 3$, with k the number of classifications. If one assigns a value to n before the experiment, as in considering some characteristic of hog litters of 9 piglets each, then n is not determined from the experiment and v is $k - 2$.

There are other applications to two-way contingency tables and further refinements of this test which are described in the articles contained in the bibliography.

6. Bibliography.

6.1 Hoel, P., Introduction to Mathematical Statistics, Wiley and Sons, New York.

6.2 Kenney, J. F., Mathematics of Statistics, D. Van Nostrand Co.

6.3 Kendall, M. G., The Advanced Theory of Statistics, Vol. 1, Griffin and Company.

6.4 Rider, P., An Introduction to Modern Statistical Methods, John Wiley and Sons.