Notes on the chi-squared distribution

Ernie Croot

October 7, 2008

1 Introduction

Know the material in your book about chi-squared random variables, in addition to the material presented below.

1.1 Basic properties of chi-squared random variable

A chi-squared random variable χ^2_n with n degrees of freedom is a continuous random variable taking on values in $[0, \infty)$. It has the probability density function

$$f(x) \ = \ \begin{cases} \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$

Here, $\Gamma(x)$ is the function

$$\Gamma(x) \ = \ \int_0^\infty e^{-t}t^{x-1}dt.$$

Note that this integral converges for all $x > 0$, because the e^{-t} decays so quickly to 0 as $t \to \infty$ that it more than compensates for the facts that (a) t^{x-1} tends to infinity with t when $x > 1$; and (b), for $0 < x < 1$, t^{x-1} is near infinity when t is near 0. Furthermore, this "gamma function" enjoys a number of useful properties, among them:

- For $x > 0$, $x\Gamma(x) = \Gamma(x+1)$. This is easily proved upon using some integration by parts.
- $\Gamma(1) = 1$.
- $\Gamma(1/2) = \sqrt{\pi}$. It turns out that this is a consequence of the fact that $\int_0^\infty e^{-t^2/2}dt = \sqrt{2\pi}$, together with a change of variable (i.e. "u substitution").

• And, combining together the first two above, one sees that for an integer $x \geq 1$, $\Gamma(x) = (x-1)!$, where $0! = 1$ by convention.

The mean and variance of χ^2_n are given by

$$\mathrm{E}(\chi^2_n) \;=\; n, \text{ and } \mathrm{V}(\chi^2_n) \;=\; 2n.$$

1.2 Additional properties

It turns out that χ^2_n has the same distribution as

$$X_1^2 + X_2^2 + \cdots + X_n^2,$$

where each of the X_i are independent, standard normal random varibles. In other words,

$$\mathrm{P}(\chi^2_n \leq a) \;=\; \mathrm{P}(X_1^2 + \cdots + X_n^2 \leq a).$$

This will be useful in developing the "chi-squared test statistic": One can think of these $X_1^2, ..., X_n^2$ as the "squares of errors in n independent measurements". So, if one develops a model to describe some data, and one decides that the "errors" in each measurement are approximately $N(0,1)$ in distribution, then one can get some idea of how well ones estimate for certain model parameters match the measured (noisy) parameters. And, one would like to know when the discrepancy between these values (model and measured) makes the model appear to be improbable, so that one can reject it.

1.2.1 An important formula

Using a tedious integration-by-parts computation, one can further prove the following beautiful formula, which holds in the case for n even: Let Y be a Poisson random variable with parameter $\lambda = a/2$. Then,

$$\mathrm{P}(\chi^2_n \;\geq\; a) \;=\; \mathrm{P}(Y \leq n/2 - 1).$$

So, in the case $n = 2$ we have

$$\mathrm{P}(\chi^2_n \;\geq\; a) \;=\; \mathrm{P}(Y = 0) \;=\; e^{-a/2}.$$

There is similarly a formula for the case n odd, but it is more complicated.

2 Two applications

2.1 Goodness of fit of a model

The following application is not the usual "chi-squared test statistic", but gives some indication of how chi-squared random variables may be used. Later in the course we will discuss the chi-squared test statistic in depth, so be patient.

Suppose that $p(t)$ is the measured position of an object, where we will only work with the times $t = 1, 2, ..., 6$. We wish to test the hypothesis that "the actual position of the object at time t is t." Assuming that this is the case, we further make the assumption that the discrepancy between observed and actual position is

$$p(t) - t \;=\; N(0, 1),$$

for each of the times $t = 1, 2, ..., 6$. Furthermore, we assume that all the $p(i) - i$ are independent of each other (i.e. the errors between actual and observed positions at different times are independent of each other).

Before we delve deeper into this application, we note that it is not a bad assumption that this discrepancy has a normal distribution, since often errors in measurement are the result of many little errors acting cumulatively on the observed value (such as millions of air particles deflecting slightly a laser beam used to find the position of an object).

We now want to test the hypothesis that "the actual position at time t is t", so we perform an experiment and obtain the following observed position values:

$$p^*(1) = 0.5,\ p^*(2) = 1.5,\ p^*(3) = 3.0,\ p^*(4) = 3.0,\ p^*(5) = 5.5,\ p^*(6) = 6.0.$$

The reason for using $p^*(t)$ here in place of $p(t)$ is that we think of $p(t)$ as a random variable, while $p^*(t)$ are observed instances of the random variable (alsed called "exposures of $p(t)$"). This notation is fairly common.

Now we compute

$$E \;=\; (p^*(1) - 1)^2 + (p^*(2) - 2)^2 + \cdots + (p^*(5) - 5)^2 \;=\; 1.75;$$

that is, E is the sum of the square errors between the predicted location of the object and the observed location of the object for times $t = 1, 2, ..., 6$. Since we have assumed that $p(t) - t$ is $N(0, 1)$ it follows that

$$(p(1) - 1)^2 + \cdots + (p(6) - 6)^2 \;\text{ is chi} - \text{squared with 6 degrees of freedom.}$$

So, to see whether the hypothesis that "the actual position at time t is t" is a good one, we compute, with the aid of the formula in section 1.2.1,

$$\begin{aligned} P(\chi^2 \geq 1.75) &= e^{-1.75/2} \sum_{j=0}^{2} \frac{(1.75/2)^j}{j!} \\ &= 0.41686202(1 + 0.875 + 0.3828125) \\ &\approx 0.94. \end{aligned}$$

Thus, there is about a 94% chance that one will get a sum-of-squares error that is at least as big as our observed error $E = 1.75$. In other words, the error we observed was actually quite small, so we don't reject the hypothesis "the actual position of the object at time t is t".

2.2 Population sampling: the chi-squared test statistic

Another type of problem where a chi-squared distribution enters into hypothesis testing is population sampling; indeed, this problem is one where the chi-squared test statistic is absolutely critical in checking claims about a population makeup. Here is the setup: Suppose you have a population that is divided into k different categories. Further, you hypothesize that the fraction of individuals in the jth category is p_j. Note that $p_1 + \cdots + p_k = 1$. You now wish to test this hypothesis by picking a large number of individuals, and checking to see which category they fall into. If you let X_j denote the number of people in category j, then note that $X_1 + \cdots + X_k$ is the total number of people in your sample. It is clear then that the X_i's are not independent, because if one knows any $k-1$ of the X_i's, then the remaining X_i is determined.

Now fix an integer $j = 1, 2, ..., k$. We would like to know the distribution of X_j under the hypothesis that p_j fraction of the population is in the jth class. This distribution turns out the be binomial, as can be seen as follows: Suppose that the sample size is N. Then, we define the Bernoulli random variables $B_1, ..., B_N$ such that

$$B_i = \begin{cases} 1, & \text{if person } i \text{ is in } j\text{th category} \\ 0, & \text{if person } i \text{ is not in } j\text{th category.} \end{cases}$$

Furthermore, $P(B_i = 1) = p_j$. Clearly,

$$X_j = B_1 + \cdots + B_N.$$

So, the mean of X_j is p_jN and its variance is

$$V(X_j) \;=\; V\left(\sum_{i=1}^{N} B_i\right) \;=\; \sum_{i=1}^{N} V(B_i) \;=\; NV(B_1) \;=\; Np_j(1-p_j).$$

We have used here the fact that the B_i's are independent to write this variance of a sum of B_i's as a sum of variances of the B_i's.

Next, we normalize X_j by letting

$$Y_j \;=\; \frac{X_j \;-\; p_jN}{\sqrt{Np_j(1-p_j)}}.$$

Note that Y_j has mean 0 and variance 1. Furthermore, from the Central Limit Theorem, as N tends to infinity, Y_j approaches $N(0,1)$ in the following sense: For every real number c, we have that

$$\lim_{N\to\infty} P(Y_j \le c) \;=\; P(N(0,1) \le c).$$

Now consider the following sum of square errors:

$$E \;=\; \sum_{j=1}^{k} Y_j^2. \tag{1}$$

This function will be "large" precisely when several of the Y_j's stray "too far" from 0, and this happens precisely when several of the X_j's stray "too far" from their conjectured mean p_jN. Also, if p_j is the correct percent of the population belonging to class j for all $j = 1, 2, ..., k$, then we expect that X_j should be "close" to p_jN for all $j = 1, 2, ..., k$, and thus Y_j should be "close" to 0 for all $j = 1, 2, ..., k$; and so, we should have that E is "small". In order to be able to check our hypothesis that "p_j fraction of the population belongs to class j, for all $j = 1, 2, ..., k$", we need to know the probability distribution of E, and the following result gives us this needed information:

Theorem. For large values of N, the random variable E given in (1) has approximately a chi-squared distribution with $k-1$ degrees of freedom.

A natural question here is: Why only $k-1$ degrees of freedom, why not k? The reason is that, as was stated earlier, the X_j's, and therefore the Y_j's, are not independent: We have that any X_j is completely determined by the other $k-1$ values X_i's since $X_1 + \cdots + X_k = N$.

We will not prove the above theorem, as its proof is long and technical, but we will apply it here to a simple example:

Example: Suppose you read in a newspaper that likely voters in Florida break down according to the following distribution: 40% will vote Republican, 40% will vote Democrat, 10% will vote Libertarian, 5% will vote Green, and 5% will vote "other".

You decide to test this by doing a poll of your own. Suppose that you ask 10,000 likely Florida voters which group they will vote for, and suppose you receive the following data:

4,200 will vote Republican;
3,900 will vote Democrat;
1,000 will vote Libertarian;
700 will vote Green; and,
200 will vote "other".

So, we let

$$X_1^* = 4200,\ X_2^* = 3900,\ X_3^* = 1000,\ \text{and } X_4^* = 700,\ \text{and } X_5^* = 200.$$

The $*$ indicates that the X_j^* is an observed value, rather than a random variable.

Then, we compute

$$\begin{aligned}
Y_1^* &= \frac{4200 - 4000}{100\sqrt{0.24}} = \frac{2}{\sqrt{0.24}} \approx 4.0825 \\
Y_2^* &= \frac{3900 - 4000}{100\sqrt{0.24}} = -\frac{1}{\sqrt{0.24}} \approx -2.0412 \\
Y_3^* &= 0 \\
Y_4^* &= \frac{700 - 500}{100(0.2179)} = \frac{2}{0.2179} \approx 9.177 \\
Y_5^* &= \frac{200 - 500}{100(0.2179)} = -\frac{3}{0.2179} \approx -13.768.
\end{aligned}$$

So, the sum-of-square errors is

$$E^* = (Y_1^*)^2 + \cdots + (Y_5^*)^2 \approx 294.608.$$

Now, to test whether our conjectured probabilities for the types of likely voters is correct, we select a parameter α, which is often taken to be 0.05.

Then, we ask ourselves: What is the probability that a chi-squared random variable having $k - 1 = 4$ degrees of freedom has value ≥ 294.608; that is,

$$P(\chi_4^2 \geq 294.608) \ = ?.$$

If this value of less than $\alpha = 0.05$, then we reject the hypothesis on the makeup of likely Florida voters; otherwise, we do not reject the hypothesis, which is not the same as saying that we accept it.

After doing a table lookup, it turns out that

$$P(\chi_4^2 \leq 294.608) \ = \ 1.000..., \ \text{ is very close to 1;}$$

and so, the probability we seek is much smaller than α. So, we reject our hypothesis on the makeup of likely Florida voters.