Notes on the Chi-Squared Distribution

October 19, 2005

1 Introduction

Recall the definition of the chi-squared random variable with k degrees of freedom is given as

$$\chi^2 \;=\; X_1^2 + \cdots + X_k^2,$$

where the X_i's are all independent and have $N(0,1)$ distributions. Also recall that I claimed that χ^2 has a gamma distribution with parameters $r = k/2$ and $\alpha = 1/2$: Let

$$f(x) \;=\; \left(\frac{1}{2}\right)^{k/2} \frac{x^{k/2-1}e^{-\alpha x}}{\Gamma(k/2)},$$

where $\Gamma(t)$ is the gamma function, given by

$$\Gamma(t) \;=\; \int_0^\infty x^{t-1}e^{-x}dx, \quad \text{for } t > 0.$$

Then, we are saying that

$$P(\chi^2 \geq a) \;=\; \int_a^\infty \left(\frac{1}{2}\right)^{k/2} \frac{x^{k/2-1}e^{-\alpha x}}{\Gamma(k/2)}dx.$$

In this set of notes we aim to do the following two things:

1) Show that the chi-squared distribution with k degrees of freedom does indeed have a gamma distribution;

2) Discuss the chi-squared test, which is similar to the one you have already seen in class.

2 Proof that chi-squared has a gamma distribution

We first recall here a standard fact about moment generating functions:

Theorem 1. Suppose that X and Y are continuous random variables having moment generating functions $M_X(t) = E(e^{tX})$ and $M_Y(t) = E(e^{tY})$, respectively. Further, suppose that these functions exist for all t in a neighborhood of 0, and that they are continuous at $t = 0$. Then,

$$P(X \leq a) \ = \ P(Y \leq a) \text{ for all } a \quad \Longleftrightarrow \quad M_X(t) = M_Y(t).$$

Note: Stronger version of this theorem are possible, but this is good enough for our purposes.

We will not prove this theorem here, as it is long and technical. However, let us now use it to determine the moment generating function for the chi-square distribution; first, we determine the m.g.f. for a gamma distribution: Suppose Z is a random variable having a gamma distribution with parameters $r > 0$ and $\alpha > 0$. Then, it has pdf given by

$$g(x) \ = \ \frac{\alpha(\alpha x)^{r-1}e^{-\alpha x}}{\Gamma(r)}, \quad \text{where } x \geq 0.$$

Since it is a pdf, we know that

$$\int_0^\infty g(x)dx \ = \ 1,$$

regardless of the parameters α and r. Now, then, we have that $M_Z(t)$ is given by

$$\begin{aligned} M_Z(t) \ &= \ E(e^{tZ}) \ = \ \int_0^\infty e^{tz}g(z)dz \\ &= \ \int_0^\infty \frac{\alpha^r z^{r-1}e^{-z(\alpha-t)}}{\Gamma(r)}dz \\ &= \ \frac{\alpha^r}{(\alpha-t)^r}\int_0^\infty \frac{(\alpha-t)((\alpha-t)z)^{r-1}e^{-z(\alpha-t)}}{\Gamma(r)}dz \end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha^r}{(\alpha - t)^r} 1 \\
&= \frac{1}{(1 - \alpha^{-1} t)^r}.
\end{aligned}
$$

Notice here that we used the fact that the integral

$$\int_0^\infty \frac{(\alpha - t)((\alpha - t)x)^{r-1} e^{-x(\alpha - t)}}{\Gamma(r)} dx \;=\; 1,$$

so long as $t < \alpha$.

Also, notice that $M_Z(t)$ exists in a neighborhood of $t = 0$, namely the neighborhood $|t| < \alpha$.

Next, let us compute the moment generating function for the χ^2 random variable: We have that

$$
\begin{aligned}
M_{\chi^2}(t) &= E(e^{\chi^2 t}) \\
&= E(e^{(X_1^2 + \cdots + X_k^2)t}) \\
&= E(e^{X_1^2 t} \cdots e^{X_k^2 t}) \\
&= E(e^{X_1^2 t}) \cdots E(e^{X_k^2 t}) \\
&= E(e^{X_1^2 t})^k.
\end{aligned}
$$

The facts we have used here are that 1) The X_i's are independent, which implies that the $e^{X_i^2 t}$'s are all independent, and so allows us to write $E(e^{X_1^2 t} \cdots e^{X_k^2 t})$ as $E(e^{X_1^2 t}) \cdots E(e^{X_k^2 t})$; and, 2) The X_i's all have the same same distribution, which gives $E(e^{X_i^2 t}) = E(e^{X_1^2 t})$.

We now compute $E(e^{X_1^2 t})$: We note that since X_1 is $N(0,1)$, this expectation is

$$
\begin{aligned}
E(e^{X_1^2 t}) &= \int_{-\infty}^{\infty} e^{x^2 t} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \\
&= \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{x^2}{2(1-2t)^{-1}}\right)}{\sqrt{2\pi}} dx \\
&= \sqrt{\frac{1}{1-2t}} \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{x^2}{2(1-2t)^{-1}}\right)}{\sqrt{2\pi}\sqrt{(1-2t)^{-1}}} dx \\
&= \sqrt{\frac{1}{1-2t}}.
\end{aligned}
$$

The last line was gotten here by realizing that the last integral is the integral over the whole real line of a pdf for $N(0,(1-2t)^{-1})$. Notice that this moment generating function exists for $|t| < 1/2$.

Now, then, we have that

$$M_{\chi^2}(t) \;=\; E(e^{X_1^2 t})^k \;=\; \left(\frac{1}{1-2t}\right)^{k/2}.$$

This moment generating function is the same as the one for a gamma distribution with parameters $r = k/2$ and $\alpha = 1/2$ provided $|t| < 1/2$. So, our theorem above then gives us that:

Theorem 2. If χ^2 is a chi-squared random variable having k degrees of freedom, and if Z is a random variable that obeys a gamma distribution with parameters $r = k/2$ and $\alpha = 1/2$, then we have that

$$P(\chi^2 \geq a) \;=\; P(Z \geq a).$$

Now recall that in the case where k is a positive even integer we get that $\Gamma(k/2) = (k/2-1)!$, which gives us that

$$P(\chi^2 \geq a) \;=\; P(Z \geq a) \;=\; \frac{1}{(k/2-1)!}\left(\frac{1}{2}\right)^{k/2}\int_a^\infty x^{k/2-1}e^{-x/2}dx.$$

Now, it was shown (or stated) in class that through a tedious integration-by-parts calculation, this right-most expression equals the sum of probabilities of a certain Poisson random variable. Specifically, let Y be a Poisson random variable with parameter $a/2$. Then, we have that

$$\begin{aligned} P(\chi^2 \geq a) \;&=\; P(Z \geq a) \;=\; \sum_{j=0}^{k/2-1} P(Y=j) \\ &=\; e^{-a/2}\sum_{j=0}^{k/2-1}\frac{(a/2)^j}{j!}. \end{aligned}$$

In the case where k is odd, there exists a similar formula, but it is much more involved. In our applications if k is odd we will either use a table lookup or else approximate the chi-squared random variable by a normal distribution via the Central Limit Theorem.

3 The chi-squared test statistic

We have seen one form of the chi-squared test already, which involved measurements of positions of objects that varied with time. In that case, we supposed that an object had a given velocity v (in some fixed direction away from the observer) and that at times $t = 1, 2, ..., 6$, a measurement of the position was made. We are assuming that we can repeat the experiment of measuring the particle many many times, and that $p(t)$ is a random variable giving the observed position of the particle at time t. Suppose we make the following predicition: The acutal position of the particle is at time t is t, and the discrepancy between the actual position and the observed position is due to faulty measuring equipment. Further suppose that the error $p(t) - t$ is $N(0, 1)$ for each of the times $t = 1, 2, ..., 6$. It is not a bad assumption that this error 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 a laser beam slightly, where that laser beam was 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 an obtain the following observed position values:

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

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

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

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.

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 percent 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 N of individulas, and checking to see which category they fall into. The expected number of individuals in class j is $e_j = p_j N$; and, suppose that the actual number observed is X_j. Note that $X_1 + \cdots + X_j = N$.

Define the parameter

$$E \;=\; \sum_{j=1}^{k} \frac{(X_j - p_j N)^2}{p_j N}. \tag{1}$$

Then, $E \geq 0$, and will be "large" if too many of the classes contain a number of individuals that are far away from the expected number.

In order to be able to check our hypothesis that "p_j percent 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 3. 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 the X_j's are not independent: We have that any X_j is completely determined by the other $k - 1$ values X_i's through $X_1 + \cdots + X_k = N$.

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 Independent, 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 Independent;
700 will vote Green; and,
200 will vote "other".

So, we have that

$$\begin{aligned} E &= \frac{(4200-4000)^2}{4000} + \frac{(3900-4000)^2}{4000} + 0 + \frac{(700-500)^2}{500} + \frac{(200-500)^2}{500} \\ &= 10 + 2.5 + 8 + 180 \ > \ 200. \end{aligned}$$

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 ≥ 200; that is,

$$P(\chi_4^2 \geq 200) \ = ?.$$

If this value of less than $\alpha = 0.05$, then we reject the hypothesis on the makeup of likely Florida voters.

After doing a table lookup, it turns out that

$$P(\chi_4^2 \leq 200) \ = \ 1.000..., \ \ \text{is very close to 1;}$$

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