Math 4581 Additional Notes I: The χ^2 Distribution and Two-way Tables

The χ^2 distribution is part of the family of gamma distributions and comes from the sum of the square of n independent standard normal random variables. Like many of the distributions in statistics, its distribution is difficult so probabilities for it are found using tables or algorithms. We will look at two applications that use it.

The χ^2 distribution with k degrees of freedom is: $\chi^2 = \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}$ for $x \geq 0$

Remember that we showed that Z^2 has a χ^2 distribution with k=1 (using $\Gamma(1/2) = \sqrt{\pi}$) and that the χ^2 distribution is a special case of the gamma distribution. We know that the sum of gamma distributions is also a gamma distribution which leads to:

$\sum_{i=1}^{n} Z^2$ is a χ^2 distribution with n degrees of freedom.

Goodness of Fit

Here we will be looking at probabilities using the **multinomial** distribution (you don't need to know the multinomial distribution, but I include it if you're interested).

Definition: Given n independent trials each of which leads to a success for exactly one of k categories, a distribution is multinomial if:

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_n = n_k) = \frac{n!}{n_1!n_2!\cdots n_k!}p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$$

where p_i is the probability of a success in category i.

Example: You roll a die 10 times and get the following outcomes:

outcome	1	2	3	4	5	6
Number of times (actual)	2	1	0	3	1	3

The probability of this is: $\dfrac{10!}{2!1!0!3!1!3!}(1/6)^{10}$

Given the results from a sample we will want to know if a process follows a given model (for example: is the die in the above example fair). The following theorem gives the criteria for the decision process.

Theorem: If there are k possible outcomes from n independent trials where X_i is the number of outcomes in category i, then $X^2 = \sum_{i=1}^{k} \dfrac{(X_i - np_i)^2}{np_i}$ is approximately χ^2 with $k-1$ degrees of freedom.

Note: this is an asymptotic result so we want each outcome to occur frequently (some texts say at there should be at least 5 in each category).

Note: Given the null hypothesis, you can see that the idea of a one-sided test doesn't make sense so we always look up the significance level as given.

Example: You will test H_0: $p_1 = .2, p_2 = .5, p_3 = .3$ against H_1: {not all equal} at the 5% level of significance. Suppose a sample of 100 finds:

outcome	1	2	3
Number of times	15	60	25
Expected number $(= 100 * p_i)$	20	50	30

The test statistic is: $X^2 = \dfrac{(15-20)^2}{20} + \dfrac{(60-50)^2}{50} + \dfrac{(25-30)^2}{30} = 4.083$

From the χ^2 table using 2 degrees of freedom, we get the critical value is 5.99.

Since the test statistic is smaller than the critical value, we fail to reject the null hypothesis and accept that the probabilities could be as given.

Two-way Tables

It turns out that this works almost exactly like the goodness of fit test.

Example: We look to see if three cities have similar age breakdowns. We get the following (with the numbers in thousands):

age/city	A	B	C	total
0-19	60	30	10	100
20-29	25	10	25	60
30-49	10	5	5	20
50-	5	5	10	20
total	100	50	50	200

To check to see if there is a difference, we will test to see if the variables (age and city) are independent. If they are independent then, for example, $P((0-19) \cap A)$ would equal $P(0-19)P(A)$.
From the example, P(0-19)=.5 and P(A)=.5 so we would expect $P((0-19) \cap A) = (.5)(.5) = .25$ and thus we would expect $200(.25) = 50$ in that cell. We can continue this to get a two-way table with expected numbers.

Note: the expected number $= \dfrac{(\#\ in\ row)(\#\ in\ column)}{n}$.

Using this we get the matrix with the expected numbers:

age/city	A	B	C	total
0-19	50	25	25	100
20-29	30	15	15	60
30-49	10	5	5	20
50-	10	5	5	20
total	100	50	50	200

Theorem: Given a two-way table with n observations partitioned into r rows and c columns with entries X_{ij}, then $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \dfrac{(X_{ij}-E_{ij})^2}{E_{ij}}$
is approximately χ^2 with $(r-1)(c-1)$ degrees of freedom

where $E_{ab} = (\sum_{j=1}^{c} X_{aj})(\sum_{i=1}^{r} X_{ib})/n$

Note: this is again an asymptotic result, so we again want the sample to be large. One criteria is that the average entry should be at least 5 and each expected count is at least one (for a 2×2 table each entry should be at least 5).

Back to the example. Let's suppose we are testing at the 5% level to see if the two variables are independent, so H_0: city and age are independent, H_1: they are not independent.

The test statistic is $X^2 = (60-50)^2/50 + (30-25)^2/25 + (10-25)^2/25 + (25-30)^2/30 + \ldots$
$= 28.67$.

The χ^2 table using 6 degrees of freedom, gives a critical value of 12.59.

Since the test statistic is bigger than the critical value we reject the null hypothesis and believe the cities have different age distributions.

Homework:

1. Test $H_0 : p_1 = .6, p_2 = .3, p_3 = .1$ against H_1: they are not all equal at the 5% level given the sample:

outcome	1	2	3
Number of times	100	40	10

2. Test $H_0 : p_1 = .4, p_2 = .3, p_3 = .2, p_4 = .1$ against H_1: they are not all equal at the 1% level given the sample:

outcome	1	2	3	4
Number of times	85	70	25	20

3. A sample of 100 is taken from a distribution that is thought to have the density: $f_X(x) = 2x$ for $0 \leq x \leq 1$. Test at the 5% level to see if the sample data fits this model

outcome	0-.25	.25-.5	.5-.75	.75-1
Number of times	10	20	30	40

4. Test at the 5% level to see if the distribution of class size (with X=number of students) is the same for the different semesters given the data:

	$X \leq 19$	$20 \leq X \leq 49$	$X \geq 50$
Spring	12	80	15
Summer	7	20	2
Fall	30	90	20

5. A random sample of 300 households is chosen and asked where they live and their income (in $thousands). Test at the 1% level to see if the variables are independent.

	≤ 50	$50 - 100$	$X \geq 100$
Inside Metro	131	74	37
Outside Metro	38	15	5

6. Test at the 5% level to see if the variables are independent.

	A	B	C	D
a	30	20	15	7
b	20	30	30	18

7. The two way table below has X^2=21.07 (show this). If you are testing at the 5% level of significance, decide if you accept or reject the null hypothesis (give the critical value).

	A	B	C	D	E
red	4	8	10	12	15
green	14	12	15	9	5
blue	15	2	16	8	9