TOPIC: The χ^2 (chi-square) distribution

The chi-square distribution allows us to test different hypotheses about frequencies (or proportions).

$$\chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

$f_o \equiv$ observed frequency
$f_e \equiv$ expected frequency

Since χ^2 is computed using a sum of squares its value can never be negative. The minimum value it can take on is 0. When the observed frequencies are exactly equal to the expected frequencies ($f_o = f_e$), the value of the χ^2 statistic is 0. The larger the value of the χ^2 statistic, the greater the discrepancy between the observed and expected frequencies. How large is too large? For that we have to compare the statistic we compute with the critical value from the χ^2 table.

As with the Student's t distribution, χ^2 is a series of distributions, a family of curves. There is a different χ^2 distribution for each value of degrees of freedom.
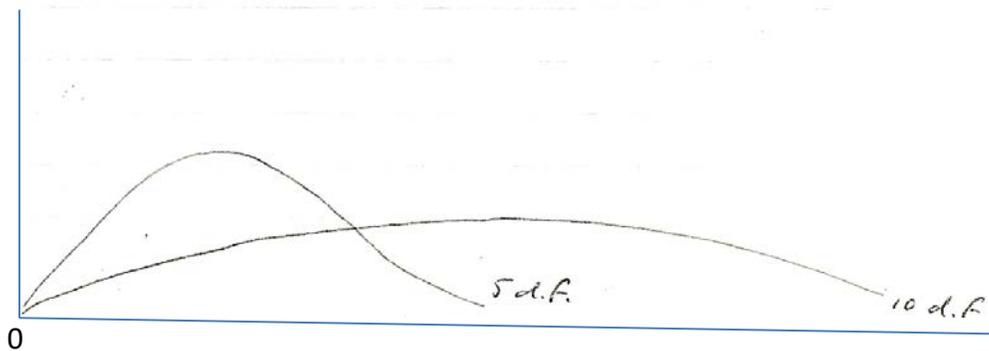
Essentially the hypothesis for a χ^2 test is always that the observed and expected frequencies are exactly the same, no difference between them. In practice, can use this test statistics for 3 types of tests:

χ^2 Test for Differences among the Proportions of C Populations

χ^2 Test for Independence – to determine whether two nominal variables are related

χ^2 Goodness of Fit Test – to determine whether a sample may be regarded as a random sample drawn from a population with a specified distribution (e.g., normal, binomial, etc.)

χ² Test for Differences among the Proportions of C Populations

We start with C=2 (compare with Z test for two proportions.)

In a sample of 200 individuals, researchers are interested in the proportions of males and females who use a particular brand of shampoo. Is there a difference between males and females with regard to their usage of the product? Test at α=.05

	MALE	FEMALE	
USER	20	50	70
NON-USER	80	50	130
	100	100	200

Hypotheses:
H_0: $P_1 = P_2$ (alternatively, sex and usage are independent)
H_1: $P_1 \neq P_2$ (alternatively, sex and usage are related)

Critical value of the χ² test statistic:
First we compute the degrees of freedom for the χ² test statistic, using the number of rows (r) and the number of columns (c).
Degrees of freedom = (r-1)(c-1)
So in this example, df = (2-1)(2-1) = 1

From the χ² table, the critical value of the χ² test statistic, with 1 degree of freedom and a tail probability (α) of .05, is: 3.841.

Calculated value of the χ² test statistic:
$$\chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

We can compute the expected frequencies (f$_e$) for each cell by:
$$f_e = \frac{row\ total * column\ total}{n}$$
Expected frequencies (f$_e$) are sometimes called theoretical frequencies (f$_t$).

Expected Frequencies (in red):

	MALE	FEMALE	
USER	35	35	70
NON-USER	65	65	130
	100	100	200

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/ f_e$
20	35	-15	225	6.43
80	65	+15	225	3.46
50	35	+15	225	6.43
50	65	-15	225	3.46
		0		19.78

Validity check: Make sure all the f_e are ≥ 5 or you may not have a valid χ^2 test.

Conclusion:

The calculated value of the test statistic (19.78) is greater than the critical value at the .05 level of significance (3.841). Hence,

Reject H_0

Women are more likely to use the product than men.

NOTE: When testing for differences between proportions of two populations, we can use either the Z test or the χ^2 test with exactly the same results. In this case, we could have done the problem using the Z test for the difference between 2 proportions:

$$Z = \frac{.20-.50}{\sqrt{(.35)(.65)\left(\frac{1}{100}+\frac{1}{100}\right)}} = \frac{-.30}{.06745} = -4.448$$

At $\alpha=.05$ and a two-tail test, the critical values of Z are ±1.96. Hence, Reject H_0
Also note that $\chi^2=(Z)^2$ so
$3.841 = (1.96)^2$ and $19.78 = (4.448)^2$]