Chi-squared distribution

Say you have some observed data and you want to know if the values you observed differ "significantly" from the values you expected to observe.

This is a form of hypothesis testing. You will have two hypotheses:

(1) The null hypothesis, denoted H_0, which states that the observed data does not differ "significantly" from our expected value.
(2) The alternative hypothesis, denoted H_1, which states that the observed data does differ "significantly" from our expected value.

You need to perform a "goodness of fit" test in order to which hypothesis you should accept/reject. But first let's define what we mean by "significant".

Definition 0.1. *Statistical significange* is, simply put, is a measure of the probability that you obtained your results by coincidence. Typically you want this probability to be low (usually less than 0.05 or 0.0, known as the level of significance). If your result meet your chosen level of significance, your results are said to be statistically significant.

To perform a goodness of fit test we use the χ^2 distribution.

Definition 0.2. Let $X_1, \ldots, X_n$ be independent random variables which have a standard normal distribution. Then the sum of their squares

$$Y = \sum_{i=1}^{n} X_n^2$$

has a chi-squared distribution with n degrees of freedom, and we write $Y \approx \chi_n^2$.

Example 0.3. Suppose that the number of males and females in the Masterson's school of Statistics is exactly equal. However over the past five years the number of males and females who received first class honours from in Biomedical Science is 40 and 80 respectively. Are these figures significantly different from what we expected?

Solution: Our "null" hypothesis is what we expected to observe, i.e. males and females receive the same number of first class honours.

The "alternate" hypothesis is that males and females don't receive the same number of first class honours. The following formula is used to calculate where we are in the χ^2 distribution, where O stands for observed values and E stands for expected values.

$$\chi^2 = \frac{(O - E)^2}{E}$$

	Female	Male	Total
Observed	80	40	120
Expected	60	60	120
$O - E$	20	-20	0
$(O - E)^2$	400	400	
$(O - E)^2 / E$	6.67	6.67	$13.34 = \chi^2$

The last calculation we need to do before is for our degrees of freedom. The degrees of freedom are equal to

$$n - 1,$$

where n is the number of categories we have. Here we have two categories, so we have one degree of freedom.

DEGREES OF FREEDOM	PROBABILITY										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
				Nonsignificant					Significant		

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., Table IV, Oliver & Boyd, Ltd., Edinburgh, 1963, by permission of the authors and publishers.

Our calculation for χ^2 exceeds the value for a p-value $= 0.05, 0.01$ and 0.001. So this means our observed data "significantly" differs from our expectations. We can reject our null hypothesis and accept the alternate.