χ^2 Test for Goodness of Fit

We have seen theoretical models of probability distributions, e.g. Binomial, Poisson, Exponential, Uniform etc, but need to ask how real data upon which such models are based, measure up in reality, i.e. how do the data fit the model.

On p. 34, consider 1000 random numbers, all uniformly generated over [0,1]. The real live observations are put into 20 class intervals.

We measure the disparity 'O−E', noting that the differences all sum to zero. Instead we take $(O-E)^2/E$ to allow for positive differences to be suitably normalised.

If our observations are :

x_i, $i = 1(1)n$ and $\bar{x}$ their mean, then the statistic $\sum_{i=1}^{n}(x_i-\bar{x})^2$ or sum of squares of differences from the mean, is the sum of squares of Normally distributed random variables. (note p 26/7)

(assumes $x_i-\bar{x}$ is Normally distributed)

In fact $\sum_{i=1}^{n}(x_i-\bar{x})^2 \sim \sum(O-E)^2/E$ is the sum of squares of 'n-1' random variables, all of which are $N(0,1)$.

This is a $\chi^2_{(n-1)}$ distribution with 'n-1' degrees of freedom. SMP handout p. 47.

Think of it as losing a degree of freedom because the contents of the bottom box are known when the preceeding 19 are known.

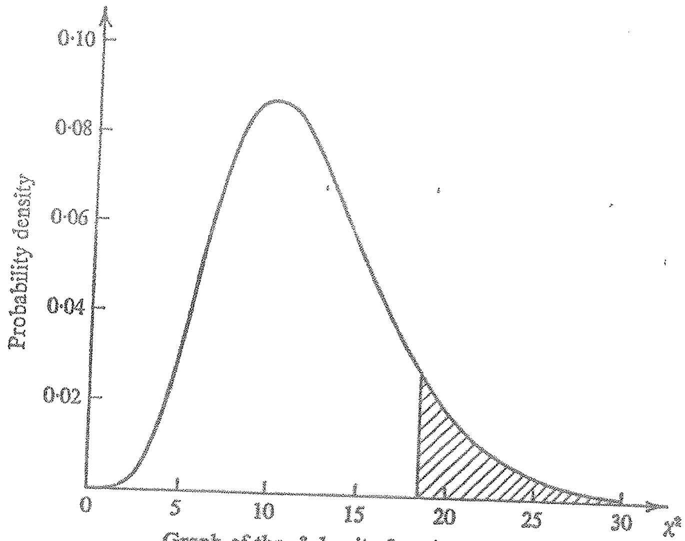
χ^2 PROBABILITY DENSITY FUNCTION

P	99	95	10	5	1	0·1
$\nu = 1$	0·000157	0·00393	2·71	3·84	6·64	10·83
2	0·0201	0·102	4·61	5·99	9·21	13·75
3	0·115	0·352	6·25	7·81	11·34	16·27
4	0·297	0·711	7·78	9·49	13·28	18·47
5	0·554	1·15	9·24	11·07	15·09	20·51
6	0·873	1·64	10·64	12·59	16·81	22·46
7	1·24	2·17	12·02	14·07	18·47	24·32
8	1·65	2·73	13·36	15·51	20·09	26·12
9	2·09	3·33	14·68	16·92	21·67	27·88
10	2·56	3·94	15·99	18·31	23·21	29·59
11	3·05	4·57	17·27	19·68	24·72	31·26
12	3·57	5·23	18·55	21·03	26·21	32·91
13	4·11	5·89	19·81	22·36	27·69	34·53
14	4·66	6·57	21·06	23·68	29·14	36·12
15	5·23	7·26	22·31	25·00	30·58	37·70
20	8·26	10·85	28·41	31·41	37·57	45·31
30	14·96	18·49	40·26	43·77	50·89	59·66
40	22·17	26·51	51·81	55·76	63·69	73·40
50	29·71	34·76	63·17	67·50	76·15	86·68
60	37·49	43·19	74·40	79·08	88·38	99·72

The function tabulated is χ^2_P defined by

$$\frac{P}{100} = \frac{\int_{\chi^2_P}^{\infty} x^{\frac{1}{2}\nu-1} e^{-\frac{1}{2}x} dx}{\int_0^{\infty} x^{\frac{1}{2}\nu-1} e^{-\frac{1}{2}x} dx}.$$

If x is a random variable with probability density function that of χ^2 with ν degrees of freedom then $P/100$ is the probability that $x \geqslant \chi^2_P$.



Graph of the χ^2 density function for $\nu = 12$

The graph illustrates that for $\nu = 12$, $\chi^2_{10} = 18{\cdot}55$. The shaded area is $\frac{10}{100}$ of the total.

For $10 < \nu < 70$ linear interpolation or extrapolation in ν is adequate. The percentage accuracy is better for small P than large P and for large ν than small ν.

Example. To estimate χ^2_P for $\nu = 70$, $P = 10$.

The difference between χ^2_{10} for $\nu = 50$ and $\nu = 60$ is 11·23. Our estimate is $\chi^2_{10} = 85{\cdot}63$. (The correct value is 85·53.)

BINOMIAL example

P 37/8

100 guns fired 100 times.
so 0 to 6 hits
The probability of a successful hit is <u>not known</u>, so it must be estimated from the data

$$\bar{x} = \frac{8\times0 + 95\times1 + 624\times2 + \cdots 1121\times6}{10,000}$$

$$= 4.20 = 6\times p\ , \text{ so } p = 0.7$$

Expected Frequencies = 'E'

$$10.000 \binom{6}{x}(0.7)^x(0.3)^{6-x},$$

$$x = 0(1)6$$

Hits	'O'	'E'
0	8	7
1	95	102
2	624	595
3	1786	1852
4	3257	3241
5	3109	3025
6	1121	1176
	10,000	10,000

Degrees of Freedom = $7 - 1 - 1^{(*)}$ $= 5$ $(\chi^2_{(5)})$ as $\bar{x}$ is used to estimate 'p'.

Poisson Example

No of calls	0	1	2	3	4	5
Freq of days	34	37	25	3	1	0

$$\bar{X} = \frac{34\times 0 + 37\times 1 + 25\times 2 + 3\times 3 + 4\times 1}{100}$$

$$= 1{,}00.$$

We assume that the expected values follow a Poisson distribution with $\mu = 1$. In other words we need the 'statistical' parameter $\bar{X} = 1$. This accounts for 1 degree of freedom.

Note $E_j = 100 \times P(X=j)$
and $P(X=j) = \dfrac{e^{-\mu}\mu^j}{j!}$.

We get:

j	O_j	E_j
0	34	36.79
1	37	36.79
2	25	18.39
3+	4	8.03

Computing $\sum (O_j - E_j)^2 / E_j$ gives 4.61 for $\chi^2_{(2)} < 5.99$.

Hypothesis Testing

We can define a Null Hypothesis H_0, that the data fit the distribution, subject to using statistical parameter etc. An alternative hypothesis H_1 might be that they do not fit the distribution. By definition this is <u>one-tailed</u>, see p. 37.

Contingency Tables

O_{11}	O_{12}	O_{1n}	R_1
O_{21}			R_2
$\vdots$	$\boxed{O_{ij}}$		
O_{m1}	. . . .	O_{mn}	R_m
C_1	C_2	C_n =	T

With no bias, $E_{ij} = \dfrac{R_i C_j}{T}$ (overall total)

$\chi^2 = \sum_{i,j} (O_{ij} - E_{ij})^2 / E_{ij}$ is distributed with $(m-1)\times(n-1)$ degrees of freedom.

In a drug trial, effectiveness is measured by administering the new drug to some patients but not others. 70 patients are split into two groups of 35. The results are as follows:

	Recover	Die	
Drug	20	15	35
No drug	13	22	35
	33	37	

Compute $E_{ij} = \frac{R_i C_j}{T}$

then detemine $\chi^2_{(1)}$.

Is the value significant?

Ho: no difference.