Chapter 33
Goodness of Fit Tests for Identifying Nonnormal Data

1 General Purpose

Goodness of fit for assessing normal distribution of a data file is an important requirement for normal and t-distributed tests to be sensitive for statistical testing the data. Data files that lack goodness of fit can be analyzed using distribution free methods, like Monte Carlo modeling and neural network modeling (SPSS for Starters, Part 2 from the same authors, Chaps. 18 and 19, Springer New York, 2012, from the same authors).

The chi-square and the Kolmogorov-Smirnov goodness of fit tests are pretty much similar, but one uses the differences between all observed and expected observations, while the other uses the single largest difference between observed and expected observations, and, so, results may not be identical. One test may, however, very well be used as a complementary test or contrast test to the other.

2 Schematic Overview of Type of Data File

Outcome
.
.
.
.
.
.
.
.

(continued)

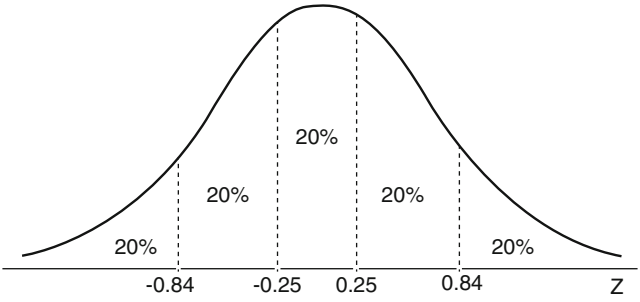
T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_33

Outcome
.

3 Primary Scientific Question

Can goodness of fit tests adequately identify nonnormal/non-t-distributed data.

4 Chi-Square Goodness of Fit Test, Data Example



With the help of the t-table the areas under the curve (AUCs) of 5 intervals of the null-hypothesis of a normal frequency distribution can be computed. The cut-off results (z-values) for the 5 intervals with an AUC of 20 % are −0.84, −0.25, 0.25, and 0.84 (AUC = area under the curve). These cut-off results can, subsequently, be used for chi-square goodness of fit testing. A data example is given.

In random not-too-small populations body-weights follow a normal distribution. Is this also true for the body-weights of 50 patients treated with a weight reducing compound?

Individual weight (kgs)

85	57	60	81	89	63	52	65	77	64
89	86	90	60	57	61	95	78	66	92
50	56	95	60	82	55	61	81	61	53
63	75	50	98	63	77	50	62	79	69
76	66	97	67	54	93	70	80	67	73

As the area under the curve (AUC) of a normal distribution curve is divided into 5 equiprobable intervals of 20 % each, we will expect approximately 10 patients per interval. From the data a mean and standard error (SE) of 71 and 15 kg are calculated. In order to compute the numbers of patients of our example in each interval, we will use the underneath equation.

$$z = \text{standardized result} = \frac{\text{unstandardized result} - \text{mean result}}{\text{SE}}$$

For example for the cut-off value of -0.84 the unstandardized result of 58.40 kg can be computed.

$-0.84 = (\text{unstandardized result} - 71)/15$
$\text{unstandardized results} = (15 \times -8.84) + 71 = -58.40$ kg.

All of the unstandardized results (kgs) are given underneath:

$-\infty$.... 58.40... 67.25... 74.25... 83.60... ∞

As they are equiprobable,

As they are equiprobable, we expect per interval:	10 pts	10 pts	10 pts	10 pts	10pts
We do, however, observe the following numbers:	10 pts	16 pts	3 pts	10 pts	11pts

The chi-square value is calculated according to

$$\sum \frac{(\text{observed number} - \text{expected number})^2}{\text{expected number}} = 8.6$$

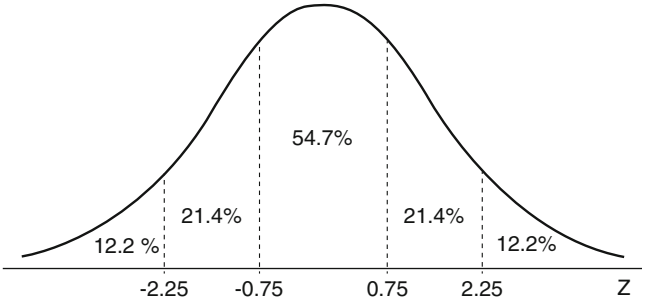
This chi-square table is given underneath. It has an upper row with areas under the curve, a left-end column with degrees of freedom, and a whole lot of chi-square values.

Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

The chi-square value of 8.6 means that for the given degrees of freedom of $5–1=4$ (there are 5 different intervals) the null-hypothesis of no-difference-between-observed-and-expected can not be rejected. However, our p-value is <0.10, and, so, there is a trend of a difference. The data may not be entirely normal, as expected. This may be due to lack of randomness.

5 Kolmogorov-Smirnov Goodness of Fit Test, Data Example



With the help of the t-table the areas under the curve (AUCs) of again 5 intervals of the null-hypothesis of a normal frequency distribution can be computed. The cut-off results (z-values) for the 5 intervals are calculated to be −2.25, −0.75, 0.75, and 2.25. The corresponding AUCs are given in the above graph (AUC = area under the curve).

In random not-too-small populations plasma cholesterol levels follow a normal distribution. Is this also true for the plasma cholesterol levels of the underneath patients treated with a cholesterol reducing compound? A data sample of 750 patients is given.

Cholesterol (mmol/l)	<4.01	4.01–5.87	5.87–7.73	7.73–9.59	>9.59
Numbers of pts	13	158	437	122	20

The cut-off results for the 5 intervals must be standardized to find the expected normal distribution for these data according to

$$z = \text{standardized cut-off result} = \frac{\text{unstandardized result} - \text{mean result}}{\text{SE}}$$

With a calculated mean (SE) of 6.80 (1.24) we must compute the unstandardized results corresponding with the z-values −2.25, −0.75, 0.75 and 2.25. For example, with z = −2.25, the unstandardized z-value is calculated.

$$-2.25 = (\text{unstandardized result} - 6.80)/1.24$$

unstandardized result

$$= (1.24 \times -2.25) + 6.80 = 4.01 \text{ mmol/l.}$$

Similarly all unstandardized z-values are computed.

With 750 cholesterol-values in total the expected frequencies of cholesterol-values in the subsequent intervals are

$12.2 \times 750 = \quad 9.2$
$21.4 \times 750 = 160.8$

$54.7 \times 750 = 410.1$
$21.4 \times 750 = 160.8$
$12.2 \times 750 = 9.2$

The observed and expected frequencies are, then, listed cumulatively (cumul = cumulative):

Frequency observed	cumul	relative (cumul/750)	expected	cumul (cumul/750)	relative	cumul observed-expected
13	13	0.0173	9.2	9.1	0.0122	0.0051
158	171	0.2280	160.98	170.0	0.2266	0.0014
437	608	0.8107	410.1	580.1	0.7734	0.0373
122	730	0.9733	160.8	740.9	0.9878	0.0145
20	750	1.000	9.2	750	1.000	0.0000

According to the Kolmogorov-Smirnov table below, the largest cumulative difference between observed and expected should be smaller than $1.36/\sqrt{n} = 1.36/\sqrt{750} = 0.0497$, while we find 0.0373. This means that these data are well normally distributed.

Level of statistical significance for maximum difference between cumulative observed and expected frequency (n = sample size)

	Areas under the curve				
n	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.463
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356

25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Over 35	1.07 $\sqrt{n}$	1.14 $\sqrt{n}$	1.22 $\sqrt{n}$	1.36 $\sqrt{n}$	1.63 $\sqrt{n}$

6 Conclusion

Goodness of fit for assessing normal distribution of a data file is an important requirement for normal and t-distributed tests to be sensitive for statistically testing the data. The chi-square and the Kolmogorov-Smirnov goodness of fit tests are adequate for the purpose, and are pretty much similar, but results need not be identical.

7 Note

More background, theoretical and mathematical information of goodness if fit testing is given in Statistics applied to clinical studies 5th edition, Chap. 42, Springer Heidelberg Germany, 2012, from the same authors.