

Testing for Normality – Prof. Richard B. Goldstein

How does one know if the data fits a normal distribution?

Cholesterol Data from the Framingham Heart Study

Stem-and-leaf plot	Freq	Cumul Freq
16 7	1	1
17	0	1
18 4	1	2
19 28	2	4
20 02	2	6
21 0125678	7	13
22 0556	4	17
23 0000122244668	13	30
24 03678	5	35
25 444668	6	41
26 347778	6	47
27 00288	5	52
28 35	2	54
29	0	54
30 008	3	57
31	0	57
32 7	1	58
33 46	2	60
34	0	60
35 3	1	61
36	0	61
37	0	61
38	0	61
39 3	1	62

Descriptive Statistics:

$$n = 62, \bar{X} = 250.03, s = 41.44, \sqrt{b_1} = \frac{m_3}{m_2^{3/2}} = 1.024, b_2 = \frac{m_4}{m_2^2} = 4.577 \text{ where } m_k = \frac{\sum (X_i - \bar{X})^k}{n}$$

Tests to Consider

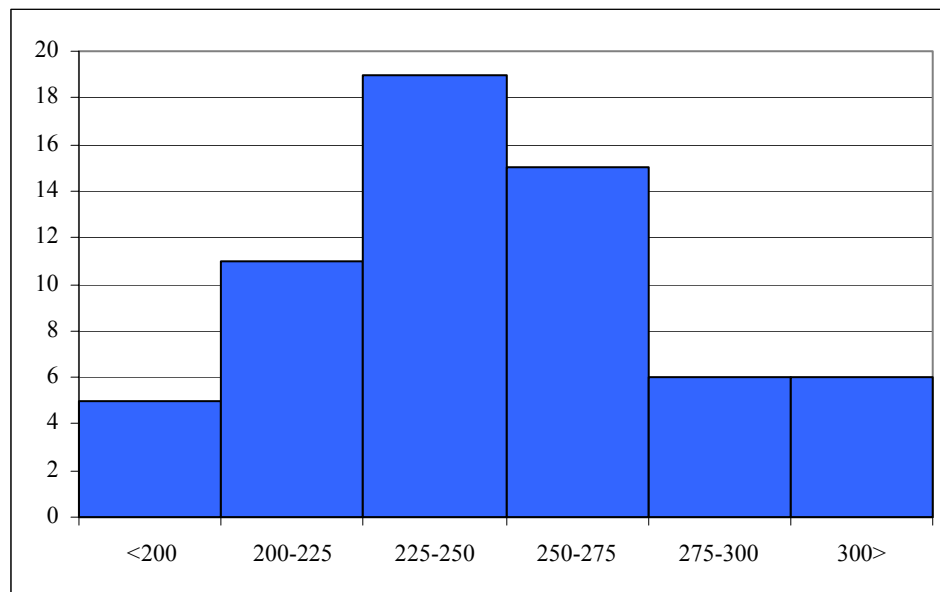
- [1] Chi-Square Goodness-of-Fit Test
- [2] Graphics – Normal Probability Plot
- [3] Non-Parametric Kolmogorov-Smirnov Test
- [4] Geary's Test
- [5] D'Agostino-Belanger-Pearson's Skewness, Kurtosis, and Omnibus Tests
- [6] Others – Shapiro-Wilk, Anderson-Darling, Cramér-von Mises Tests

[1] Chi-Square Goodness-of-Fit Test

mean 250.0323
 stdev 41.44321

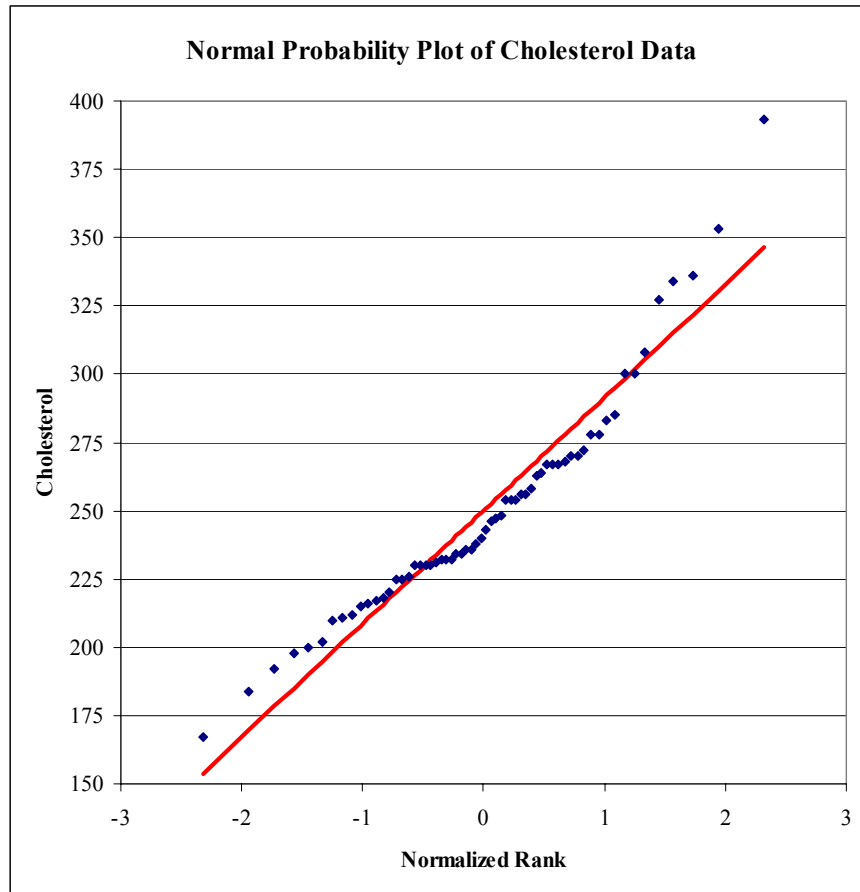
Bin	Obs Freq	z	Exp Freq	$(O_i - E_i)^2 / E_i$
<200	5	-1.20725	7.04743	0.594822
200-225	11	-0.60401	9.873446	0.128539
225-250	19	-0.00078	14.05987	1.735782
250-275	15	0.602457	14.06628	0.061981
275-300	6	1.205692	9.886948	1.528112
300>	6		7.066027	0.160828
sum	62		62	$\chi^2 = 4.210064$ p = 0.239656

For example $-1.2075 = (200 - \text{mean})/\text{stdev}$ and $7.04743 = 62 * \text{NORMDIST}(7.04743, 0, 1, 1)$



By this test the data would be accepted as fitting a normal distribution.
 Note that the values in the tails had to be combined to assure at least 5 in each cell.

[2] Normal Probability Plot



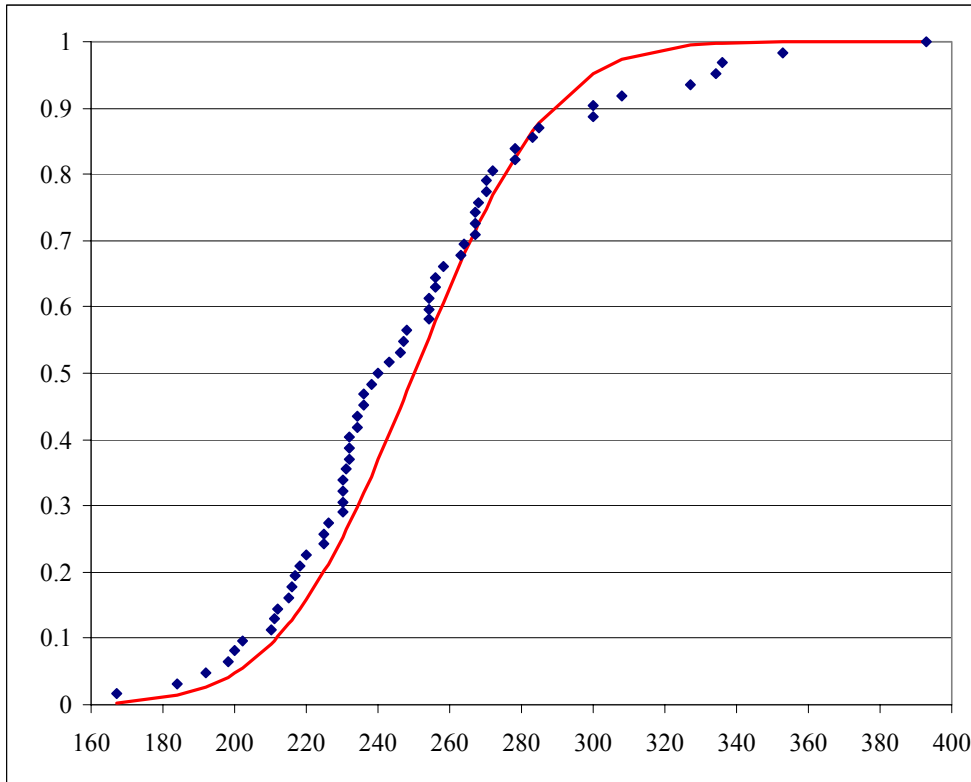
i	$(i-0.375)/(n+0.25)$	$Z_i = \Phi^{-1}[(i-0.375)/(n+0.25)]$	Data	$m + Z_i s$
1	0.010040	-2.324844	167	153.6833
2	0.026104	-1.941408	184	169.5741
3	0.042169	-1.726056	192	178.4990
...
61	0.973896	1.941408	353	330.4904
62	0.989960	2.324844	393	346.3812

The normal probability plot is a plot of $Z = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$ on $X_{(i)}$ where $X_{(i)}$ is the i^{th}

ordered sample $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ and Z is the value such that $\frac{i-3/8}{n+1/4} = \int_{-\infty}^Z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

for $I=1, 2, \dots, n$. In the above plot the skewness to the right is very evident and indicates that the data set may not be normally distributed.

[3] Non-parametric Kolmogorov-Smirnov Test



Data	cumul f	normal	D+	D-
167	0.016129	0.002832	0.0133	0.0028
184	0.032258	0.013903	0.0184	-0.0022
192	0.048387	0.026598	0.0218	-0.0057
...				
236	0.467742	0.320369	0.1474	-0.1312
...				
353	0.983871	0.999702	-0.0158	0.0320
393	1	0.999999	0.0000	0.0161

$$D_n^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F(x_{(i)}, \theta) \right\}, D_n^- = \max_{1 \leq i \leq n} \left\{ F(x_{(i)}, \theta) - \frac{i-1}{n} \right\}, D_n = \max(D_n^+, D_n^-)$$

$$P\{\sqrt{n}D_n > t\} \rightarrow K(t) = 2\left(e^{-2t^2} - e^{-8t^2} + e^{-18t^2} - \dots \pm e^{-2k^2t^2} \pm \dots\right)$$

Using a normal distribution with an arbitrary parameters $\mu = 150$ and $\sigma = 28$ gives the graph above, a value of $D_n = 0.1474$ for $n = 62$ and a p-value of 0.0863. This is only a marginal indication of non-normality. This test may be used for any probability distribution. This test requires presumed values of μ and σ and should not be based upon the sample data. Other tests are more powerful because they are for made for a specific distribution.

[4] **Geary's Test**

$$U = \frac{\sqrt{\frac{\pi}{2}} \sum_{i=1}^n |X_i - \bar{X}| / n}{\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}} \quad Z = \frac{U - 1}{\frac{0.2661}{\sqrt{n}}} \quad \text{for a normal distribution } U = 1$$

$$U = \frac{38.979}{41.108} = 0.9482 \quad Z = -1.532 \quad \rightarrow \quad P = 0.1255 \text{ on a two-sided test}$$

[5] **D'Agostino-Belanger-Pearson's Skewness, Kurtosis, and Omnibus Tests**

For large n skewness, $\sqrt{b_1}$, has an approximately normal distribution with $\sigma \approx \sqrt{\frac{6}{n}}$

For large n kurtosis, b_2 , has an approximately normal distribution with $\sigma \approx \sqrt{\frac{24}{n}}$

Quick approximation for large n

$\sqrt{b_1}$	1.023548	b_2	4.57739
N	62	n	62
$\sqrt{(6/n)}$	0.31109	$\sqrt{(24/n)}$	0.62217
Z	3.29025	z	2.53530
P	0.00100	p	0.01124

These results show skewness is highly significant above normal ($p = 0.001$) and kurtosis is significant ($p = 0.011$) as well. But n should be at least 100 for such an approximation. For skewness when $n = 125$, the value of p is accurate within 1% (0.01) and for kurtosis that accuracy is achieved at $n = 105$. Since $n = 62$ a more accurate approach is suggested below.

In a paper by R. D'Agostino, A. Belanger, and R. D'Agostino, Jr. the following approach is used:

Test of Skewness

1. Compute $\sqrt{b_1}$ from the sample data.
2.
$$Y = \sqrt{b_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$
3.
$$\beta_2(\sqrt{b_1}) = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$
4.
$$W^2 = -1 + \sqrt{2\{\beta_2(\sqrt{b_1}) - 1\}}$$
5.
$$\delta = 1/\sqrt{\ln W}$$
6.
$$\alpha = \sqrt{\frac{2}{W^2 - 1}}$$
7.
$$Z(\sqrt{b_1}) = \delta \ln \left\{ \frac{Y}{\alpha} + \sqrt{\left(\frac{Y}{\alpha}\right)^2 + 1} \right\} = \delta \sinh^{-1} \left(\frac{Y}{\alpha} \right)$$

Z is approximately normally distributed under the null hypothesis of population normality.

Test of Kurtosis

1. Compute b_2 from the sample data
2.
$$E(b_2) = \frac{3(n-1)}{n+1}$$
3.
$$\text{var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$
4.
$$x = \frac{b_2 - E(b_2)}{\sqrt{\text{var}(b_2)}}$$
5.
$$\sqrt{\beta_1(b_2)} = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$
6.
$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{1 + \frac{4}{\beta_1(b_2)}} \right]$$
7.
$$Z(b_2) = \frac{\left(1 - \frac{2}{9A}\right) - \left(\frac{1 - \frac{2}{A}}{1 + x \sqrt{\frac{2}{A-4}}}\right)^{1/3}}{\sqrt{\frac{2}{9A}}}$$

Z is approximately normally distributed under the null hypothesis of population normality.

Omnibus Test

$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$ has approximately a χ^2 distribution with 2 df when the population is normally distributed.

Fisher g statistics (unbiased estimates used in Excel, SPSS) vs. moments

Skewness: $g_1 = \frac{n \sum (X - \bar{X})^3}{(n-1)(n-2)S^3}$ $\sqrt{b_1} = \frac{(n-2)}{\sqrt{n(n-1)}} g_1$

Kurtosis: $g_2 = \frac{n(n+1) \sum (X - \bar{X})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$ $b_2 = \frac{(n-2)(n-3)}{(n+1)(n-1)} g_2 + \frac{3(n-1)}{n+1}$

where $S^2 = \frac{n}{n-1} m_2 = \frac{\sum (X - \bar{X})^2}{n-1}$ is the unbiased estimate of variance

Results for Framingham Heart Study

D'Agostino-Belanger-Pearson			
$\sqrt{b_1}$	1.023548	b_2	4.577388
N	62	n	62
Y	3.452104	E(b_2)	2.904762
$\beta_2(\sqrt{b_1})$	3.398433	var(b_2)	0.304745
W^2	1.190175	x	3.029915
δ	3.38934	$\sqrt{\beta_1}(b_2)$	1.494267
α	3.242935	A	22.11069
Z($\sqrt{b_1}$)	3.139392	Z(b_2)	2.212628
p	0.001693	p	0.026923

Omnibus	K^2	14.7515
	p	0.000626

Here the skewness has a probability $p = 0.0017$ of being from a normal distribution and kurtosis has a p-value of 0.0269 with the combined test (Omnibus Test) with a p-value of 0.000626.

Conclusion: not a normal population distribution

Other Tests

Shapiro-Wilk http://en.wikipedia.org/wiki/Shapiro-Wilk_test
Anderson-Darling http://en.wikipedia.org/wiki/Anderson-Darling_test
Cramér-von-Mises http://en.wikipedia.org/wiki/Cram%C3%A9r-von-Mises_criterion
or http://en.wikipedia.org/wiki/Cramér-von-Mises_criterion

References

- [1] A Suggestion for Using Powerful and Informative Tests of Normality. Ralph B. D'Agostino; Albert Belanger; Ralph B. D'Agostino, Jr., *The American Statistician*, Vol. 44, No. 4 (Nov., 1990), pp. 316-321.
- [2] Thode, H.C., *Testing for Normality*, Marcel Dekker, New York (2002).
- [3] Moments of the Ratio of the Mean Deviation to the Standard Deviation for Normal Samples, R.C. Geary, *Biometrika*, Vol. 28, No. 3/4., (Dec. 1936), pp. 295-307.
- [4] Tests for Normality, <http://www.vub.ac.be/BFUCC/sas/sasdoc/qc/chap1/sect19.htm>