MedicalBiostatistics.com

Checking Gaussianity

Requirement of Gaussian pattern of original values is not rigorous for most sampling distributions. Yet, many times you would want to know whether the pattern is Gaussian or not – for example to decide whether median would be more appropriate central value or mean. For sampling distribution also, knowledge about the pattern of original values is helpful. For example, there is a tendency for small sample $\bar{x}$ to follow the same kind of distribution as the individual *x*s. The distribution of duration of labor in childbirth is known to be skewed to the right. That is, a long duration is more common than a short duration relative to mode. The distribution of mean duration in a small sample of, say, eight women is also likely to follow the same pattern, although in attenuated form. In such cases, if the pattern is not known, it is worthwhile to investigate.

Many times the biological process underlying a medical measurement provides sufficient clue that the distribution of a particular measurement is Gaussian or not. In all other cases, you will be required to make a judgment on the basis of the data you have on a sample of subjects. In this situation, you can use any of the following methods. These methods work well for large *n* but may fail for small *n*. When *n* is small, you may have to use your subjective judgment.

A. Simple but Approximate Methods

First two methods given next may fail to detect the deviation in peakedness (kurtosis) of the distribution though they are adequate for skewness.

1) If you are calculation oriented, just calculate coefficient of skewness. Actual procedure for calculating coefficient of skewness is complex as it requires sum of the cubes of deviations from mean. A simple procedure is to calculate

$$\text{Coefficient of skewness} \approx \frac{\text{mean} - \text{mode}}{\text{SD}}. \qquad (1)$$

This works reasonably well for unimodal (single-peak) distributions. Negative value indicates left skewness and positive value right skewness. For a symmetric distribution, this coefficient is zero. A value < -1 or $> +1$ indicates highly skewed distribution.

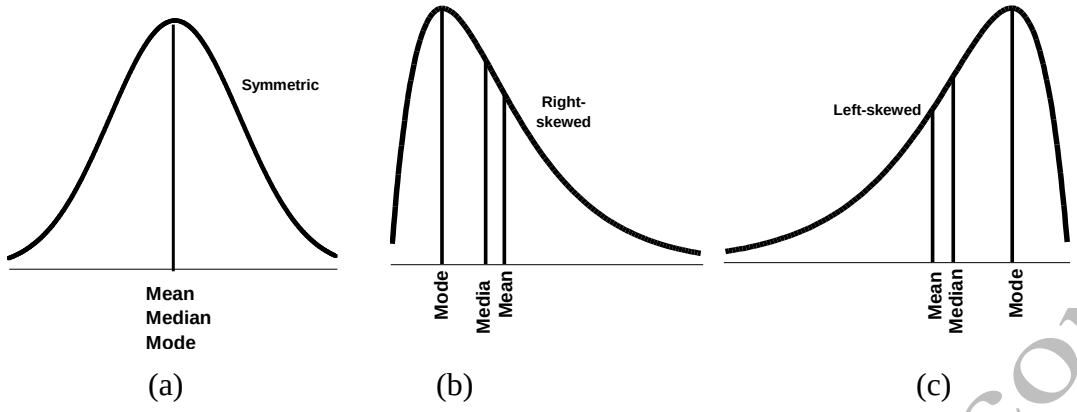

FIGURE 1: Location of mean, median and mode in (a) symmetric, (b) right-skewed and (c) left-skewed distributions

In a Gaussian distribution – in fact in all symmetric unimodal distributions – mean, median and mode are equal (Figure 1a). In sample values, this could be approximately so. For others, note the following:

Right skewed distribution: Mode < Median < Mean

Left skewed distribution: Mean < Median < Mode

These are also shown in Figure 1b and 1c. Incidentally, these words appear in a dictionary in the order seen for left skewed distribution and reverse in the right skewed distribution. Also the distance between Mean and Median in a dictionary is small relative to distance between Median and Mode. The coefficient of skewness (1) is also based on these considerations. Thus the first method to find that a distribution is symmetric or not is to calculate mean, median and mode, and see if they follow any of the above mentioned patterns. In samples, the difference between mean, median and mode must be substantial for the distribution to be considered skewed.

2) If you are graph oriented, most basic to check Gaussianity is histogram. Draw it for frequencies in different class intervals and see if it largely follows a bell-shape or not. An alternative to histogram is stem-and-leaf plot. Second approximate method is **quartile plot** of the type shown in Figures 2a, b, c. For this, compute Q_1, Q_2 and Q_3 and plot them on Minimum to Maximum axis. If the distance between Q_1 and Q_2 is nearly the same as between Q_2 and Q_3, you can safely assume symmetry and possibly Gaussian. If the pattern is different as in Figures 2b and 2c, the distribution is either left-skewed or right-skewed. If the sample size *n* is really large, you may like to try this type of plot with deciles instead of quartiles.

An alternative to quartile plot is the box plot. For symmetry, the boxes above and below the median as well as the whiskers on both sides should be nearly equal.

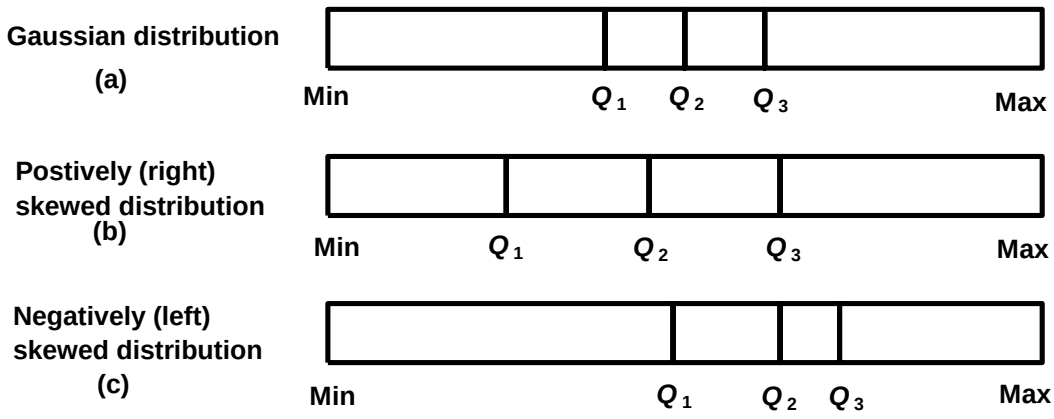

FIGURE 2: Pattern of quartiles in (a) symmetric (Gaussian), (b) right-skewed and (c) left-skewed distributions

3) Whereas the methods just described are for symmetry, the following methods check all aspects of Gaussianity including kurtosis. First such graphic method is ogive. You know that ogive is the plot of cumulative frequencies against the *x*-values. In a Gaussian distribution, this takes the shape of a sigmoid. If the shape is substantially different, take it as an indication of nonGaussian shape. You can also try to plot the cumulative relative frequencies in your sample vs. cumulative probabilities of Gaussian distribution. This is called **P-P (proportion-by-probability) plot**. If the distribution is Gaussian, this will be nearly a straight line. If substantially different, suspect that the distribution is not Gaussian. In place of P-P, you can also try **quantile-by-quantile (Q-Q) plot**. This plots observed quantile for each distinct value against expected quantile for that value under Gaussian pattern. This should also provide nearly a straight line.

4) The other alternative is to plot the standardized deviate against the observed values on Gaussian probability paper. Good statistical software will give such a plot. If this plot is a straight line or nearly so, then a Gaussian pattern can be safely assumed. You can also check if mean ± 1SD covers nearly two-thirds and mean ± 2SD nearly 95% of the values. The range should be nearly 6SD.

More exact methods for checking statistical significance of the departure from Gaussian. require intricate calculations.

B. Significance Tests for Assessing Gaussianity

Although the following tests are described in the context of assessing Gaussianity, the methods are general and can be used to assess whether the observed values fall into any specified pattern. Ironically, all these methods require large *n* in which case the sampling distribution of $\bar{x}$ tends to be Gaussian anyway. The methods would still be useful if your interest is in assessing the distribution of original values rather than of sample mean.

Among several statistical tests for Gaussianity, a useful method is goodness-of-fit test based on chi-square. This is based on proportions in various class-intervals. Among others, more popular are Shapiro-Wilk test, Anderson-Darling test and Kolmogorov-Smirnov test. All these are mathematically complex that are being avoided in this site. Statistical software packages generally have a routine for these tests that you can easily apply. However, it is important that you understand the implications.

Shapiro-Wilk test focuses on lack of symmetry particularly around the mean. This test is not much sensitive to differences present towards the tails of the distribution. Opposed to this, **Anderson-Darling test** emphasizes lack of Gaussian pattern in the tails of the distribution. This test performs poorly if there are many ties in the data. That is, for this test, the values must be truly continuous. **Kolmogorov-Smirnov test** works well for large *n* when mean and SD are known a priori and do not have to be estimated from the data. This also tends to be sensitive near the center of the distribution than at the tails.

Critical value beyond which the hypothesis is rejected in Anderson-Darling test is different when Gaussian pattern is being tested than when another distribution such a lognormal is being tested. Shapiro-Wilk critical value also depends on the distribution under test. But Kolmogorov-Smirnov test is distribution-free as the critical values do not depend on whether Gaussianity is being tested or some other form.

May sound strange to some but all these statistical tests cannot confirm Gaussianity although they confirm, with reasonable confidence, lack of it when present. Gaussianity is presumed when its lack is not detected. For reasonable assurance of Gaussianity, equivalence tests can be possibly devised.