

ENDGAMES

STATISTICAL QUESTION

The normal distribution

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, Tooting, London, UK

Researchers investigated whether antidepressant drug prescribing and referrals to specialist services for depression by general practitioners were associated with patients' scores on depression severity questionnaires measured at the start of treatment. Anonymised data from the medical records of 38 general practices were investigated. The nine item patient health questionnaire (PHQ-9) was used in 1658 patients and the depression subscale of the hospital anxiety and depression scale (HADS) was used in 548 patients.¹

The distribution of HADS questionnaire scores for the sample was approximately normal, whereas the PHQ-9 scores were negatively skewed. The mean PHQ-9 score was 15.5 (standard deviation 6.0) and the mean HADS score was 11.1 (4.6). It was reported that patients with higher depression severity scores were more likely to receive prescriptions for antidepressants and be referred to specialist services. However, other factors were independently associated with treatment and referral, including patient age and concurrent physical illness and geographical area.

Which of the following statements, if any, are true?

- a) The normal distribution approximated by the HADS scores can be uniquely described by the sample mean and standard deviation
- b) The normal distribution approximated by the HADS scores is symmetrical about the sample mean value
- c) For the distribution of PHQ-9 scores, the tail on the right was longer than the tail on the left
- d) The sample mean PHQ-9 score was smaller than the median PHQ-9 score

Answers

Statements *a*, *b*, and *d* are true, whereas *c* is false.

Figure 1 shows the distribution of the HADS and PHQ-9 questionnaire scores in the above study. The distribution of HADS questionnaire scores for the sample was approximately normal. The normal distribution is theoretical and described by an equation representing a mathematical model. The properties of this theoretical distribution are used to describe and answer questions about the empirical distribution of the HADS

questionnaire scores. The normal distribution is a probability distribution that describes the probability of each score occurring in the population as estimated by the sample.

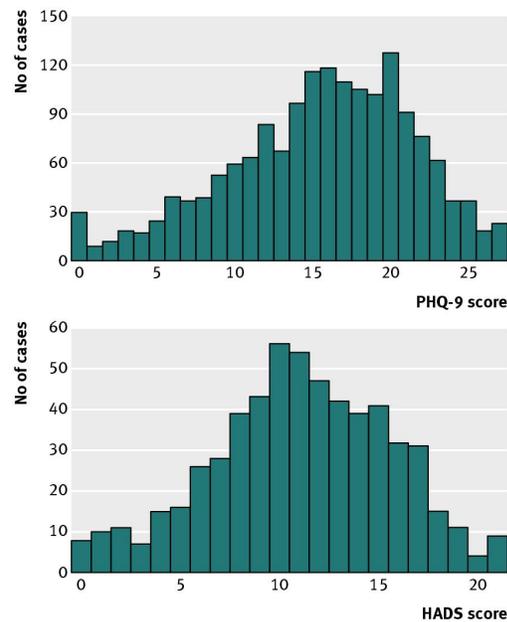


Fig 1 Distributions of depression severity scores among general practice patients with depression. Depression was assessed using the nine item version of the patient health questionnaire (PHQ-9) (n=1658) or the depression subscale of the hospital anxiety and depression scale (HADS) (n=584)

Figure 2 shows the theoretical normal distribution approximated by the patients' HADS questionnaire scores. It resembles the curve that would be obtained if a smooth line was drawn through the top of each bar in the histogram of the scores in fig 1. This normal distribution is uniquely described by its mean and standard deviation, equivalent to the mean and standard deviation of the HADS questionnaire scores for the patients in

the above example (a is true). The mean describes the average HADS questionnaire score for the sample, whereas the standard deviation describes the spread of scores about the sample mean. The normal distribution, also known as the Gaussian distribution, is bell shaped and symmetrical about its mean (b is true). Because the distribution of test scores is symmetrical, the median equals the mean.

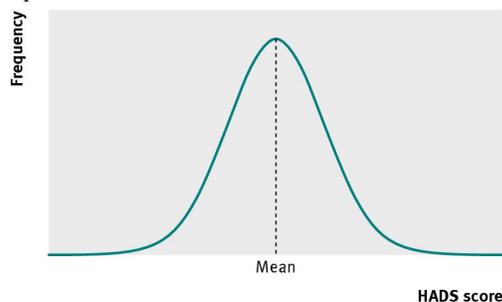


Fig 2 Theoretical normal distribution approximated by the hospital anxiety and depression scale (HADS) questionnaire scores for the sample of patients, uniquely described by the arithmetic mean of 11.1 and standard deviation of 4.6

The normal distribution is typically described by reference to the central hump that contains most of the observations, plus the tails or extremes of the distribution. If the standard deviation of the sample's HADS questionnaire scores increased, the hump of the normal distribution would flatten and more scores would be obtained in the tails of the distribution. If the standard deviation of the scores decreased, fewer patients would have a score in the extremes of the distribution, resulting in a greater proportion of scores in the central hump and an increase in its height.

The distribution of PHQ-9 questionnaire scores (fig 1) is not symmetrical about its mean and it is referred to as skewed. The tail to the left of the distribution incorporates low scores and is longer than the tail on the right (c is false). The bulk of scores is concentrated to the right of the distribution. The distribution

of PHQ-9 questionnaire scores is termed negatively skewed or skewed to the left. Conversely, if the distribution had a long tail to the right, with the bulk of observations concentrated to the left, then the distribution would be said to be positively skewed or skewed to the right. The name given to a skewed distribution is dictated by the side of the longer tail—negatively skewed or skewed to the left if the left hand tail is longer, positively skewed or skewed to the right if the tail on the right is longer.

The distribution of PHQ-9 questionnaire scores was skewed to the left and therefore the median score would not equal the mean value. The median PHQ-9 score is unaffected by extreme values in the left tail of the distribution because, by definition, half the patients have a score less than the median score and half have a score greater than it. However, these low scores decrease the mean value. Therefore, because the distribution of PHQ-9 questionnaire scores is negatively skewed, the mean is smaller than the median value (d is true). If the distribution had been positively skewed, the mean would be larger than the median value.

Many physical, biological, and social variables are normally distributed, with each distribution having a unique mean and standard deviation. The normal distribution plays an important role in statistical inference. Many statistical tests rely on the properties of the normal distribution, including the Student's t test and analysis of variance, described in previous questions.^{2 3} Such statistical tests are referred to as parametric tests.⁴ Some statistical procedures, including Pearson's correlation coefficient, described in a previous question,⁵ also depend on the properties of the normal distribution.

Competing interests: None declared.

- 1 Kendrick T, Dowrick C, McBride A, Howe A, Clarke P, Maisey S, et al. Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* 2009;338:b750.
- 2 Sedgwick P. Independent samples t test. *BMJ* 2010;340:c2673.
- 3 Sedgwick P. One way analysis of variance. *BMJ* 2012;344:e2427.
- 4 Sedgwick P. Parametric v non-parametric statistical tests. *BMJ* 2012;344:e1753.
- 5 Sedgwick P. Pearson's correlation coefficient. *BMJ* 2012;345:e4483.

Cite this as: *BMJ* 2012;345:e6533

© BMJ Publishing Group Ltd 2012