*BMJ* 2015;350:h831 doi: 10.1136/bmj.h831 (Published 17 February 2015) 

ENDGAMES

STATISTICAL QUESTION

# Standard deviation or the standard error of the mean

Philip Sedgwick *reader in medical statistics and medical education*

Institute for Medical and Biomedical Education, St George's, University of London, London, UK

The effects of a diet with a low glycaemic index during pregnancy on maternal and neonatal morbidity for women at risk of fetal macrosomia (large for gestational age infants) were investigated. A randomised controlled trial was performed. The intervention consisted of a low glycaemic index diet from early pregnancy. The control treatment was no dietary intervention. Participants were women without diabetes, all in their second pregnancy, who had previously delivered an infant weighing greater than 4000 g. In total, 800 women were recruited and randomised to the intervention (n=394) and control treatment (n=406) groups.[1]

The baseline characteristics for the treatment groups were presented; these included body mass index (BMI) (intervention: mean 26.8 (standard deviation 5.1); control 26.8 (4.8)). The outcome measures included birth weight and gestational weight gain. Of those women allocated to the intervention, 372 provided data at follow-up, compared with 387 of those allocated to the control. A per protocol analysis was performed. Mean birth weight was greater in the intervention group than in the control group, although the difference was not significant (mean 4034 (standard error 26.4) *v* 4006 (25.3) g; mean difference 28.6 g, 95% confidence interval −45.6 to 102.8; P=0.449). Mean gestational weight gain was significantly less for the intervention arm (12.2 (standard error 0.23) *v* 13.7 (0.25) kg; mean difference −1.35 kg, −2.45 to −0.24; P=0.01). The researchers concluded that a low glycaemic index diet in pregnancy did not significantly reduce birth weight for large for gestational age infants but it did have a significant effect on reducing gestational weight gain for women at risk of fetal macrosomia.

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

a) The standard deviation of the BMI quantified the variation in measurements at baseline for the sample members allocated to a treatment group

b) The standard error of the birth weight quantified the variation in measurements of birth weight in the population

c) At baseline, about 66% of sample members had a BMI that was within one standard deviation of the sample mean

d) If the sample size increased, the size of the standard error would be expected to decrease

## Answers

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

Standard deviation and standard error are often confused. The standard deviation is used to describe the variation in measurements of a variable for the members of a sample (*a* is true). The standard error describes the precision of the sample mean as an estimate of the population parameter—the population mean (*b* is false). The standard error, sometimes referred to as the standard error of the mean, is used, for example, to make inferences about population parameters using confidence intervals. One way of remembering when to use each of these measures is that standard deviation is for description and standard error is for estimation.

The aim of the trial was to ascertain the effects of a low glycaemic index diet during pregnancy on maternal and neonatal morbidity for women at risk of fetal macrosomia. A randomised controlled trial study design was used. The purpose of randomisation was to achieve groups similar in baseline characteristics, and thereby minimise confounding. To assess the success of the randomisation process, the researchers presented descriptive statistics for the baseline characteristics of the intervention and control groups. The treatment groups were compared using a visual inspection rather than statistical significance testing. Randomisation is expected to produce treatment groups with similar baseline characteristics, so statistical hypothesis testing is generally considered inappropriate because it has the potential for type I errors and may produce misleading results.[2 3] Presenting descriptive statistics for the baseline characteristics permitted readers to assess whether the results of the trial could be generalised to the patients in their clinical practice.

The baseline characteristics presented included BMI. The sample standard deviation of BMI quantified the variation in BMI—in particular, for each treatment group it provided a measure of how much on average the BMI of the sample members varied about the sample mean BMI at baseline (*a* is true). The derivation of the sample standard deviation has been described in a previous question.[4] The sample standard deviation of BMI at baseline may be used to calculate a series of ranges in BMI

p.sedgwick@sgul.ac.uk

containing approximate percentages of the sample members. Three ranges are often derived. For the intervention group, for example, about 68% of the sample had a BMI at baseline that was no further than one sample standard deviation away from the sample mean—that is, between (26.8–5.1; =21.7) and (26.8+5.1; =31.9). Furthermore, about 95% of the intervention group had a BMI at baseline that was no further than two sample standard deviations away from the sample mean—that is, between (26.8−2(5.1); =16.6) and (26.8+2(5.1); 37.0). Finally, about 99% of the intervention group had a BMI at baseline that was no further than three sample standard deviations away from the sample mean—that is, between (26.8−3(5.1); 11.5) and (26.8+3(5.1); 42.1).

The derivation of the three ranges described above is based on the properties of the theoretical normal distribution.[5] These ranges can be derived for any variable measured on a continuous scale. As long as the distribution of the variable in the sample is not too skewed, the series of ranges will generally provide useful approximations of the spread of measurements in the sample members. Typically only the ranges based on one and two sample standard deviations are considered. The proportion of sample members contained in each range is an approximation. For that reason, authors often state that the proportion of sample members with a measurement in the range based on two sample standard deviations is about two thirds—66%—rather than 68% (*c* is true). No doubt two thirds is easier to remember than 68%. The range based on two sample standard deviations is often used to derive so called normal ranges.[6] Sometimes the derived ranges can be used to ascertain whether the distribution of measurements for a variable is skewed. In particular, if the lower limit of a range is not permissible or is unlikely, this suggests that the distribution of measurements is skewed to the right (positively skewed).[7]

The outcomes for the above trial included birth weight. The mean birth weight of the sample was an estimate of the population parameter, and that for each treatment group estimated a different population parameter. The population parameter is the mean birth weight that would be seen if all mothers in the population from which the sample was taken received the intervention or control treatment. Although it was essential that the sample estimate of mean birth weight was similar in size to the population parameter, it was unlikely to have been exactly equal. Any inaccuracy in the sample estimate would be the result of it being based on a sample of mothers from the population—that is, it would be caused by sampling error. The accuracy of the sample mean birth weight as an estimate of the population parameter is quantified by the standard error of the mean. The standard error of the mean for a treatment group was derived by dividing the sample standard deviation of birth weight by the square root of the sample size of the treatment group. Therefore, in general, if sample size increased, the size of the standard error of the mean would be expected to decrease (*d* is true). This is intuitive, because as the sample size for a treatment group approaches that of the population, the sample mean will become closer in value to the population mean and therefore become a more accurate estimate of the population parameter.

The difference between treatment groups in mean birth weight was 28.6 g and was the sample estimate of the population parameter of the difference in mean birth weight. The standard error of the mean difference is derived in a similar way to that described above for the standard error of the sample mean. For each treatment group, the sample variance is divided by the sample size; the resulting values are then summed together and the square root of this value will equal the standard error of the mean difference. The standard error of the mean difference is used to derive the confidence interval for the population parameter of the mean difference in birth weight. The confidence interval is an interval estimate for the population parameter, and it quantifies the accuracy of the sample mean difference in birth weight as an estimate of the population parameter. A percentage is attached to the confidence interval, typically 95%. The 95% confidence interval for the population mean difference in birth weight was derived as the interval 1.96 standard errors either side of the sample mean difference in birth weight—that is, from (28.6−1.96(37.86); =−45.6 g) to (28.6+1.96(37.86); 102.8 g). It can be inferred that the derived confidence interval contains the population parameter with a probability of 0.95 (95%).

As described above the standard error can be derived for the sample mean plus the sample mean difference. The standard error can also be calculated for other sample estimates including proportions, the difference between two proportions, relative risks, and odds ratios. The standard error of each estimate is used in a similar way to that described above to derive a 95% confidence interval for the population parameter.

Competing interests: None declared.

1 Walsh JM, McGowan CA, Mahony R, Foley ME, McAuliffe FM. Low glycaemic index diet in pregnancy to prevent macrosomia (ROLO study): randomised control trial. *BMJ* 2012;345:e5605.
2 Sedgwick P. Randomised controlled trials: balance in baseline characteristics. *BMJ* 2014;349:g5721.
3 Sedgwick P. Pitfalls of statistical hypothesis testing: multiple testing. *BMJ* 2014;349:g5310.
4 Sedgwick P. Describing the spread of data I. *BMJ* 2010;340:c1116.
5 Sedgwick P. The normal distribution. *BMJ* 2012;345:e6533.
6 Sedgwick P. Normal ranges. *BMJ* 2013;346:f1343.
7 Sedgwick P. Skewed distributions. *BMJ* 2012;345:e7534.

Cite this as: *BMJ* 2015;350:h831