BMJ

BMJ 2012;345:e7767 doi: 10.1136/bmj.e7767 (Published 21 November 2012)

ENDGAMES

STATISTICAL QUESTION

What is a P value?

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 rapid rather than standard intravenous rehydration resulted in improved hydration and clinical outcomes when administered to children with gastroenteritis. Treatments were compared in a parallel randomised controlled trial. Children were recruited if aged 3 months to 11 years, had a diagnosis of dehydration secondary to gastroenteritis, had not responded to oral rehydration, and had been prescribed intravenous rehydration. Intervention was rapid (60 mL/kg) or standard (20 mL/kg) rehydration with 0.9% saline over an hour.[1]

The primary outcome was the proportion of children with clinical rehydration, assessed on a validated scale, within two hours of start of treatment. In total, 226 children were recruited, of whom 114 were randomised to rapid and 112 to standard rehydration. The proportion of children rehydrated at two hours was higher in the rapid rehydration group, although the difference was not significant (41/114 (36%) *v* 33/112 (29.5%); P=0.32).

Which one of the following statements best describes the P value?

a) It is the probability that the null hypothesis is true.

b) It is the probability that the alternative hypothesis is true.

c) It is the probability of obtaining the observed difference in the outcome measure, or a larger one, given that no difference exists between treatments in the population.

d) It is the probability that the observed difference in the outcome measure was due to random chance.

Answers

Statement *c* best describes the P value.

The trial investigated whether rapid rather than standard intravenous rehydration resulted in improved rehydration and clinical outcomes when administered to children with gastroenteritis. The statistical null hypothesis stated that in the population from which the sample was taken there was no difference between rapid and standard intravenous rehydration in the proportion of children with clinical rehydration within two hours of starting treatment. The alternative hypothesis was two sided: in the population rapid rehydration, when compared with standard intravenous rehydration, would result in either a larger or smaller proportion of children with clinical rehydration within two hours of starting treatment. The population was children aged 3 months to 11 years, who had a diagnosis of dehydration secondary to gastroenteritis, had not responded to oral rehydration, and had been prescribed intravenous rehydration.

The P value is the probability that the observed difference of 6.5% between treatments in the proportion of children with clinical rehydration within two hours of starting treatment—or a larger difference—would have been obtained if there were no difference between treatments in the population (statement *c*). The P value is a probability between zero and one (inclusive) and represents the strength of evidence provided by the sample data in support of the null hypothesis. A large P value indicates that the sample data support the null hypothesis, whereas a small P value indicates that they do not. The cut-off between a large and small P value is typically set at 0.05 (5%), termed the critical level of significance.

The χ squared test, described in a previous question,[2] could have been used to derive the P value for the test of the statistical hypotheses. The resulting P value was 0.32, which is larger than 0.05. Therefore, there was no evidence to reject the null hypothesis in favour of the alternative. The conclusion was that there was no evidence of a difference between treatments in the proportion of children with clinical rehydration within two hours of starting treatment.

Hypothesis testing and derivation of the P value are based on the theoretical situation of sampling an infinite number of times from the population. The above trial would be repeated infinitely with the same sample size and under the same conditions. As the critical level of significance was set at 0.05 (5%), the null hypothesis would be rejected in favour of the alternative in 5% of this infinite number of samples. In particular, this 5% of samples would be those with the largest difference between treatments in the main outcome measure, regardless of whether the proportion of children rehydrated at two hours is greater or smaller in the rapid than in the standard intravenous rehydration group—the sign of the difference is ignored. It is for these 5%

p.sedgwick@sgul.ac.uk

of samples that the null hypothesis is rejected in favour of the alternative and significance is inferred.

Even though the proportion of children showing clinical rehydration was greater in the rapid than in the standard intravenous rehydration group (36% versus 29.5%), the difference was not significant at the 5% level of significance. The sample difference between treatments was consistent with what would be expected with samples of the same size when taken from the population if there was no population difference between treatments in the proportion of children with clinical rehydration within two hours of starting treatment (that is, under the null hypothesis). Even if there were no difference between treatments in the population, we would expect differences between treatments in the trial sample, simply because of sampling error. Sampling error has been described in a previous question.[3]

A P value is often misinterpreted in a variety of ways, including as in statements *a* and *b*. The P value does not indicate the probability that the null hypothesis or alternative hypothesis is true or false. Instead the P value indicates whether the data support the null hypothesis or lend support to the alternative. This distinction is important, because in theory it would be difficult to prove that a hypothesis is true or false. A sample is one of a theoretical infinite number taken from a population and as such is prone to sampling error. Small samples are more likely to result in a type I or II error. Such errors have been described in a previous question.[4] A type I error occurs when the null hypothesis is rejected when it should not have been—that is, there is no difference between treatment groups in the population. A type II error occurs when the null hypothesis is not rejected when it should have been—that is, there is a difference between treatment groups in the population. Therefore, inferring that the null or alternative hypothesis is true or false on the basis of a single sample may be misleading.

The P value is a conditional probability, it being conditional on the null hypothesis: that there is no difference in the population between treatments in the proportion of children that are clinically rehydrated within two hours of starting treatment. Statement *d* is not a conditional statement and therefore does not best describe the P value. However, it is not obvious that statement *d* is confusing—in particular, the notion that the difference in treatments occurred by random chance. The words random and chance are closely related, so much so in ordinary English that they are near enough synonymous: to say that something occurred by chance is to imply it occurred at random. Hypothesis testing and derivation of the P value are based on the theoretical situation of sampling at random an infinite number of times from the population. Therefore, associating the P value with a random event would seem logical. However, the P value has nothing to do with the observed difference in the outcome measure between treatments occurring by chance or at random. The P value is equal to the proportion of the infinite number of samples that would give a difference between treatments in the proportion of children clinically rehydrated within two hours of starting treatment that was as large as or bigger than that observed.

Competing interests: None declared.

1 Freedman SB, Parkin PC, Willan AR, Schuh S. Rapid versus standard intravenous rehydration in paediatric gastroenteritis: pragmatic blinded randomised clinical trial. *BMJ* 2011;343:d6976.
2 Sedgwick P. Statistical tests for independent groups: categorical data. *BMJ* 2012;344:e344.
3 Sedgwick P. What is sampling error? *BMJ* 2012;344:e4285.
4 Sedgwick P. Errors when statistical hypothesis testing. *BMJ* 2010;340:c2348.

Cite this as: *BMJ* 2012;345:e7767