Kolloquium „Statistische Methoden in der empirischen Forschung“

Wann: 12. November 2019, 17:00 – 18:30 Uhr

Wo: Robert Koch-Institut | Nordufer 20 | 13353 Berlin (Wedding)

Cornelia Ursula Kunz (Boehringer Ingelheim, Biberach an der Riß)

P-values – A self-fulfilling prophecy?

Recent articles (for example, in Nature and The American Statistician) have reignited the debate around the meaning and interpretation of p-values. The controversy usually evolves around the interpretation of p-values, whether or not to use 5% as the default, or the difference between significance and relevance. Furthermore, arguments arise about the question whether or not small p-values provide evidence against the null hypothesis.

What is often overlooked is that p-values can only be calculated once an ordering of the sample space has been defined with different orderings leading to different p-values for the same outcome. Yet, very little attention is paid to this. Instead simple examples are being used, like the z-test for testing the mean against a fixed value, where only one sensible ordering of the sample space exists. For less simple examples, the problem of the ordering is often missed and instead the type I error rate is mistaken as the p-value, ignoring the fact, that these two quantities are not the same.

In this talk, we will show that p-values by their definition cannot provide evidence against the null per se. Using some commonly known test problems, we demonstrate that the evidence does not stem from the p-value being small but from the definition of the ordering. Furthermore, we illustrate how choosing a convenient ordering yields any arbitrarily small p-value for a given outcome. Hence, claiming that a small p-value provides evidence against the null hypothesis becomes a self-fulfilling prophecy.

Standard text books and introductory classes on statistics often introduce the idea of statistical testing using a normal or a t-distribution, both of which are unimodal, symmetric, and continuous distributions. The null hypothesis is usually stated in a two-sided way and the test decision is then based on either the test statistic being less than -1.96 or larger than +1.96 using a two-sided significance level of 5%. If statistical software packages like for example, SPSS, SAS, Stata or R are being used, it is then pointed out that the decision is the same as calculating the two-sided p-value and comparing it to the significance level alpha.

If one-sided testing is mentioned at all, then usually by stating that the one-sided p-value is half the two-sided p-value and the significance level is set to 2.5% (i.e. half of the significance level for the two-sided test). What is not discussed is that this only applies to certain test statistics and does not hold true in general.

In this talk, we will also go back to the roots of statistical testing by looking into the underlying definition of the p-value, the underlying ordering of the sample space and the implications these have on one- and two-sided testing. We will show that the results based on one- and two-sided test can be inconsistent with each other and sometimes even be illogical. Furthermore, we will show that doubling the one-sided p-value can substantially inflate the type I error. Hence, it is highly recommended not to derive one- or two-sided p-values from each other but to conduct the appropriate test instead.