**Universidade Federal de Alfenas**
**Programa de Pós-graduação em Estatística Aplicada e Biometria**
English Proficiency Exam

**Name:** _____

In this exam we present fragments of a scientific paper and ask questions regarding their interpretation. The paper is:
LEW, M. J. To P or not to P: on the evidential nature of P-values and their place in scientific inference. *arXiv*, November, 2013. Available on: `http://arxiv.org/abs/1311.0081v1`. Access: 10 Nov 2013.

Read the fragments carefully and answer the questions. Don't forget, you must answer only in Portuguese! Answers in English will not be considered.

**Fragment 1**:

## Abstract

The customary use of P-values in scientific research has been attacked as being ill-conceived, and the utility of P-values has been derided. This paper reviews common misconceptions about P-values and their alleged deficits as indices of experimental evidence and, using an empirical exploration of the properties of P-values, documents the intimate relationship between P-values and likelihood functions. It is shown that P-values quantify experimental evidence not by their numerical value, but through the likelihood functions that they index. Many arguments against the utility of P-values are refuted and the conclusion is drawn that P-values are useful indices of experimental evidence. The widespread use of P-values in scientific research is well justified by the actual properties of P-values, but those properties need to be more widely understood.

**Question 1**: What does this paper review about? _____

_____

_____

**Question 2**: P-values show to have intimate relationship with..._____

_____

_____

**Question 3**: According to the abstract, which conclusion is drawn about P-valeus?_____

_____

_____

**Fragment 2**:

## 1.1 What is a significance test and what is a P-value?

A significance test is not a hypothesis test [11]. That will will be self-evident to many readers, but not all. Consider the likely responses by non-statistically sophisticated users of statistics to this question: which of those two types of procedure is referred to by the common phrase 'null hypothesis significance test'?

A significance test yields a P-value whereas a hypothesis test yields a decision about acceptance of the null hypothesis or an alternative hypothesis. Frameworks exist that attempt to amalgamate significance and hypothesis tests [e.g. 14] or to append desirable inferential aspects of significance testing onto hypothesis testing [e.g. 15, 16] but those frameworks are controversial and have no been widely adopted. Nonetheless there is mixed approach in very widespread use. Unfortunately it is not an intentional mixture but an accidental hybrid that has been called a mishmash [17, 18], and it is a dysfunctional mishmash because the two approaches are incompatible [19, 11, 20]. The phrase 'null hypothesis significance test' should be avoided because it is confusing and, arguably, is itself a product of confusion.

An essential role for P-values is a core difference between significance tests and hypothesis tests. P-values are conventionally defined with reference to the null hypothesis.

**Question 4**: What is the main difference between a significance test and a hypothesis test?

**Question 5**: Which statement the author recommends to be avoided since it is confusing and arguable?

**Question 6**: What P-values are defined with reference to?

**Fragment 3**:

The other common style of definition specifies tail areas under sampling distributions, which amounts to the same thing. However, judging from the obvious confusion in many publications regarding the properties of P-values, neither style of definition serves well as an explanation. The introductory listing of alleged shortcomings of P-values may give the impression that confusion about P-values takes many forms but, while that may be true to a degree, one form of confusion leads more or less directly to the others. That primary confusion is that P-values measure error rates.

The idea that P-values measure type I error rates is as pervasive as it is erroneous, and it comes hand in hand with the significance test-hypothesis test hybrid. It might be seen as a natural extension or corollary of the P-value definition quoted above and, given that many introductory level textbooks actually introduce P-values within the hybrid framework, such a misunderstanding is itself understandable. However, even though deficiencies of textbooks in that regard have been noted many times [e.g. 18, 21, 5, 11] and sometimes analyzed in depth [19, 22, 23], textbooks are not entirely to blame. It can reasonably be said that R.A. Fisher himself was a contributor to the adoption of the hybrid approach. His writings are often difficult to fathom, his approach to argument was often to 'play the man rather than the ball', and even while promoting P-values as indices of evidence against the null hypotheses he advised:

> It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense they are prepared to ignore all results which fail to reach this standard [24 p. 13]

The dichotomization implied by that statement gives the impression that P-values fit into the error-decision framework of Neyman and Pearson. In same vein, Neyman and Pearson may also have contributed to the hybridization. They wrote:

> We may accept or reject a hypothesis with varying degrees of confidence; or we may decide to remain in doubt. [25 pp. 295-296]

**Question 7**: What the author calls a "primary confusion" about the P-values in this fragment?

**Question 8**: Who the author claims to be himself a contributor to the adoption an hybrid approach?

**Question 9**: What did Neyman and Pearson write that the author considered a contribution to the hybridization?

**Fragment 4**:

from this oft-quoted passage from Neyman and Pearson's original publication of their framework:

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.
>
> But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. [25 pp. 290-291]

**Question 10**: What may we search rather hoping to know whether each separate hypothesis is true of false?