

# PS1\_Answers

vkvats

## **Paragraph 1: Explain why the use of P-value-based “statistical significance” for decision- making has become controversial in recent years.**

When it comes to decision making, especially if the decision has to be binary in nature or establishing a cause-and-effect relation, then using only p-value-based statistical significance comes with lot of caveats. For p-values to be useful in a hypothesis testing, we need power in hypothesis testing which requires large number of samples ( to lower the chance of Type I and/or Type II errors). To Establish a cause and effect relation, we need to have a randomised controlled experimental setup ( with control group and experimental group) which nullify any biasness in the experiment. A biased sample might give the result we want but wont give the result that it should, then there comes the problem of choosing an  $\alpha$  value for making a binary decision, the conventional threshold value is set to 0.05 (the threshold value of  $\alpha$  is problem dependent, there is nothing conventional about it!) When it comes to make a binary decision and it varies from problem to problem on what could be the value of  $\alpha$  for a decision to be on the right side (which isn't really decided only by p-value alone). Taking a uniform value of  $\alpha$  will never be applied to all problems alike. Also, if we don't have an experimental setup then we can't really comment on a cause and effect relation based on any hypothesis testing or make a binary decision based on  $\alpha$  value. The biomedical and social sciences are facing a widespread crisis with published findings failing to replicate at an alarming rate (Blakeley et al, Abandone statistical significance, 2019). In recent years in a series of well-publicized examples (eg; Carney, Cuddy, and Yap 2010; Bem 2011) coupled with theoretical work has made it clear that statistical significance can easily be obtained from pure noise (Blakeley et al, 2019).

## **Paragraph 2 (the long one): Find ONE real-life study in which P-values are used. What is the study's argument, and are the P-values important to it? Do the P-values distract from more important things? Do the P-values show they're supposed to show? If not, why not?**

reference from paper: Carney et al, 2010. “Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance”

Study's argument: Humans and other animals express power through open, expansive postures, and they express powerlessness through closed, contractive postures.

Experimental setup: To achieve this task they used a group of 42 person (26 female, 16 male), who were made to pose in two different postures (expansiveness like taking up more space or less space and openness like keeping limbs open or closed) for one minute. The assignment of the persons to different groups were random ( but the selection of the experimental subjects were not random). After posing, the participants risk taking was measured with a gambling task and saliva samples were collected to measure the cortisol and testosterone levels. The feeling of “empowered” and “in-charge” was recorded on scale of 1 to 4 (self reported). Based on the three factors, risk taking ability, level of testosterone and feeling of empowered, the p-values were reported as.

Measurement output and conclusion in paper:

1. high-power poses caused an increase in testosterone compared with low-power poses, which caused a decrease in testosterone, p-value  $< .05$ .
2. high-power poses caused a decrease in cortisol compared with low-power poses, which caused an increase

in cortisol,  $p\text{-value} < .02$ .

3. high-power posers reported feeling significantly more “powerful” and “in charge” ( $M = 2.57$ ,  $SD = 0.81$ ) than low-power posers did ( $M = 1.83$ ,  $SD = 0.81$ ),  $F(1, 41) = 9.53$ ,  $p\text{-value} < .01$ .

Yes, P-values are important to this experiment, as it has been used as an evidence to concretize the point in favour of the study’s argument. While they used p-value to reinforce their study but they didn’t even state that threshold value  $\alpha$  that they used for it or how they decided to make some value as threshold for this study, it seemed more like showing some small p-values is enough to support the experiment. But the p-value distracts from the other important, or perhaps the more significant part of the experiment, that is how the experiment was set and and was this set-up sufficient to draw a cause-and-effect relationship between the pose and the concluded results? The important things that goes under the rug due to p-values are:

1. the experimental set-up: The experimental setup was not exactly randomized controlled experiment, which is very important if we want to establish a cause-and-effect relationship in any such experiment.
2. The people those were selected for the experiment, there is no information about how they were selected. The absence of information about the process of selection may introduce a bias in the experiment (since some outputs were self reported), which may or may not be eliminated or nullified afterwards.
3. Small sample size: a total of 42 persons, (26 female, 16 male). This is very small sample size, and gives no power in the hypothesis testing. There is more chance of committing type II error and fail to reject the hypothesis even when it is wrong.
4. Feeling of power was self-reported, this is highly subjective and some person may naturally feel empowered and some don’t. It will introduce a bias in the experiment.

In such a experimental setup, we can’t really trust on the p-values alone. the hypothesis testing has no power and consequently, the testing will not be sufficient to establish any kind of cause-and-effect relationship. To establish a cause-and-effect in this experiment, first, we need a proper randomized controlled experimental setup with control group and experiment group, and then a bigger sample of randomly selected people who are then randomly categorised into two groups. Every attempt should be made to remove human biasness from the experimental setup for it to be effective and conclusive.

**Paragraph 3: Make a conclusion, and justify it. If you don’t think P-values should be used, suggest what kinds of statistical analysis should be used instead.**

Though making statistical analysis using p-values is not the newest of the methods but it does have its significance. P-value is a probability measurement calculated with assumption that the null hypothesis is true. But except in a few special cases (randomized controlled experiments, surveys with 100% response rates), it’s not plausible that a null hypothesis model is even asymptotically true. In recent years, the problem has been in the way that p-value is being used misinterpreted. There is widespread understanding of conventional threshold value  $\alpha = 0.05$  which is being used in all studies and experiments without truly understanding the meaning and relevance of it with the study it is being used in. The value of  $\alpha$  depends on the problem in hand, it can never be one standard value accepted all across for all problems. Especially when we are supposed to make a binary decision (it’s always better not to do that) based on p-value, we should be more wary of its significance. Even if it is chosen to make a decision, it is better to supplement it with confidence-intervals. Having said this, it is quite important to establish a proper experimental setup (randomized controlled experiment) if we are planning to establish a cause-and-effect relationship in the experiment.

But it has been observed that it is very difficult to achieve an ideal condition for setting up an experiment, but then, we should always try to get close to it while being honest. But for more real world problems, where it is very difficult to achieve an ideal experimental set up, it makes more sense to do exploratory data analysis (EDA) than confirmatory data analysis. EDA is not hard-bund to make binary decisions based on some hypothesis, the flexibility is to develop a model which can better explain the data than to assume a model (hypothesis) and try to confirm the data with the model through significance testing. Though we are bound to face difficulties (overfitting, underfitting) in this approach as well but we can follow my professor’s advice

and can either be honest with the analysis or Bayesian.