P-value primer

The intent of this is to help everybody get a better feel for two variables: alpha (α) and the p-value. From a practical standpoint (at least from BCPS test standpoint) it would be easiest to consider these two variables as synonymous – the alpha (α) being sort of the p-value you are designing the study to achieve as you develop the study design. To simplify it, when you see alpha (α), think p-value.

Definition of p-value:

A measure of the probability (p) that the difference between two estimates could have occurred by chance, if the estimates being compared were really the same. The larger the p-value, the more likely the difference could have occurred by chance. The lower the P-value, the less likely that something occurred by chance. For example, for a p-value of <0.01 – there is less than one chance in one hundred (since 0.01→ 1%) that the results of the experiment were obtained by chance. This assumes good study design, random sampling, etc…but for the purposes of the BCPS test we will make those assumptions. The full range of the p-value is from zero to one. 0.05 means a 5% chance that the results are due to chance (5/100 or 1/20). A p-value of 0.5 means that there is a 50% chance that the results are due to chance (5/10). A p-value of 0.1 means that there is a 10% chance that the results were due to chance (1/10). Why is 5% (p=0.05) or less acceptable while 6% (p=0.06) or more is not?

History of the p-value:

From Fisher's writing in 1933[1]
(note: the quotes below were found in Lemuel Moye's book, Statistical Reasoning in Medicine)

"…the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 percent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials."

Note that this is rooted in what might have been expected to occur naturally under the influence of only random variation. Fisher goes on,

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point) or one in a hundred (the 1 percent point). Personally, the writer prefers to set the low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level."

The significance of 0.05 was born. There is no mathematical reason that 0.05 is the optimum type I error level – it simply became the accepted standard. Is a p-value of 0.06 essentially almost the same as 0.05? Yes. The scientific community spent decades arguing about what the cutoff should be and ultimately came to general agreement to accept Fisher's cutoff.

1. Fisher, RA The Arrangement of Field Experiments. Journal of the Ministry of Agriculture 1933; Sep 503-513.

Application of the p-value

Question: A new drug is being compared to an existing drug for the treatment of chronic myelogenous leukemia. Currently, the study is designed to detect a minimum 15% difference in response rates between the groups, if one exists, with a $p \leq 0.05$. If the study needed to detect a minimum 10% difference in response, which one of the following changes to study parameters would help ensure this?

A. Decrease the sample size.
B. Increase the sample size.
C. Select an $\alpha \leq$ of 0.01 as a cutoff for statistical significance.
D. Select an $\alpha \leq$ of 0.001 as a cutoff for statistical significance.

Sample size and alpha (α). **notice that I'll be using p-value and alpha interchangeably – for practical purposes they are interchangeable**

There is variation between individuals. We'll change the example above to make it more intuitive. If we know for a fact that a drug reduces the pulse of individual by an average of 1 bpm (or say 2%) and we want to show that in a study – how many people would it take? It depends. What level of significance are we trying to prove it at? Our designated cutoff (level of significance) in designing this is our alpha. Our starting point for alpha designation is always 5% or less. You could set it higher, but the scientific community would not accept the results if they were higher so you just would not do that. But, the lower you set it, the harder it is to prove. Why? Pulse (like anything) has natural variation between individuals. If we want to show a difference with this drug we certainly can't do it with 2 people (1 control, 1 experimental). We probably couldn't show it with 10 people. We could show it with one million people – then we would have the power to show this difference…even though pulse varies between individuals. In general, for any given alpha, the greater the sample size, the greater the chance of ensuring that you can show a difference. How about changing the alpha (the p-value). If we set the alpha at 0.8 we can show just about anything. Granted that if we reached our designed level of significance there would be an 80% chance that the results were due to chance…our study design is junk…but we could do it. The lower the alpha, the harder it is to show significance because you need to overcome the same amount of variability (assuming this is fixed for the variable and population in question). If you set the alpha at something like 0.000001 you would make it essentially unachievable that you would reach your designated level of significance. Why? Divide 1/1,000,000. This equals 0.000001. This means that if you ran the experiment 1,000,000 times you would only have an average of 1 occurrence where the result would be due to chance. If this does not make sense then read the Fisher quote from the previous page again. The lower the alpha, the more difficult it is to reach this designated level of significance.