STA 130: Group Exercise about Statistical Explanations

This exercise considers how best to *explain* statistical ideas. Form a group with the other students sharing your same group label. (For example, if you assigned tutorial on Blackboard is "WE76(c)", then your group label is "c".)

Working cooperatively with your group, discuss the following four explanations of the meaning of a P-value. What are the positive and negative aspects of each explanation? Which explanation do you think is best? How would you write your own explanation to make it better?

Explanation #1. (From Wikipedia[1].)

In statistics, the p-value is a function of the observed sample results (a statistic) that is used for testing a statistical hypothesis. More specifically, the p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that the null hypothesis is true. Here, "more extreme" is dependent on the way the hypothesis is tested. Before the test is performed, a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% and denoted as α.

If the p-value is less than or equal to the chosen significance level (α), the test suggests that the observed data are inconsistent with the null hypothesis, so the null hypothesis must be rejected. However, that does not prove that the tested hypothesis is true. When the p-value is calculated correctly, this test guarantees that the Type I error rate is at most α.

The p-value [...] does not in itself support reasoning about the probabilities of hypotheses but is only as a tool for deciding whether to reject the null hypothesis.

Statistical hypothesis tests making use of p-values are commonly used in many fields of science and social sciences, such as economics, psychology, biology, criminal justice and criminology, and sociology. Misuse of this tool continues to be the subject of criticism.

Explanation #2. (From dummies.com[2].)

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling

[1] https://en.wikipedia.org/wiki/P-value

[2] http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html

you about the population). The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that. You conduct a hypothesis test because you believe the null hypothesis, H_0, that the mean delivery time is 30 minutes max, is incorrect. Your alternative hypothesis (H_a) is that the mean time is greater than 30 minutes. You randomly sample some delivery times and run the data through the hypothesis test, and your p-value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the pizza place's claim that their delivery time is less than or equal to 30 minutes. Since typically we are willing to reject the null hypothesis when this probability is less than 0.05, you conclude that the pizza place is wrong; their delivery times are in fact more than 30 minutes on average, and you want to know what they're gonna do about it! (Of course, you could be wrong by having sampled an unusually high number of late pizza deliveries just by chance.)

Explanation #3. (From minitab.com[3].)

The P value is used all over statistics, from t-tests to regression analysis. Everyone knows that you use P values to determine statistical significance in a hypothesis test. In fact, P values often determine what studies get published and what projects get funding. [. . .]

P values evaluate how well the sample data support the [. . .] argument that the null hypothesis is true. It measures how compatible your data are with the null hypothesis. How likely is the effect observed in your sample data if the null hypothesis is true?

- High P values: your data are likely with a true null.
- Low P values: your data are unlikely with a true null.

A low P value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population.

In technical terms, a P value is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. For example, suppose that a vaccine study produced a P value of 0.04. This P value indicates that if the vaccine had no effect,

[3] http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values

you'd obtain the observed difference or more in 4% of studies due to random sampling error.

P values address only one question: how likely are your data, assuming a true null hypothesis? It does not measure support for the alternative hypothesis.

Explanation #4. (From a professor's note on his own web page[4].)

We wish to test a null hypothesis against an alternative hypothesis using a dataset. The two hypotheses specify two statistical models for the process that produced the data. The alternative hypothesis is what we expect to be true if the null hypothesis is false. We cannot prove that the alternative hypothesis is true but we may be able to demonstrate that the alternative is much more plausible than the null hypothesis given the data. This demonstration is usually expressed in terms of a probability (a P-value) quantifying the strength of the evidence against the null hypothesis in favor of the alternative.

We ask whether the data appear to be consistent with the null hypothesis or whether it is unlikely that we would obtain data of this kind if the null hypothesis were true, assuming that at least one of the two hypotheses is true. We address this question by calculating the value of a test statistic, i.e., a particular real-valued function of the data. To decide whether the value of the test statistic is consistent with the null hypothesis, we need to know what sampling variability to expect in our test statistic if the null hypothesis is true. In other words, we need to know the null distribution, the distribution of the test statistic when the null hypothesis is true. In many applications, the test statistic is defined so that its null distribution is a "named" distribution for which tables are widely accessible; e.g., the standard normal distribution, the Binomial distribution with n = 100 and p = 1/2, the t distribution with 4 degrees of freedom, the chi-square distribution with 23 degrees of freedom, the F distribution with 2 and 20 degrees of freedom.

Now, given the value of the test statistic (a number), and the null distribution of the test statistic (a theoretical distribution usually represented by a probability density), we want to see whether the test statistic is in the middle of the distribution (consistent with the null hypothesis) or out in a tail of the distribution (making the alternative hypothesis seem more plausible). Sometimes we will want to consider the right-hand tail, sometimes the left-hand tail, and sometimes both tails, depending on how the test statistic and alternative hypothesis are defined. Suppose that large positive values of the test statistic seem more plausible under the alternative hypothesis than under the null hypothesis. Then we want a measure of how far out our test statistic is in the right-hand tail of the null distribution. The P-value provides a measure of this distance. The P-value (in this situation) is the probability to the right of our test statistic calculated using the null distribution. The further out the test statistic is in the tail, the smaller the P-value, and the stronger the evidence against the null hypothesis in favor of the alternative.

---

[4] http://www.stat.ualberta.ca/ hooper/teaching/misc/Pvalue.pdf