One of the most misunderstood concepts in statistics is that of p values: what they are, how you get them, and what they mean. The purpose of this short handout is mainly to clarify what the first 2 items are (what they are and how you get them) and also hopefully clarify a bit the third (what they mean and DON’T mean).

GENERAL CONTEXT

Inferential statistics has as one of it’s cornerstones the process of formulating and testing various null hypotheses. For example, we formulate a null that says there is 0 difference between 2 treatments. We implement our experiment, collect data, analyze the data, then proceed to make some decision with respect to that null; that is let the analysis help us decide if in fact the treatments are different or not. Modern day statistics typically uses the p value from the analysis to help us make that decision. Let’s see how this process works.

The rest of this handout will present several different situations where nulls are formulated and tested, p values found, and then decisions made about the null. While I can’t cover all possible scenarios that APstat will likely encounter, I want to cover examples of the major types, which will include here: 1) a single proportion, 2) a difference in means, 3) the slope of a regression line, and finally 4) a test of independence with categorical data using chi square. I would encourage you after looking at these to find your own “pet” examples that you can put into your own context to help reinforce these ideas. Since the focus of this handout is on p values, please assume that whatever assumptions or conditions that are necessary to check for with these examples ... have been done and no red flags are waving.

SINGLE PROPORTION

A simple poll might be the most common example in this area. A large foods company wants to gather data about how well potential customers like one of their products ... KrispyPuffs Cereal. From their own anecdotal hunches, they have a feeling that about 70% of kids between 10 and 15 like KrispyPuffs, but want to verify that with some well done study. They hire a polling firm, who plans and conducts a polling study with 300 kids from both the city and suburbs in a large metro area having them try the cereal and then making a simple response of if they like the cereal or not. (NOTE: for sure, there are sampling issues here but ignore those for the moment). When all the data are collected, it is found that 65% of the kids say they like KrispyPuffs. Does this evidence suggest that the 70% value that the company had posited was too high or wrong?

To test any null, we formulate a model for the null case and then see where our sample data fits into that model. The basic idea is that IF the sample data are not too unusual in the null model case, then we are not able to reject the null (as usually we want to do). This in no way proves the
null, it just says that data are not different enough from the null model to reject it outright. So, what is the null model in this case where we have hypothesized that kids like KrispyPuffs at about the 70% rate?

With a SRS of about 300, we find that a normal approximation model for the null would be reasonable. We say the null is .7 and we need to find the appropriate error term like the SE proportion so as to show how far sample proportions might vary around the null of .7.

The SE proportion for a z test we will do and is \( SE \text{ proportion} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{.7 \cdot .3}{300}} = .03 \). So, for our null model of .7, we will build our approximately normal error model around .7 using the .03 value as our standard deviation. Since the normal model will go about 3 SE units on either side of .7, we would have a model like the following:

If the null had really been true under these conditions, then we would expect that SAMPLE proportions could range (in the extreme) from about .6 up to about .8. That means for example, that it would be very unlikely that we could see a sample proportion of .5 or .9 if this null model had been true (with \( n = 300 \) and this level of sampling error).

Now, given this, where do the SAMPLE DATA FIT INTO THIS NULL MODEL? Remember, we got a sample proportion of .65 and we see that .65 would be below the null value of .7. But, keep in mind that our data “could” help us to reject the null model IF we had an actual sample proportion that was equally distant from .7 BUT at the other end. That is a value of .75 would be
a sample value equally distant from .7, just in the other direction from .7. Thus, both .65 AND .75 are equally different from .7 and we need both of these to figure the p value. Here is the graph.

![Distribution Plot](image)

If we had obtained a sample proportion as far out as .65 (as we did) or .75 that would have been equally far from .7 would either of these values be different enough for us to reject the null? What would be the chances of getting a sample proportion as far out as .65 or .75 IF THE NULL HAD BEEN TRUE? Finding sample proportions at least as extreme as .65 on the low end or .75 at the high end would occur in this null model, in the $A + B$ of the total area of the null model. If we find the area in $A + B$, we have the probability of getting values as extreme as .65 or .75, in this null model. To find these areas, we just need to use our calculators or software, or a normal curve table, to find the area from .65 down AND .75 up. Since this model is symmetrical, if we find the A area and double it, we have the $A + B$ area.

What would the z value be for .65 ... or .75? Let’s do the .64 value and find the z.

$$Z (.65) = (.65 - .70)/.03 = -1.67.$$  

The area in a normal distribution from a $z = -1.67$ to the left would be about:.047. And, we would also see that this same area of .047 is in the B area too. So, the total area in the null model (assuming it is true) that is more extreme than .65 and down PLUS .75 and UP. is .047 + .047 = .094, just a bit below 10%.
The value in this particular situation of .094 ... is called the p value. The p value is nothing more ... nor less ... than the area at the extremes given the sample proportion we got of .65 (or would have applied if we had gotten .75) IF THE NULL HAD BEEN TRUE.

If the null had really been true in our scenario, we would have expected to see sample proportions (if we had taken other samples of n=300 of kids) of less than .65 or greater than .75 ... about a little less than 10% of the time.

Here is the result of doing this hypothesis test using Minitab. Minitab allows you to input the null value of .7 ... and the sample value of .65 ... along with the sample size of n = 300 ... and it can compute the confidence interval and the p value ... for your data. CAUTION: The value for p that Minitab gets is based on taking the proportion decimals much further out than what I rounded off to ( to make the calculations a bit simpler). For example, when I did the SE proportion, I got .03 ... ROUNDED TO 2 PLACES. But, actually, it is sqrt (.7 * .3) / 300 = .0265. Hence, the actual z value for our sample proportion of .65 would have been (.65 - .7)/.0265 = -1.89 ... not the rounded value of -1.67 I used. So, the actual z would have been a bit further out ... at both ends ... and the area from the A + B areas would have been a bit smaller than my rounded area of a little less than 10%. Here is the Minitab output for this problem, done more exactly.

Test of p = 0.7 vs p not = 0.7

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>N</th>
<th>Sample p</th>
<th>95% CI</th>
<th>Z-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195</td>
<td>300</td>
<td>0.650000</td>
<td>(0.596027, 0.703973)</td>
<td>-1.89</td>
<td>0.059</td>
</tr>
</tbody>
</table>

If you had isolated the z areas from -1.89 down and + 1.89 up you would have found a total area of just a bit less than 6%. The correct area or p value is actually closer to .06 than what I found doing a bit more rounding off.

YOU SHOULD ALWAYS LET THE TECHNOLOGY YOU ARE USING ... FIND THIS EXACT p VALUE ... RATHER THAN ROUNDING OFF YOUR OWN CALCULATIONS, WHICH AS YOU CAN SEE CAN INTRODUCE SOME ERROR INTO THIS VALUE.

So, what is this p value?

If the null had been true in this case, then we would have expected to see SAMPLE results, like we got of .65 or more extreme (for both ends of the null model) to have been about .06 or 6% of the time.

The probability of getting sample results like we did ... assuming that some null model IS TRUE, is the p value. It is nothing more than this ... nor nothing less than this. It is a conditional probability ... p for getting results we did IF THE NULL IS TRUE. It is very important that you keep in mind this conditional nature of the p value: the probability ONLY has meaning in the context of having a true null condition.
No matter how small or large this p value turns out to be, it is ALWAYS given the assumption that the null is true.

So, what do we make of this p value? Remember, the company had posited that the true percentage of kids who liked KrispyPuffs was 70% or .7 as a proportion. We set up a hypothetical null model with that being the case. Given our data of .65 and our p value of .06 (or about 6%) should we keep the null of .7 in abeyance ... ie, retain it ... or should we let the sample data of .65 say to us REJECT THE NULL. Is the probability of getting the result we got low enough that ... we are thinking that the null could not have been true in our case? .06 is a pretty small probability but ... is it small enough?

Most statistical guidelines will say that if that p value or probability is > .05 or 5% ... that we should NOT reject the null. So, in our case with a p value of close to .06 we probably would not reject the null. This in NO way says that the null is true nor proves that the null is true, it is just “suggesting” to us that what we got was not a rare enough event to say that the null could NOT have been true. So, what we do in this case is to hold on to the null until perhaps we have more definitive data that might suggest otherwise.

DIFFERENCE BETWEEN 2 MEANS

Another common inference situation is that of doing a simple experiment ... perhaps with 2 treatments. An educational researcher thinks that a new method of teaching sampling methods will improve the knowledge students have about these techniques. To test this out, 70 students are randomly assigned to the 2 methods, n=35 in each group ... and then at the end a 30 item MC criterion test is given. Method A is the new method and method B is the old method. Here is the data based on the results of the study.

Descriptive Statistics: New, Old

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>35</td>
<td>23.318</td>
<td>0.366</td>
<td>2.165</td>
</tr>
<tr>
<td>Old</td>
<td>35</td>
<td>21.669</td>
<td>0.428</td>
<td>2.531</td>
</tr>
</tbody>
</table>

As you can see, the new method group answered approximately 2 items more, on average, than the old method group. Is this sufficient evidence that we are willing to say that the new method is different AND better than the old method?

Again, we usually in this situation would formulate a null of no difference or 0 difference if the two methods produced the same result. We are hoping of course to be able to reject this null based on the experimental data. To do that, we need to form a null model and see where our results fall into that null model. If our results seem not unusual given the null case, we will retain the null. If the results look very odd or unusual under the null model, we reject the null of no effect for one of the two methods producing different results.
To form the null model, we will assume that the assumptions for doing a t test on the differences in means are satisfied. We will build a model with 0 in the center and complete that normal distribution looking model by going out about 3 SE differences in mean values on either side of 0. Since these are different samples, we do NOT use a pooled model for getting our SE estimate.

SE difference in means = sqrt ( VarA/nA + VarB/nB) = sqrt (4.187/35 + 6.406/35) = .55

So here is the null model given this level of estimated sampling error.
If the null were true, our expectations would be that we could get differences in the means of the two treatment groups that go from about -2 (B better than A by about 2 points on the test) to about +2 (A better than B by about 2 points on the test). If we got a difference say around -1 ... or +1.3 ... these would not be rare events in the null model. But, what did we get? The difference we got was about 1.6 points on the test. Where would that fit into our null model? Again, keep in mind that we could have gotten a difference of about -1.6 ... favoring the B or old method ... and be just as extreme on the left side as 1.6 is on the right side ... of 0.

What you see is that sample difference in means results ... either + (like we got) or - (which would have been the same but at the other extreme) are quite far out in the null model. What is the probability of getting sample results that extreme ... or even further out IF THE NULL HAD BEEN TRUE? Since we are using a t test for differences in means, we would have to find the t value for our test statistic and then see how much area there was at each end and add them together to get the total of the extremes, which will be our p value. Again, the A + B principle applies here ... we would have to let our technology or a t table help us find this area value. Here is the Minitab results doing a 2 sample t test on the difference in means.

The calculated t value for this problem would have been:

\[ t = \frac{\text{mean difference} - \text{null value}}{\text{SE difference}} \]

\[ t = \frac{1.649}{.55} = \text{about} \ 2.99 \]

NOTE: This value of 2.99 has a bit of roundoff error in it.

We would have to look up this t value, with the appropriate df value, and find the area BOTH below -2.99 and ABOVE + 2.99 ... add together to get the obtained p value. Here is how Minitab would do that a bit more accurately.

\[ \text{Difference} = \mu (\text{New}) - \mu (\text{Old}) \]
\[ \text{Estimate for difference:} \ 1.649 \]
\[ \text{95\% CI for difference:} \ (0.525, 2.774) \]
\[ \text{T-Test of difference} = 0 \ (\text{vs not}) =: \text{T-Value} = 2.93 \ \text{P-Value} = 0.005 \]

Note the calculated t of 2.93 ... which is a bit different than my rounded off value of 2.99. The Minitab value of 2.93 is the accurate one. Note also the p value of .005. This means if you get the area below a t of -2.93 and above +2.93 and add them together, you will have about .01 or 1% of the null model distribution.

For this problem there is only a little chance (approximately .01) in obtaining the results we did, or more extreme, IF THE NULL HAD BEEN TRUE.

So, with this probability of finding extreme sample results as we did WHEN THE NULL IS TRUE, what should we do? Should we hold on to the null, and not reject it ... OR since this p value is quite small, reject the null in favor of the alternative ... which says that the methods are not the same?
Most statistics guidelines would tell you to REJECT THE NULL in this case. Hence, we conclude ... BASED ON THE p VALUE BEING VERY LOW ... that the new method is better than the old method.

Does this prove that the null is false? NO. However, what we are saying is that given the results we got, it does not seem as though the null model of 0 difference is too plausible. Hence, we reject it.

Does this p value tell is how BIG the advantage if for the new method? NO. Does it tell us how important that effect is? NO.

All this rejecting of the null does is to say that we think the evidence is sufficient to allow us to think that the null is not true. It doesn’t prove that, it just provides some level of evidence against the null.

TEST OF SLOPE OF A REGRESSION LINE

A researcher in public health was interested if there was any relationship between the heights of college males and their normal pulse rates. At a state university, a SRS of n=57 males was taken and each Ss height and pulse rate were measured. Here is the scatterplot between the two variables.

As you can see, there does not appear to be much of a “connection” between the 2 factors. Would the pulse rate of students be predictive of how tall they were? We could do a simple linear regression and plot the regression line to see if the line were relatively flat or had a decided “slope” to it. Here it is.
While the regression line looks like it has a + slope (ie, reflects a + correlation between pulse rate and height), would that slope really be considered to be + ... if we did a significance test? That is, if we formulate a null that the slope = 0, would we be able to reject that null and conclude there is a + slope to this relationship? Again, just as a check, here is the correlation between the 2 variables.

Correlations: Pulse1, Height

Pearson correlation of Pulse1 and Height = 0.055

As you can see, while the r is +, it is very close to 0.

Here is the regression equation using pulse values to predict heights.

Regression Analysis: Height versus Pulse1

The regression equation is

Height = 69.8 + 0.0142 Pulse1

What we want to do is to test the null that the slope of the best fitting regression line = 0. Will the calculated slope of + .0142 in the regression equation, allow us to reject this null?

As we know with any statistic, from sample to sample it’s value can vary due to sampling error. That is, if we took another SRS of n=57 males from this population and measured their pulse rates and heights, and found the regression equation, that slope value of + .0142 will change: it could be somewhat higher or lower and since it is already close to 0, it could in fact be a negative value with the next sample. Before we can do the hypothesis test on the slope, we have to have some estimate of how much variation there could be in these slope values ... from random sample to random sample. That is, we need an estimate of the SE slope. While APStat does not typically work with this formula (it just uses computer output for it’s value), there is a formula:

\[ \text{SE slope} = \frac{(S \text{ residuals})}{\sqrt{(n - 1) \times S \times x}} \]
We need the SD of the errors or residuals and we need the SD of the actual X values, in this case, the SD of the pulse values. If we did the regression, we could find the residual values and get the SD of that and also get the SD of the pulse values, these are shown below.

SE slope = $2.579 / \sqrt{56} \times 9.95 = .035$

So, with our estimated sampling error for the slope, we can now test the null of 0 slope. Generally speaking, we can do this using a t test comparing the slope we got to 0, in terms of the SE slope. The calculated t is:

$t = (.0142 - 0) / .035 = .406$

We know that for using t tests, and either using 95% CIs or using alpha = .05 ... we need t values to be about 2 or greater to be able to reject the null. Let’s see where our t value of .406 fits into the overall null model. The null model we use in this case will be a t distribution with $n-2$ df or 55 degrees of freedom. Here is the null model with our calculated t value of .406 inked in. Keep in mind that since this null model is symmetrical, a t value of -.46 would be in the same position to the left as is +.46 is to the right.

As our usual process for finding the p value is, we need the extreme areas here added together ... that is A + B. If we use our technology or go to the t table, we will find that there is about .34 of the total area from -.406 to the left ... and of course there will be the same area of .34 to the right of +.406. So, adding A + B gives us an area of about .68.

The p value for this test of the null of 0 for our data is .68.
Of course, as I have said, we should let the technology do this test for us to get the most accurate p value. Here is the Minitab output for this regression line.

The regression equation is  
Height = 69.8 + 0.0142 Pulse1

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>69.752</td>
<td>2.486</td>
<td>28.06</td>
<td>0.000</td>
</tr>
<tr>
<td>Pulse1</td>
<td>0.01424</td>
<td>0.03496</td>
<td>0.41</td>
<td>0.685</td>
</tr>
</tbody>
</table>

On the Pulse1 line, you see that the slope is 0.01424 ... it's SE is about .035 (which we figured by hand before), the calculated t value is .41 ... and it’s associated p value is .685.

So, with a large p value of .685 ... what do we do with our null? Do we retain the null or reject the null? Again, most statistics sources would say that you do NOT reject the null unless the p value is about .05 OR LESS. Since our p value is so much larger than .05, WE RETAIN THE NULL.

Our conclusion based on our data is that we are not in a position to reject the notion that the slope for this regression line relating pulse rates to heights of males ... could be 0. In the future, we might have better and more definitive data on this matter but for the time being, we are not able to reject the null of 0 slope.

Does this mean that we have proved that the slope IS 0? Absolutely not. All we have done is to conclude that our data does not allow us to eliminate 0 as a possible true slope value in the larger population of male college students.

CHI SQUARE TEST OF INDEPENDENCE

The final example for finding p values is a typical one for categorical data using chi square as the analysis tool. For example, what if you are a polling company and have been asked to see if there is any relationship between whether voters favor (or not) a given policy issue AND the party that they belong to. You take an SRS of n=200 voters in the State of Pennsylvania and ask them if they favor or not the issue and you get the following results. You have sorted the data into Yes/No and party affiliation of Republican, Independent, or Democrat.

Data Display

<table>
<thead>
<tr>
<th>Row</th>
<th>R</th>
<th>I</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>34</td>
<td>26</td>
<td>60</td>
</tr>
<tr>
<td>No</td>
<td>51</td>
<td>24</td>
<td>15</td>
</tr>
</tbody>
</table>
If we just look at the table, we see that the Republicans seem to NOT favor the policy, whereas the Democrats do. What if we formulate a null hypothesis that says there is NO association or relationship between the dimensions of response and party affiliation: would we retain or reject that null?

This is the classic case where we could use chi square to examine this null. With a 3 by 2 table, we have 2 degrees of freedom. If the null of no association is true, then our data should follow a chi square distribution with 2 df ... below is a dotplot distribution using that model.

As you can see, this chi square distribution is VERY + skewed. While it is not obvious, the mean of chi square distributions will be the df value so in this case, the mean is about 2. If the sample data really fit the model of no association between the 2 dimensions, then our calculated chi square value should NOT be unusual within this model. Will it? Let’s complete the actual chi square calculations.

For our data, chi square = 26.74 .... Let’s see where that fits in to the null model for chi square with df = 2.

**Chi-Square Test: R, I, D**

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>I</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>26</td>
<td>60</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>48.57</td>
<td>28.57</td>
<td>42.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.371</td>
<td>0.231</td>
<td>6.857</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>24</td>
<td>15</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>36.43</td>
<td>21.43</td>
<td>32.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.829</td>
<td>0.309</td>
<td>9.143</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>85</td>
<td>50</td>
<td>75</td>
<td>210</td>
</tr>
</tbody>
</table>

**Chi-Sq = 26.740, DF = 2, P-Value = 0.000**
As you can see, our calculated chi square value of 26+ is WAY far out in this null model of chi square with df = 2. For chi square, we depart from our typical way of finding a p value. Usually, we find areas at the extremes at the left and right sides. That is because t values can be expected to be at either end. But, when doing the chi square calculation with these data, chi square will be expected to be larger and larger when there is more and more discrepancy between the observed and expected frequencies. So, what we do in the current chi square case is to ONLY find the extreme area at the upper end. That area beyond our calculated chi square above 26+ ... will be our p value. If we use our technology or a chi square table to find that area, we will see that it is very very small. In fact ... rounded off to 3 decimal places, that area or p value is 0. Actually, it is a bit larger than that but, not much. Shall we say .001 ... for good measure?

For the present case, the p value is less than .001. It is very small. Most statistics sources would say we would reject the null with \( p \) is .05 or less ... so clearly in this case, we will reject.

If our null is that there is no association between the two dimensions (response and party affiliation), we reject that given our sample data. From our data, it clearly looks like there IS an association. To interpret this, we have to look back at the data table where we see that the Republicans seem to disagree more often with the issue but the Democrats seem to agree more often, with the issue. Hence, party affiliation does seem to have a bearing on agreement or disagreement on the policy matter.

Does this prove that the null is false? NO
Does this prove that there IS an association? NO

What does this rejection of the null do for us then? Well, it suggests quite strongly that there is an association between the variables. It provides some evidence AGAINST THE NULL.

Does this rejecting tell us how STRONG the relationship is? NO. Does it tell us how important this relationship is? NO.
At most, this very small p value suggests that there is some non 0 association. Nothing more ... nothing less. So, if you are thinking that this rejection of the null tells us how big and important the association is, FORGET it!

SUMMARY OF p VALUES

Let’s first review the steps involved in finding a p value.

A. We formulate a null
B. Assuming the null is true, we display that null model
C. We superimpose in the null model, our sample data (usually a statistic like t or chi square)
D. We find the area(s) in the null model that are as extreme or more extreme than our sample data are
E. That extreme area ... is the p value.

The p value represents the area in the null model that is more extreme than our sample data given we assume the null is true. It’s that and nothing more than that.

A large p value means that our sample data do not look unusual given the null model. A very low p value suggests a misfit between our data and the null model. The smaller the p value is, the more discrepant our sample results look compared to what they should look like under the null model.

What does the p value prove? Nothing. If we retain the null with a large p value, it does NOT prove that the null is true. If we reject the null with a low p value it does NOT prove that the null is false. AT BEST, low p values suggest to you that the null model is not looking too good since the sample data conflict with it.

Does rejecting the null tell you how strong an effect is? NO
Does rejecting the null tell you how important an effect is? NO
So, given all of this, what DOES a low p value tell us? What DOES rejecting the null mean? Well, that’s a very good question. And, there is not any good, clear, and definitive answer to that question.

Does the p value and rejecting the null tell us something that is of value? That is highly debatable.

But, one thing is for sure, we make a lot more importance out of p values ... than the information that p values can give to us ... can provide.

My general view of how one interprets p value is this:

The lower is the p value, the more evidence we have saying that the null model is NOT looking too good. It does NOT say what is the case ... just that it does not look like the null is what is going on.

The p value does not tell you how big an effect is NOR does it tell you how important an effect is. The p value does NOT have the capacity to inform you of either of these results and of course, those 2 things are mainly what you want to know. If you want to know A) how big an effect is and/or B) how important an effect is, you will have to turn to something OTHER THAT p VALUES to do it.