TTest

December 20, 2016

1 T-Tests and P-Values

Let's say we're running an A/B test. We'll fabricate some data that randomly assigns order amounts from customers in sets A and B, with B being a little bit higher:

```
In [1]: import numpy as np
        from scipy import stats

        A = np.random.normal(25.0, 5.0, 10000)
        B = np.random.normal(26.0, 5.0, 10000)

        stats.ttest_ind(A, B)
```

```
Out[1]: Ttest_indResult(statistic=-14.075196812141339, pvalue=8.8277957363196977e-45)
```

The t-statistic is a measure of the difference between the two sets expressed in units of standard error. Put differently, it's the size of the difference relative to the variance in the data. A high t value means there's probably a real difference between the two sets; you have "significance". The P-value is a measure of the probability of an observation lying at extreme t-values; so a low p-value also implies "significance." If you're looking for a "statistically significant" result, you want to see a very low p-value and a high t-statistic (well, a high absolute value of the t-statistic more precisely). In the real world, statisticians seem to put more weight on the p-value result.

Let's change things up so both A and B are just random, generated under the same parameters. So there's no "real" difference between the two:

```
In [2]: B = np.random.normal(25.0, 5.0, 10000)

        stats.ttest_ind(A, B)
```

```
Out[2]: Ttest_indResult(statistic=0.088886198511817435, pvalue=0.92917324220169051)
```

Now, our t-statistic is much lower and our p-value is really high. This supports the null hypothesis - that there is no real difference in behavior between these two sets.

Does the sample size make a difference? Let's do the same thing - where the null hypothesis is accurate - but with 10X as many samples:

```
In [6]: A = np.random.normal(25.0, 5.0, 100000)
        B = np.random.normal(25.0, 5.0, 100000)

        stats.ttest_ind(A, B)
```

```
Out[6]: Ttest_indResult(statistic=0.20964627681745385, pvalue=0.83394397202032966)
```

Our p-value actually got a little lower, and the t-test a little larger, but still not enough to declare a real difference. So, you could have reached the right decision with just 10,000 samples instead of 100,000. Even a million samples doesn't help, so if we were to keep running this A/B test for years, you'd never acheive the result you're hoping for:

```
In [9]: A = np.random.normal(25.0, 5.0, 1000000)
        B = np.random.normal(25.0, 5.0, 1000000)

        stats.ttest_ind(A, B)
```

```
Out[9]: Ttest_indResult(statistic=-0.075342911693641518, pvalue=0.93994188742749496)
```

If we compare the same set to itself, by definition we get a t-statistic of 0 and p-value of 1:

```
In [10]: stats.ttest_ind(A, A)
```

```
Out[10]: Ttest_indResult(statistic=0.0, pvalue=1.0)
```

The threshold of significance on p-value is really just a judgment call. As everything is a matter of probabilities, you can never definitively say that an experiment's results are "significant". But you can use the t-test and p-value as a measure of signficance, and look at trends in these metrics as the experiment runs to see if there might be something real happening between the two.

1.1 Activity

Experiment with more different distributions for A and B, and see the effect it has on the t-test.

```
In [ ]:
```