The previous example demonstrates the old statistics saying:

[**Correlation Does Not Imply Causation**](https://xkcd.com/552/).

["Causality"](https://plato.stanford.edu/entries/causation-metaphysics/) is a vague, philosophical sounding word. In the current context, I am using it to mean "What is the effect on YY of changing XX?"

To be precise, XX and YY are [random variables](http://mathworld.wolfram.com/RandomVariable.html) and the "effect" we want to know is how the distribution of YY will change when we force XX to take a certain value. This act of forcing a variable to take a certain value is called an "Intervention".

In the previous example, when we make no intervention on the system, we have an observational distribution of YY, conditioned on the fact we observe XX:

P(Y|X)P(Y|X)

When we force people to wear cool hats, we are making an intervention. The distribution of YY is then given by the interventionaldistribution

P(Y|do(X))P(Y|do(X))

In general these two are not the same.

The question these notes will try and answer is how we can reason about the interventional distribution, when we only have access to observational data. This is a useful question because there are lots of situations where running an A/B test to directly measure the effects of an intervention is impractical, unfeasable or unethical. In these situations we still want to be able to say something about what the effect of an intervention is - to do this we need to make some assumptions about the data generating process we are investigating.

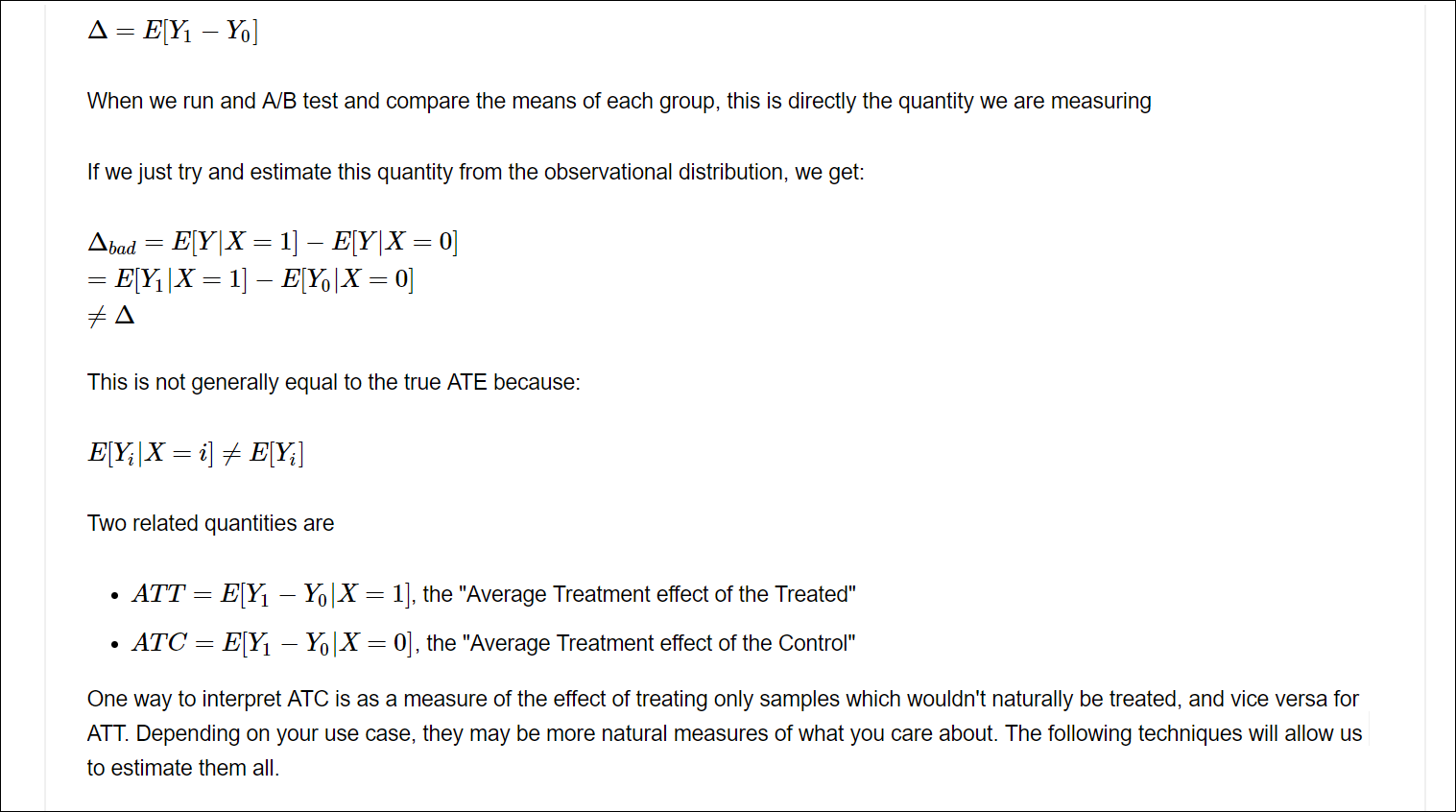
# Potential Outcomes

One way to approach this problem is to introduce two new random variables to our system: Y0Y0 and Y1Y1, known as the [Potential Outcomes](http://www.stat.unipg.it/stanghellini/rubinjasa2005.pdf). We imagine that these variables exist, and can be treated as any other random variable - the only difference is that they are never directly observed. YY is defined in terms of

* Y=Y1 when X=1
* Y=Y0 when X=0

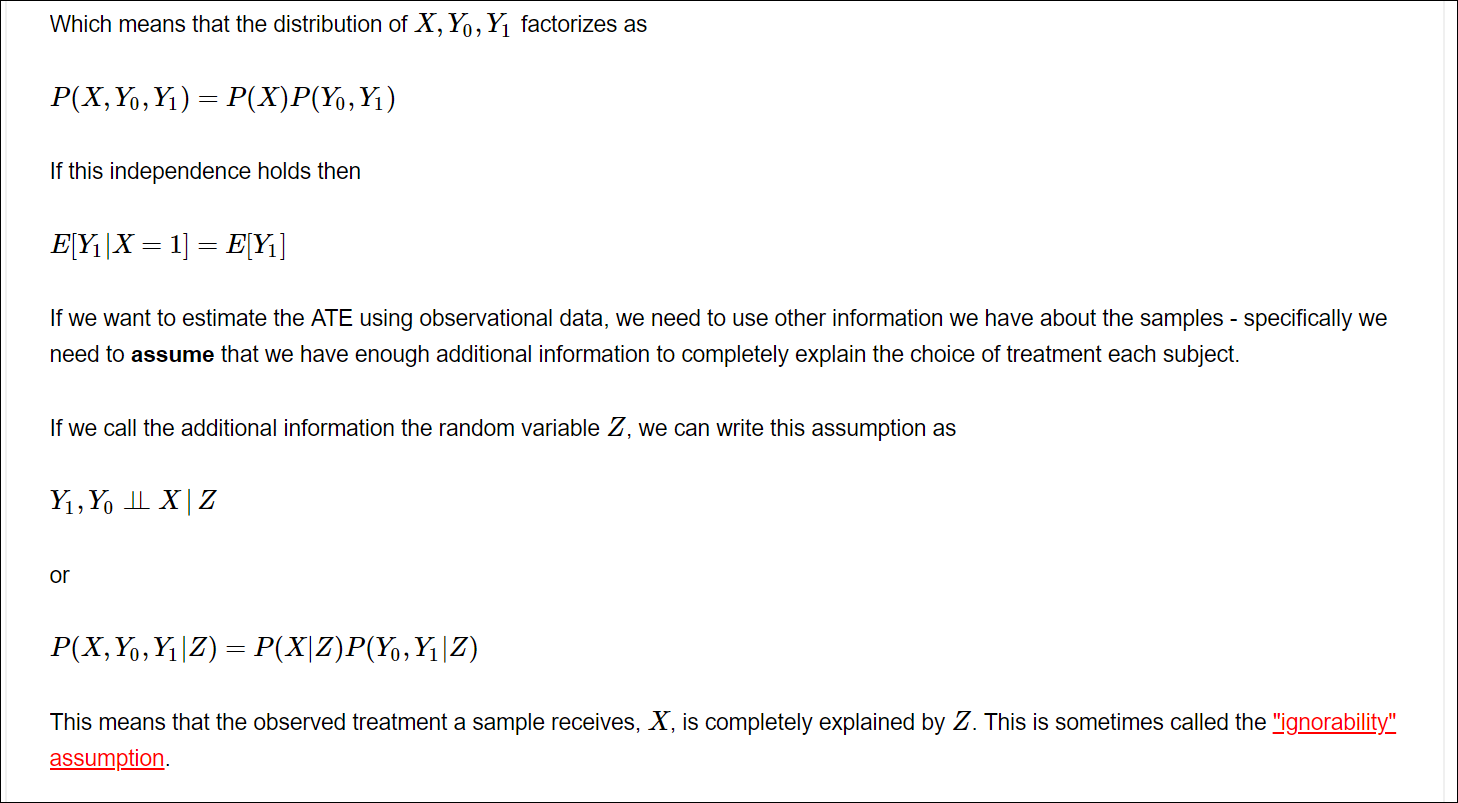
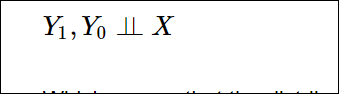
This shifts the problem from one about how distributions change under the intervention, to one about data drawn i.i.d. from some underlying distribution with [missing values](https://en.wikipedia.org/wiki/Missing_data). Under certain assumptions about why values are missing, there is well developed theory about how to estimate the missing values.

# Goals

Often we do not care about the full interventional distribution, P(Y|do(X))P(Y|do(X)), and it is enough to have an estimate of the difference in means between the two groups. This is a quantity known as the [Average Treatment Effect](https://en.wikipedia.org/wiki/Average_treatment_effect): 

# Making Assumptions

When we A/B test, we randomize the assignment of XX. This has the effect of allowing us to choose which variable of Y1Y1 or Y0Y0 is revealed to us. This makes the outcome independent of the value of XX. We write this as



In our motivating example about cool hats this would mean that there is some other factor - let's call it "skill" - which impacts both the productivity of the person and whether or not they wear a cool hat. In our example above, skilled people are more likely to be productive and also less likely to were cool hats. These facts together could explain why the effect of cool hats seemed to reverse when ran an A/B test.

If we split our data on whether or not the person is skilled, we find that for each subgroup there is a positive relationship between wearing cool hats and productivity:

In [6]:

observed\_data\_0\_with\_confounders = dg.generate\_dataset\_0(show\_z=**True**)

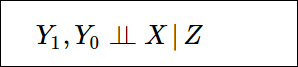
print(estimate\_uplift(observed\_data\_0\_with\_confounders.loc[**lambda** df: df.z == 0]))

print(estimate\_uplift(observed\_data\_0\_with\_confounders.loc[**lambda** df: df.z == 1]))

{'estimated\_effect': 0.21970895286301803, 'standard\_error': 0.16384029461180688}

{'estimated\_effect': 0.10869565217391297, 'standard\_error': 0.19384986576424423}

Unfortunately, because we never observe Y0Y0 and Y1Y1 for the same sample, we cannot test the assumption that



It is something we have to use our knownledge of the system we are investigating to evaluate.

The quality of any prediction you make depends on exactly how well this assumption holds. [Simpson's Paradox](http://www.degeneratestate.org/posts/2017/Oct/22/generating-examples-of-simpsons-paradox/) is an extreme example of the fact that if ZZ does not contain all confounding variables, then any inference we make could be wrong. [Facebook has a good paper comparing different causal inference approaches with direct A/B test that show how effects can be overestimated when conditional independence doesn't hold](https://www.kellogg.northwestern.edu/faculty/gordon_b/files/kellogg_fb_whitepaper.pdf).

Once we have made this assumption there are a number of techniques for approaching this. I will outline a few of simpler approaches in the rest of the post, but keep in mind that this is an area of ongoing research.

# LINK

<http://www.degeneratestate.org/posts/2018/Mar/24/causal-inference-with-python-part-1-potential-outcomes/>