# Sample Size

**Sample Size and Modelling**

**This presentation looks at how increasing the sample size increases the significance between two variables.**

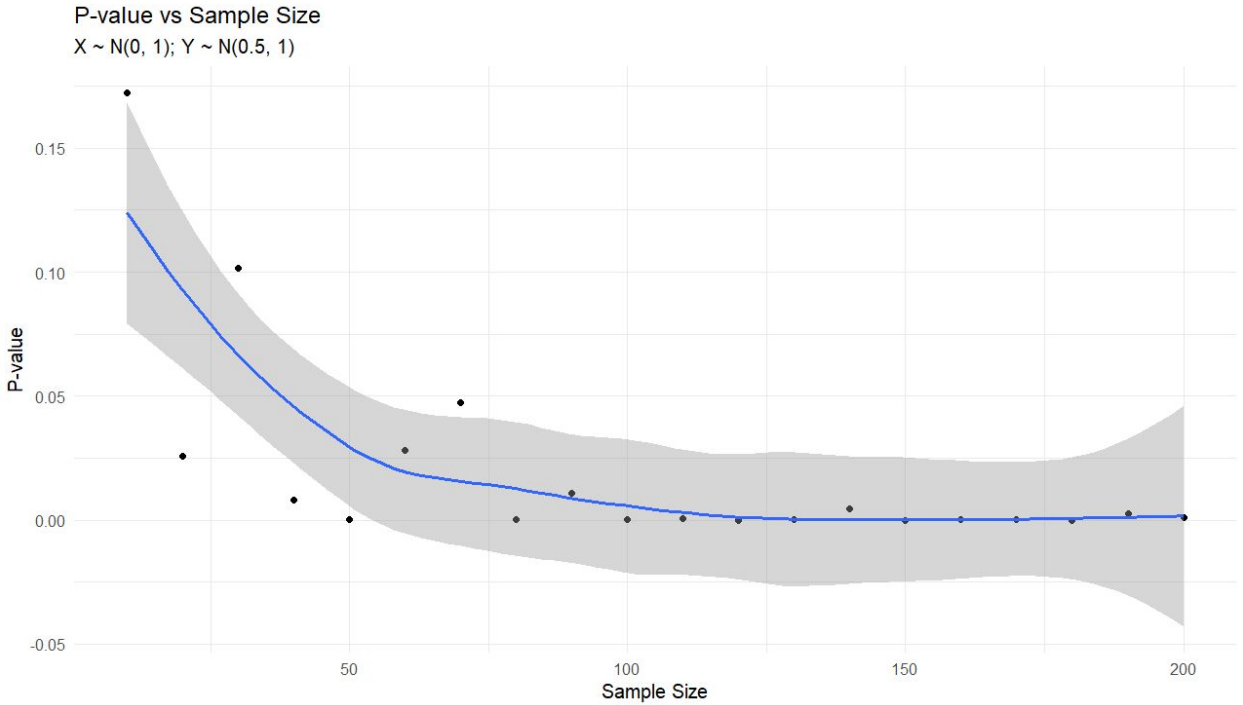
**We look at an example to demonstrate this.**

# Methodology

- Loop through increasing sample sizes from 10 to 200 in steps of 10.
- For each sample size, generate random data for two variables:
  - X is normally distributed with a mean of 0 and a standard deviation of 1
  - Y is normally distributed with a mean of 0.5 and a standard deviation of 1
- Then perform a t-test to test for a difference in means. This gives a p-value.
- Plot the Sample Size against the resulting P-Value
- The graph should show a trend of the p-value decreasing as the sample size increases.
- Keep in mind that the results will vary due to the randomness in data generation.

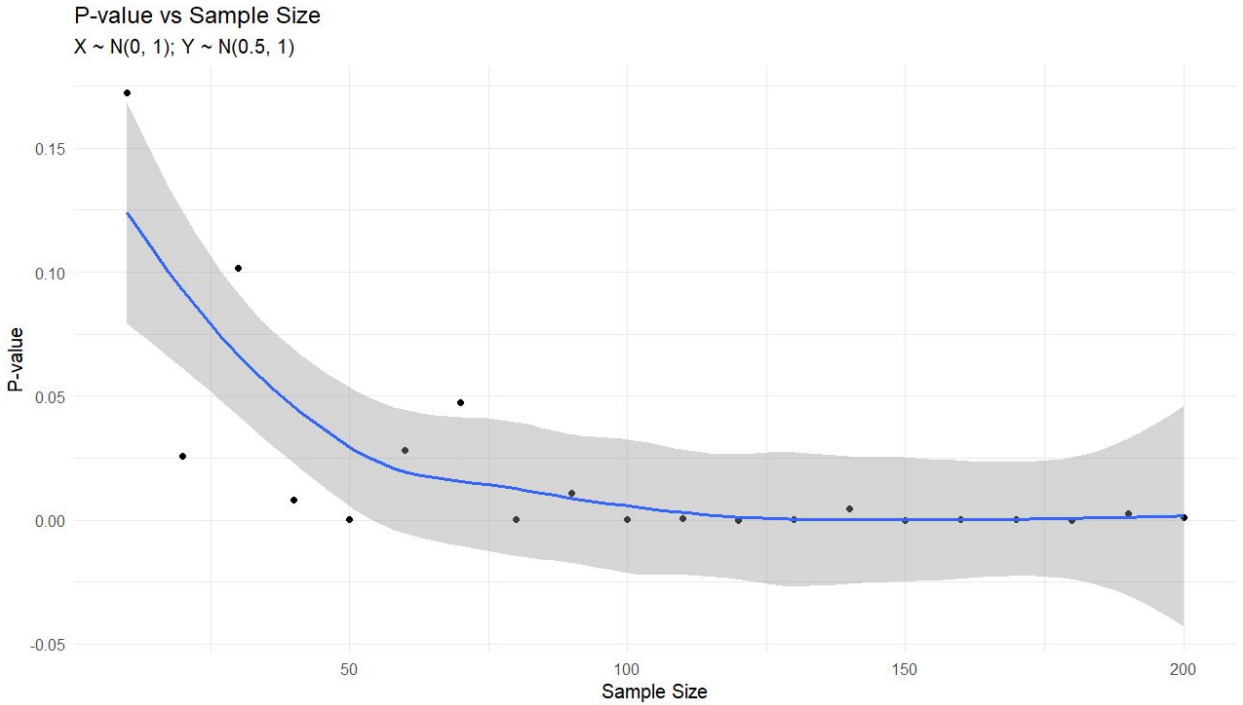
# Comparing Variables

- X: Normally Distributed
  - Mean = 0
  - SD = 1
- Y: Normally Distributed
  - Mean = 0.5
  - SD = 1



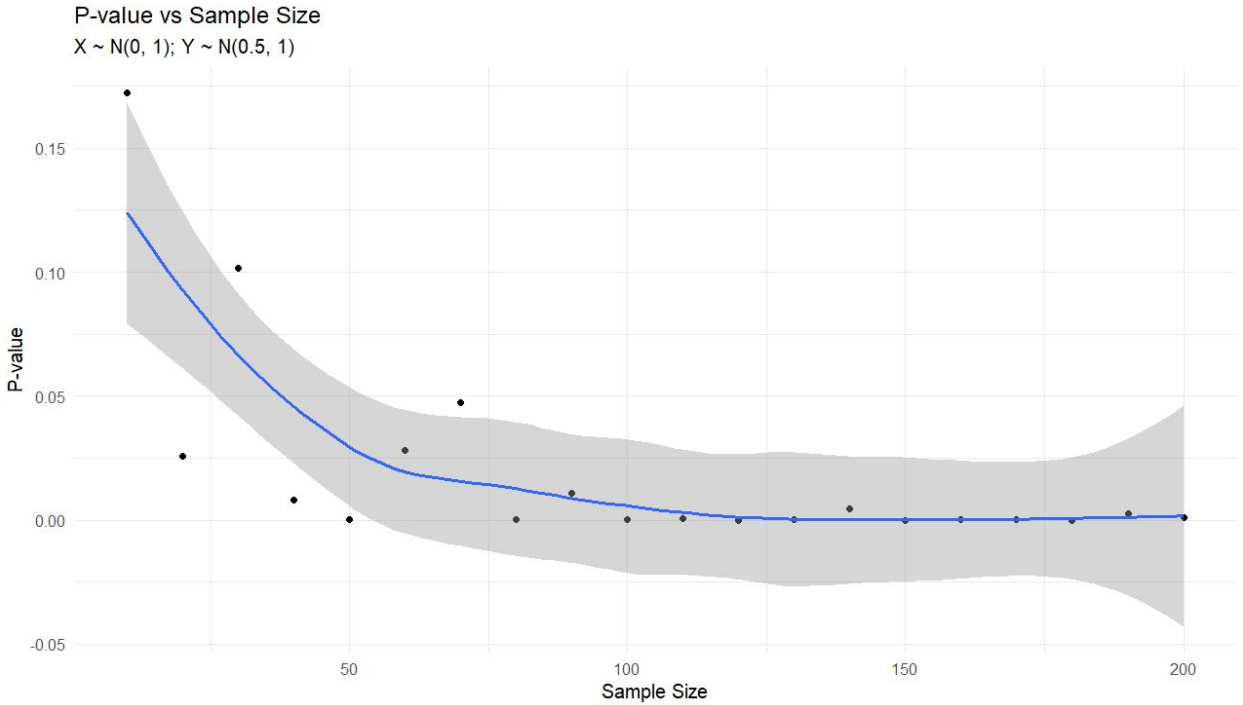
# Comparing Variables

- If the sample size is large enough then the p-value is significant.
- This does not mean that our two random values are any more related - just that the sample size in the model is large enough to give confidence that the observed differences are not merely due to chance.



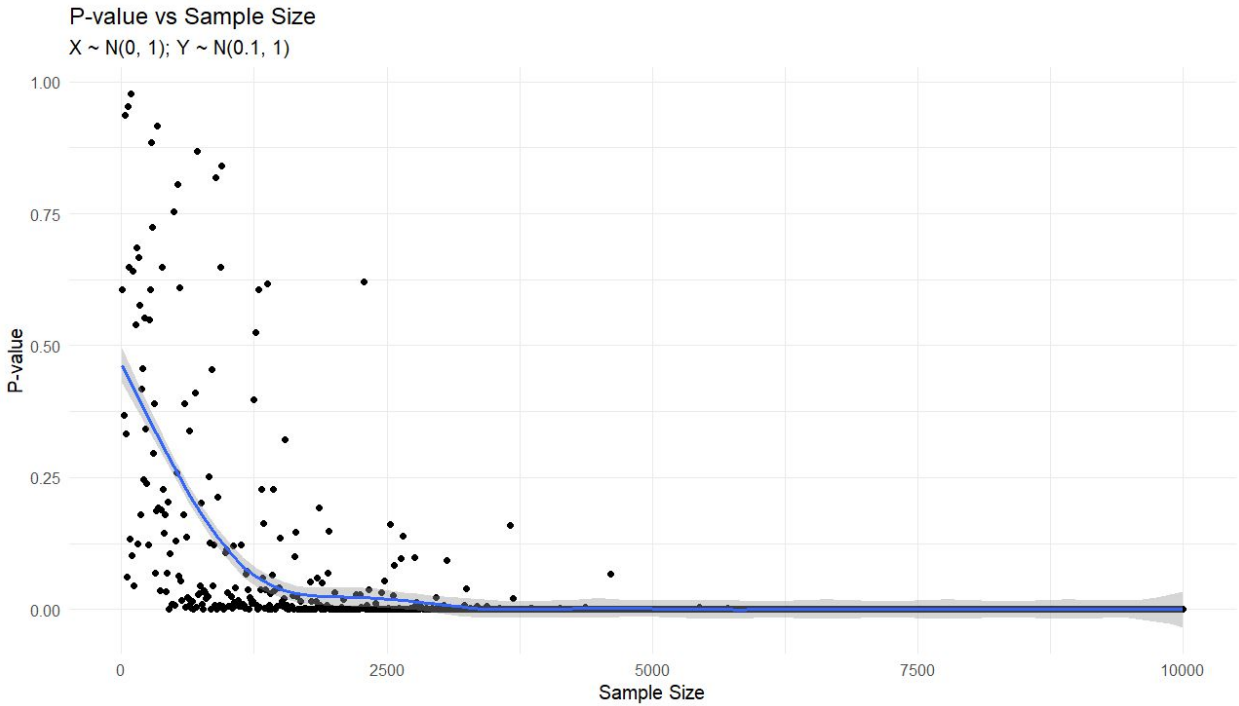
# Comparing Variables

- The threshold for a significant p is different for different scenarios.
- X: Normally Distributed
  - Mean = 0, SD = 1
- Y: Normally Distributed
  - Mean = 0.5, SD = 1
- What if we change the mean?



# Comparing Variables

- The threshold for a significant p is different for different scenarios.
- X: Normally Distributed
  - Mean = 0, SD = 1
- Y: Normally Distributed
  - **Mean = 0.1**, SD = 1
- As the means of X and Y get closer, we need more data. This is because we are testing for a difference in means.



# Comparing Variables

- The threshold for a significant p is different for different scenarios.
- X: Normally Distributed
  - Mean = 0, SD = 1
- Y: Normally Distributed
  - Mean = 0.5, **SD = 10**
- A larger SD means that it is harder to see a difference in means, so more data is needed to get a significant t-test.