# Power Analysis

Chris McClure-St. Amant

2023-02-18

**Scenario 1:** Data scientist in the Bay Area, assuming a normal distribution of salaries with approximate mean and standard deviation ($215,000 and $81000) from levels.fyi [1] and pay gap as uncontrolled gap from payscale.com (women making 90% of what men do in the tech industry) [2]. Also assuming normal distribution of salaries, which seems relatively reasonable based on the levels.fyi information. The sample size is 100. In this scenario, the power is 27.75%.

**Scenario 2** Since we plan to restrict the salary range in our study, standard deviation is likely to be smaller. Let's say we offer a range of $150,000 to $250,000 as the choices for our study participants, and the mean for the men is right in the middle at $200,000. Let's also assume that the standard deviation is $25,000. Retaining the uncontrolled difference from payscale.com used above. The sample size is 100. In this scenario, the power is 98.55%.

**Scenario 3** Same as scenario two except we'll assume the absolute (actually impossible) worst case for the standard deviation and make it $50,000, just to see what happens. The sample size is 100. Here, the power is 56.4%.

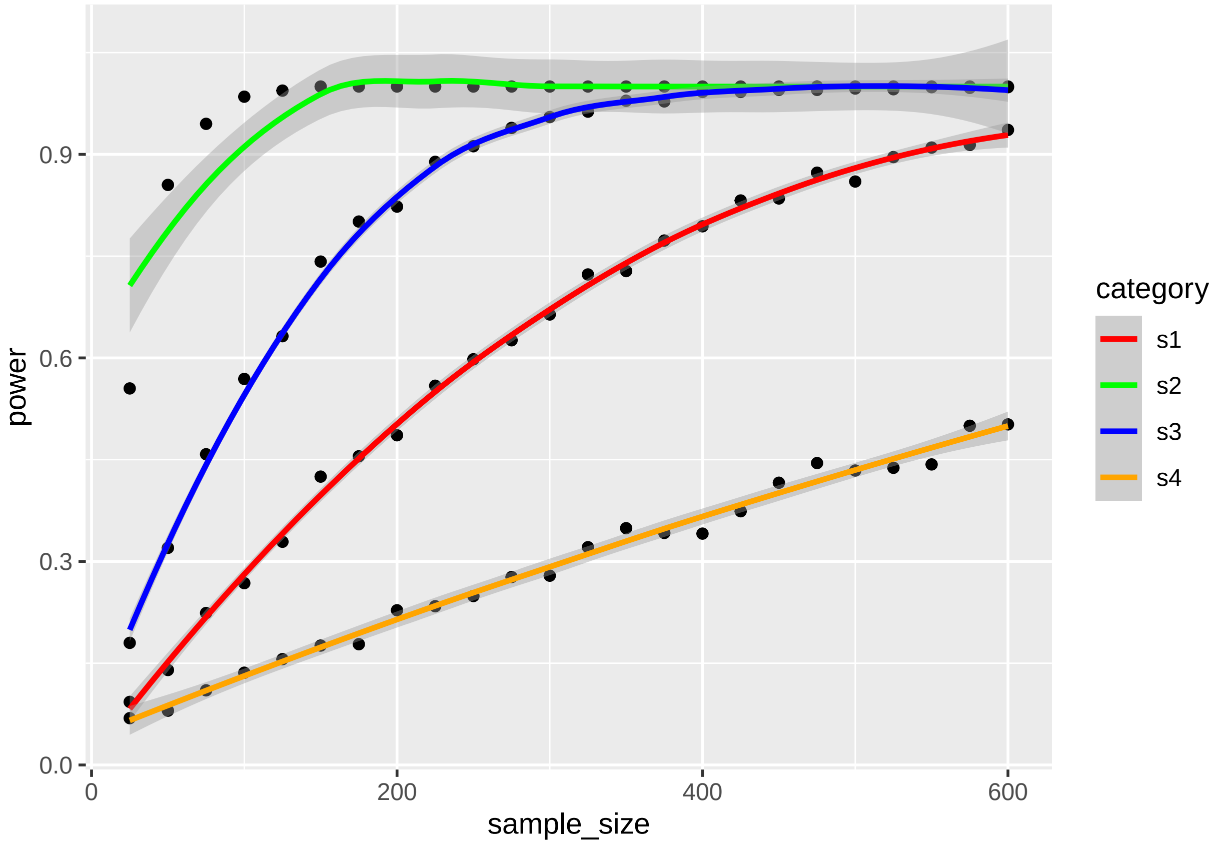
**Scenario 4** What if the treatment effect is much smaller, but we get the advantage of the restricted salary range and standard deviation from scenario 2 (mean $200,000, SD $25,000). Let's say instead of a 10% salary difference, there's a 2% difference. The sample size is 100. The power is 13.15%.

**Plot** Plotting the 4 scenarios at many different sample sizes, using a smooth line to show the trends.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

[1] https://www.levels.fyi/t/data-scientist/locations/san-francisco-bay-area

[2] https://www.payscale.com/research-and-insights/gender-pay-gap/#module-13



0.0
0.3
0.6
0.9
power
0
200
400
600
sample_size
category
s1
s2
s3
s4