

The importance of sample size & good study design

The most important aspect of a statistical analysis is not what you do with the data, it's what data you use

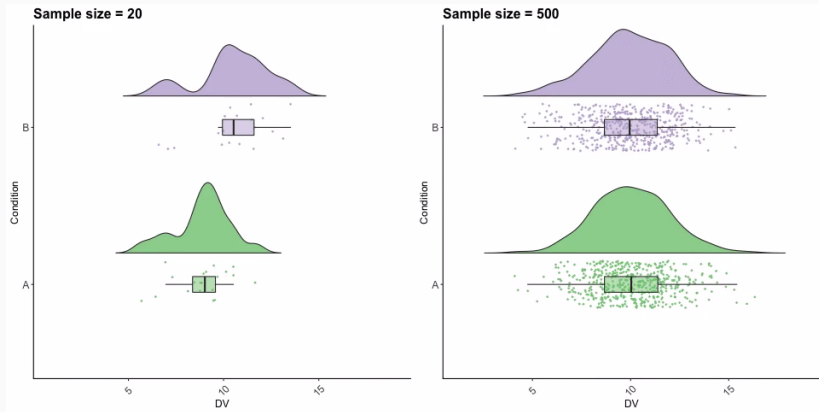
The importance of sample size

- Many studies have **too low sample sizes**.

The importance of sample size

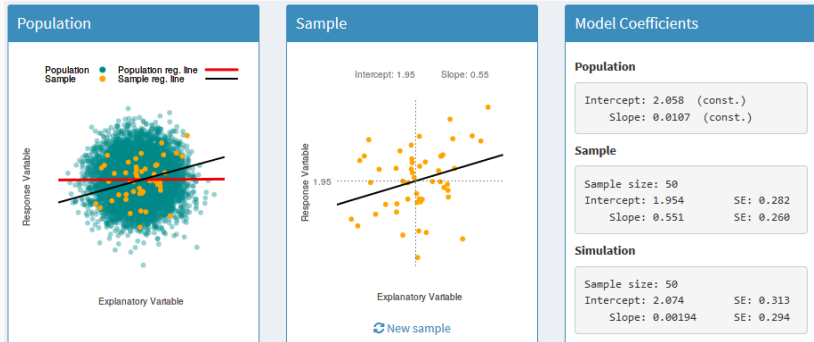
- Many studies have **too low sample sizes**.
- Low sample sizes miss subtle effects, but also **prone to bias**.

Low sample sizes very sensitive to random noise



https://twitter.com/ajstewart_lang/status/1020038488278945797

Low sample sizes may bias inferences about population



<http://statisticalgate.com/regression-simulation/>

Low sample sizes may bias inferences

See *The evolution of correlations*

Stopping rules

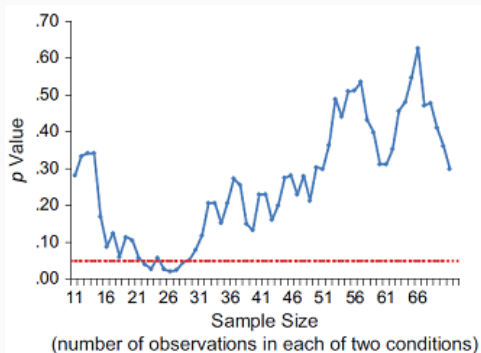


Fig. 2. Illustrative simulation of p values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a t test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

Sample size estimation

- Plan model/statistical analysis **before** data collection.

Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses (e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).

Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses (e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).
- Plan to have at least **10-30 observations per predictor**.

Sample size estimation

- Plan model/statistical analysis **before** data collection.
- **Do simulations.** Power/Sample size/Precision analyses (e.g. see papers like [this](#) & [this](#), or software like [this](#) & [this](#)).
- Plan to have at least **10-30 observations per predictor**.
- Complex models (w/ many predictors, interactions etc) require **high** sample sizes.

Sample size estimation

Calculating sample size for Gaussian (Normal) response model:

- expected mean: 30
- expected sd: 10
- 10 parameters (predictors)
- expected R-squared: 0.2

```
library(pmsampsize)
pmsampsize(type = "c", parameters = 10, intercept = 30, sd = 10, rsquared = 0.2)
```

NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared

NB: Assuming MMOE <= 1.1 in estimation of intercept & residual standard deviation

SPP - Subjects per Predictor Parameter

	Samp_size	Shrinkage	Parameter	Rsq	SPP
Criteria 1	313	0.900	10	0.2	31.3
Criteria 2	161	0.827	10	0.2	16.1
Criteria 3	244	0.876	10	0.2	24.4
Criteria 4*	313	0.900	10	0.2	31.3
Final	313	0.900	10	0.2	31.3

Minimum sample size required for new model development based on user inputs = 313

* 95% CI for intercept = (29.69, 30.31), for sample size n = 313

Sample size estimation

Calculating sample size for binary response model:

- expected prevalence: 0.1
- 20 parameters (predictors)
- expected R-squared: 0.2

```
library(pmsampsize)
pmsampsize(type = "b", parameters = 20, prevalence = 0.1, rsquared = 0.2)
```

NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared

NB: Assuming 0.05 margin of error in estimation of intercept

NB: Events per Predictor Parameter (EPP) assumes prevalence = 0.1

	Samp_size	Shrinkage	Parameter	Rsq	Max_Rsq	EPP
Criteria 1	796	0.900	20	0.2	0.48	3.98
Criteria 2	738	0.893	20	0.2	0.48	3.69
Criteria 3	139	0.900	20	0.2	0.48	0.70
Final	796	0.900	20	0.2	0.48	3.98

Minimum sample size required for new model development based on user inputs = 796,
with 80 events (assuming an outcome prevalence = 0.1) and an EPP = 3.98