DataLunch Statistical Power

Ian Dworkin

31 Jul 2025

points to think about before starting your power analysis

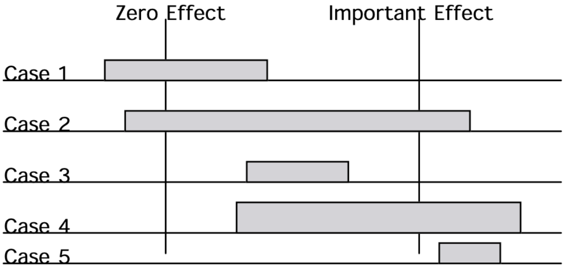


Figure 1. Confidence intervals for five different environmental scenarios.

Figure 1: Effects_Not_Pvalues

From Fox 2001 (DOI:10.1002/env.470)

points to think about before starting your power analysis

- Power analysis to maximize precision of your quantities of interest
- This paper by Daniël Lakens is a good starting place for considering how to determine sample sizes that will be sufficient for your needs.

Determining what effect sizes and variation to consider

- Small pilot studies are useful for many reasons.
- Using them to provide approximate parameter values for effect sizes and measures of among sample variation is most often not one of the reasons.
- Small studies will often have poorly estimated parameters of interest.
- Likely better to include information from published studies that give you a sense of among sample variation and plausible effect sizes.

Maximize precision of quantities of interest

- The standard error (sampling variation) is our usual measure of uncertainty in estimates.
- For a mean the s.e. is usually approximated as

$$se_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

With s_x being the standard deviation of your variable x.

So as a first approximation, you can re-write this:

$$n = (\frac{s_x}{se_{\bar{x}}})^2$$

Why is this useful

- I find this useful as I can think of how precise I want my estimate relative to the variation in the sample.

$$n = (\frac{s_x}{se_{\bar{x}}})^2$$

So if I want my estimate to be about as precise as a tenth of the variation

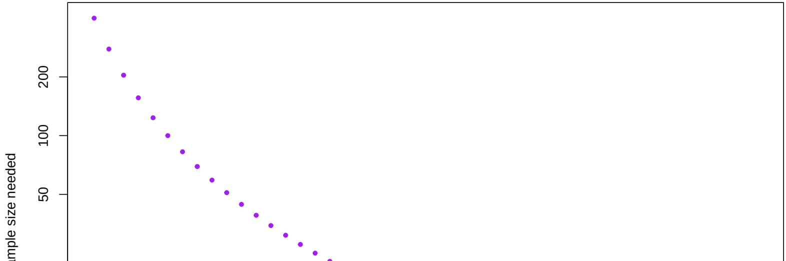
$$n = (\frac{s_x}{se_{\bar{x}}})^2 = (\frac{1}{0.1})^2 = 10^2 = 100$$

So I would need at least 100 samples to estimate the $\bar{x}$ to that level of precision.

helpful to visualize

```
precision_wanted <- seq(from = 0.05, to = 0.5, by = 0.01)
sample_size_needed <- (1/precision_wanted)^2

plot(y = sample_size_needed, x = precision_wanted,
     type = "p", pch = 20, col = "purple",
     log = "y",
     xlab = "precision desired",
     ylab = "sample size needed")
```



What if I had a minimal value I wanted to distinguish my estimate from

- Now we need to think about two pieces.
  - That our lower 95% (or whatever) confidence interval on our estimate does not overlap with this value.
  - The "power" we want to achieve.

The confidence intervals side of things

For sample sized above about 35, for the 95% SE you are looking at a value that is about $1.96se$

```r
qnorm(0.975) # two sided
```

```
## [1] 1.959964
```

```r
qnorm(0.95) # one sided
```

```
## [1] 1.644854
```

```r
qt(0.95, df = 50)
```

```
## [1] 1.675905
```

The power side of things

Say you want a power of about 0.8

```r
qnorm(0.8)
```

```
## [1] 0.8416212
```

We need to add these both

```r
multipler_we_need <- qnorm(0.8) + qnorm(0.975)

multipler_we_need
```

```
## [1] 2.801585
```

Now we include our minimal estimated effect and point of comparison

- We call this multiplier to account for precision of estimate and the power we wish to achieve m.
- Our point of comparison is θ_0 and our minimum estimated value of consideration is θ

$$n = (\frac{m \times s_x}{\theta - \theta_0})^2 = (\frac{2.8 s_x}{\theta - \theta_0})^2$$

How about if we are estimating the means of two groups

- If we can assume that the variation for each groups (A and B) is similar then it is

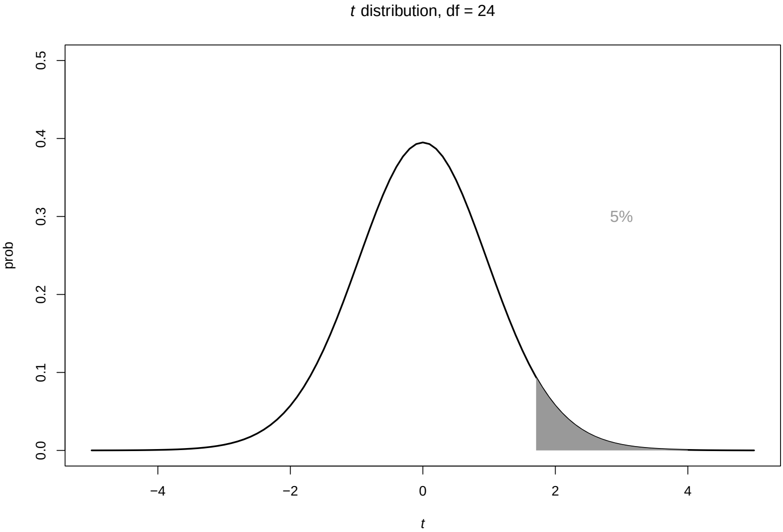
$$n = (\frac{2m \times s_x}{\theta_A - \theta_B})^2 = (\frac{5.6 s_x}{\theta_A - \theta_B})^2$$

the traditional "4 possible outcomes of a statistical test"

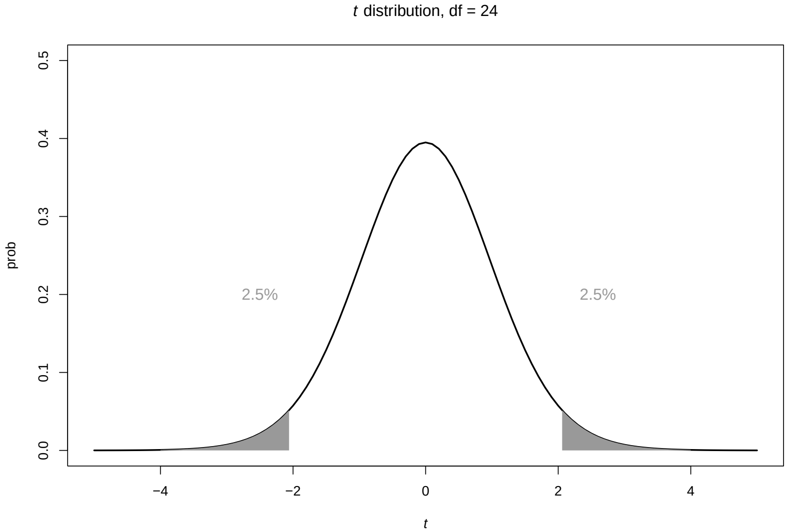
	Reject Null	Accept Null
Null True	Type I error, α	Correct, $1-\alpha$
Null False	Correct, $1-\beta$	Type II error, β

- $(1-\beta)$ is power, probability of detecting a true difference.
- $(1-\alpha)$ is confidence, probability of correctly accepting null.

Critical value for a t distribution, for a one tailed test



Critical value for a t distribution, for a two tailed test



Keep in mind

These kinds of dichotomies lead you to an *"Is there an effect?"* thinking.

Instead you should ask *"What is the effect?"* and for a power analysis, *"What precision of the effect do I want, given the resources I have?"*

What does a p-value tell you?

- Say you conduct an analysis on two different data sets, in the first $p = 0.05$, the second test has $p = 0.001$.

What does a p-value tell you?

- Say you conduct an analysis on two different data sets, in the first $p = 0.05$, the second test has $p = 0.001$.
- Does this mean the second test has a bigger effect? That the statistical model accounts for more variation (i.e. higher R^2)?

What does a p-value tell you?

- Say you conduct an analysis on two different data sets, in the first $p = 0.05$, the second test has $p = 0.001$.
- Does this mean the second test has a bigger effect? That the statistical model accounts for more variation (i.e. higher R^2)?
- Not necessarily. The magnitude of an effect could be similar, and the sample sizes differ (the second data set being much larger).

What does a p-value tell you?

- Say you conduct an analysis on two different data sets, in the first $p = 0.05$, the second test has $p = 0.001$.
- Does this mean the second test has a bigger effect? That the statistical model accounts for more variation (i.e. higher R^2)?
- Not necessarily. The magnitude of an effect could be similar, and the sample sizes differ (the second data set being much larger).
- It could also be that there is less variability in the second data set.

What does a p-value tell you?

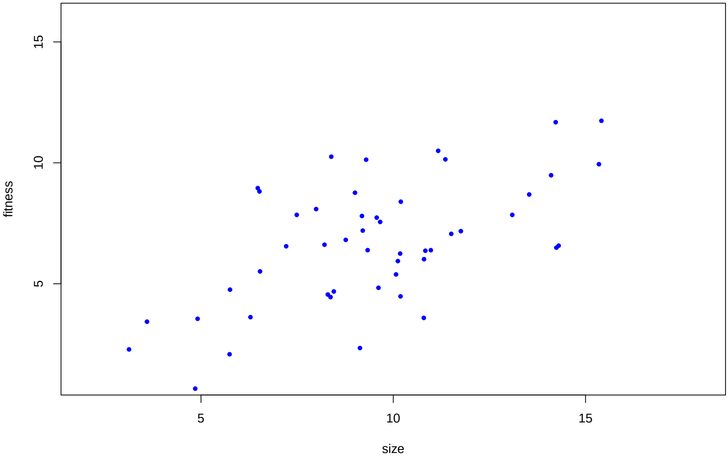
- Say you conduct an analysis on two different data sets, in the first p = 0.05, the second test has p = 0.001.
- Does this mean the second test has a bigger effect? That the statistical model accounts for more variation (i.e. higher R^2)?
- Not necessarily. The magnitude of an effect could be similar, and the sample sizes differ (the second data set being much larger).
- It could also be that there is less variability in the second data set.
- However it could also be that there is a difference in these. You need to examine (and report) all three whenever possible (include confidence intervals on estimates).

Let's compare these three data sets

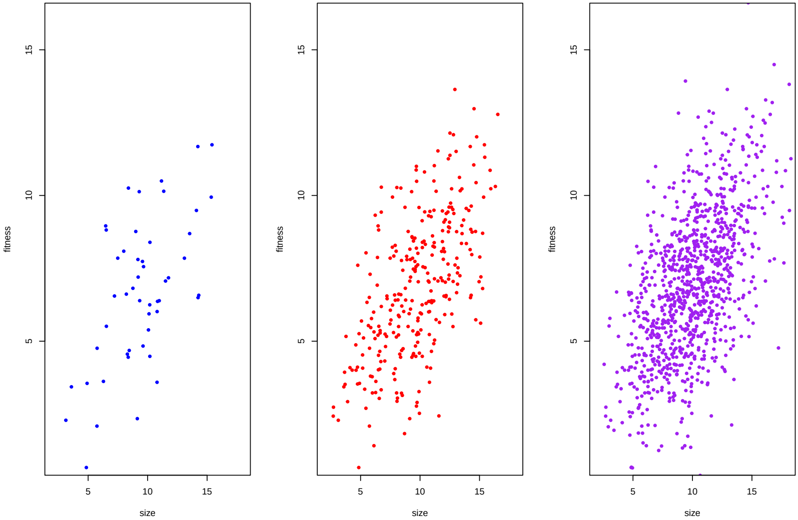
- We are examining the relationship between body size and fitness.

Let's compare these three data sets

- Is there a relationship?



Is there a relationship?



- In fact they all have the same relationship
  fitness $\sim N(2 + 0.5 * size, \sigma = 2)$, and only differ in sample

Statistical Power Analysis in R

- Most statistical software packages provide functions for simple power analyses

Statistical Power Analysis in `R`

- Most statistical software packages provide functions for simple power analyses
- In `R` there are many libraries one can use

Getting a critical value (for t distribution)

- Let's start by seeing how we get critical values for a t distribution

Getting a critical value (for t distribution)

- Let's start by seeing how we get critical values for a t distribution
- Assume we have a sample size of $n = 25$, $\alpha = 0.05$ for a two-tailed distribution.

Getting a critical value (for t distribution)

- Let's start by seeing how we get critical values for a t distribution
- Assume we have a sample size of $n = 25$, $\alpha = 0.05$ for a two-tailed distribution.
- We can use the `qt()` function for the t distribution

Getting a critical value (for t distribution)

- Let's start by seeing how we get critical values for a t distribution
- Assume we have a sample size of $n = 25$, $\alpha = 0.05$ for a two-tailed distribution.
- We can use the `qt()` function for the t distribution

```
qt(p = 0.975, df = 24)
```

```
## [1] 2.063899
```

- Why do we have df $= 24$, not 25?

Getting a critical value (for t distribution)

- Let's start by seeing how we get critical values for a t distribution
- Assume we have a sample size of $n = 25$, $\alpha = 0.05$ for a two-tailed distribution.
- We can use the `qt()` function for the t distribution

```
qt(p = 0.975, df = 24)
```

```
## [1] 2.063899
```

- Why do we have $\text{df} = 24$, not 25?
- Why is $\text{p} = 0.975$, not 0.95 (with $\alpha = 0.05$)?

How does the critical value change with sample size?

- We can make a plot looking at this across a range of sample sizes.

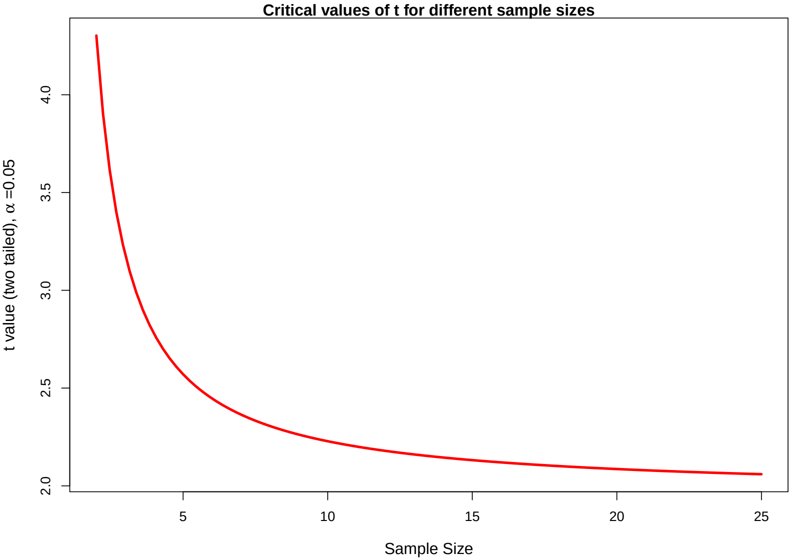
How does the critical value change with sample size?

- We can make a plot looking at this across a range of sample sizes.

```
curve(qt(p = 0.975,df = x), 2, 25,
    col = "red", lwd = 3, cex.lab = 2,
    main = "Critical values of t for different sample sizes
    xlab = "Sample Size",
    ylab = expression(paste("t value (two tailed), ", alpha
```

How does the critical value change with sample size?

- We can make a plot looking at this across a range of sample sizes.



Critical values for other distributions

- There are other distributions we can use:

Critical values for other distributions

- There are other distributions we can use:
- `qf()` for the F distribution, `qchisq()` for χ^2 etc..

tools in `R`

- Many libraries in `R` to use, depending on purposes.

tools in R

- Many libraries in R to use, depending on purposes.
- in base R, there is `power.t.test()`, `power.anova.test()`, `power.prop.test()`

tools in `R`

- Many libraries in `R` to use, depending on purposes.
- in base `R`, there is `power.t.test()`, `power.anova.test()`, `power.prop.test()`
- `pwr` is an `R` package that does many simple types of statistical models (proportions, 1-way ANOVA, regression, χ^2, glm)

tools in R

- Many libraries in R to use, depending on purposes.
- in base R, there is `power.t.test()`, `power.anova.test()`, `power.prop.test()`
- `pwr` is an R package that does many simple types of statistical models (proportions, 1-way ANOVA, regression, χ^2, glm)
- See the CRAN TASK VIEWS on experimental design for a list (and description) of more.

tools in `R`

- Many libraries in `R` to use, depending on purposes.
- in base `R`, there is `power.t.test()`, `power.anova.test()`, `power.prop.test()`
- `pwr` is an `R` package that does many simple types of statistical models (proportions, 1-way ANOVA, regression, χ^2, glm)
- See the CRAN TASK VIEWS on experimental design for a list (and description) of more.
- See this draft task view for power

tools in `R`

- Many libraries in `R` to use, depending on purposes.
- in base `R`, there is `power.t.test()`, `power.anova.test()`, `power.prop.test()`
- `pwr` is an `R` package that does many simple types of statistical models (proportions, 1-way ANOVA, regression, χ^2, glm)
- See the CRAN TASK VIEWS on experimental design for a list (and description) of more.
- See this draft task view for power
- I will show just a couple here.

Some of the functions in base R

```
## [1] "power"              "power.anova.test" "power.prop.te
```

power.t.test

- What goes into a t-test?

power.t.test

- What goes into a t-test?
- $$\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}\frac{1}{\sqrt{n}}}$$

`power.t.test`

- What goes into a t-test?
- $$\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}\frac{1}{\sqrt{n}}}$$
- $\bar{x}_A$ is the mean for group A, $\bar{x}_B$ for B

`power.t.test`

- What goes into a t-test?
- $$\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}\frac{1}{\sqrt{n}}}$$
- $\bar{x}_A$ is the mean for group A, $\bar{x}_B$ for B
- The denominator is just the *pooled standard error of the mean*

`power.t.test`

- What goes into a t-test?
- $$\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}\frac{1}{\sqrt{n}}}$$
- $\bar{x}_A$ is the mean for group A, $\bar{x}_B$ for B
- The denominator is just the *pooled standard error of the mean*
- So we see that there are 4 critical things:

power.t.test

- What goes into a t-test?
- $$\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}\frac{1}{\sqrt{n}}}$$
- $\bar{x}_A$ is the mean for group A, $\bar{x}_B$ for B
- The denominator is just the *pooled standard error of the mean*
- So we see that there are 4 critical things:
- α, the difference between means $\Delta = \bar{x}_A - \bar{x}_B$, n and $\hat{\sigma}$

power.t.test

```r
pwr_t_check <- power.t.test(delta = 0.5, sd = 2,
                            sig.level = 0.05, power = 0.8)

pwr_t_check
```

```
##
##      Two-sample t test power calculation
##
##               n = 252.1281
##           delta = 0.5
##              sd = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
str(pwr_t_check)
```

`power.t.test`

- what sample sizes we would need for a range of differences, Δ, on the interval $[0.1, 0.5]$.

`power.t.test`

- what sample sizes we would need for a range of differences, Δ, on the interval $[0.1, 0.5]$.
- $(1 - \beta) = 0.8$, $\hat{\sigma} = 2$, $\alpha = 0.05$

power.t.test

- $\Delta = 0.5$, $\hat{\sigma} = 2$, $\alpha = 0.05$

```
delta_vals = seq(from = 0.1, to = 0.5, by = 0.01)
delta_vals
```

```
##  [1] 0.10 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19
## [16] 0.25 0.26 0.27 0.28 0.29 0.30 0.31 0.32 0.33 0.34
## [31] 0.40 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49
```

- This creates a vector from 0.1 - 0.5

power.t.test

```
pow.test <- function(x){

  pow2 <- power.t.test(delta = x, sd = 2,
                          sig.level = 0.05, power = 0.8) # We

    return(pow2$n) # This pulls out the sample size we need
    }
```

power.t.test

```
power.n <- sapply(delta_vals, pow.test)
```

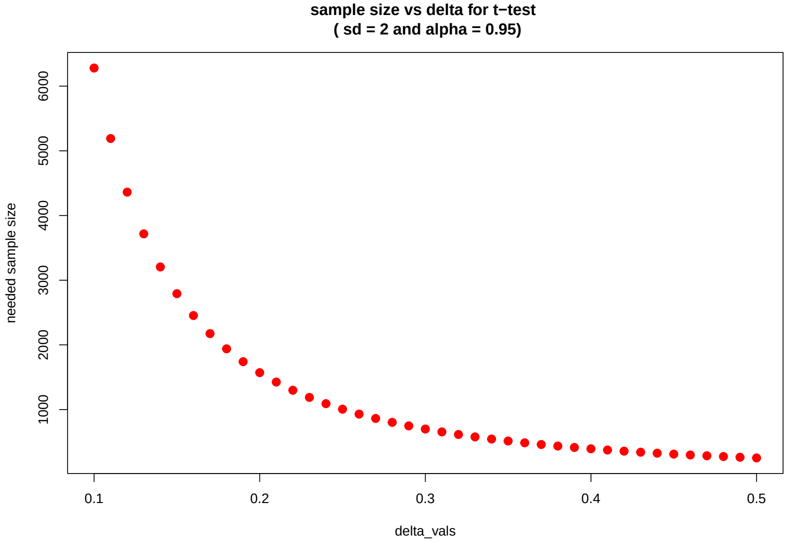
- This just uses one of the apply functions to repeat the function `pow.test` for each element of the vector “delta_vals”.

power.t.test

```
power.n <- sapply(delta_vals, pow.test)
```

- This just uses one of the apply functions to repeat the function `pow.test` for each element of the vector "delta_vals".
- Thus for each value in the vector "delta_vals" (from 0.1 to 0.5), it inputs this value into `pow.test()` and then returns the estimated n (# of observations needed to achieve this power).

power.t.test



sample size vs delta for t−test
(sd = 2 and alpha = 0.95)
needed sample size
1000
2000
3000
4000
5000
6000
0.1
0.2
0.3
0.4
0.5
delta_vals

Similarly, there are functions in base R for 1-way ANOVA

`power.anova.test` example

More complex power analyses

- `pwr` has many useful functions for experimental designs of simple to moderate complexity.
- `pwrss` does as well, and can generate some very helpful figures to help understand
- If you are designing experiments and you think it is likely you are going to use mixed models, the `simr` is a good choice to learn (relatively straightforward)
- `EMSS` has useful sample size calculators.

role your own with monte carlo simulations

- It is relatively straightforward to loop this and generate more complex power analyses.

role your own with monte carlo simulations

- It is relatively straightforward to loop this and generate more complex power analyses.
- Learning how to do simple *Monte Carlo* simulations can give you a lot of flexibility to do this.

role your own with monte carlo simulations

- It is relatively straightforward to loop this and generate more complex power analyses.
- Learning how to do simple *Monte Carlo* simulations can give you a lot of flexibility to do this.
- I have posted a series of screencasts on youtube, starting here that will teach you the basics.

Monte carlo power analysis example

- `R` code is hidden (but you can see it with the .Rmd file)

Plotting results from a power analysis

Plotting results from a power analysis

Plotting results from a power analysis