Statistics with Sparrows - 06

Julia Schroeder

2022

HO 06

Statistical power

Learning aims

- Conduct a statistical power test

Hypothetically …

After your graduation, you are approached by the Ministry of Magic (MoM) to investigate how the harvesting of horns from Romanian Longhorn dragons could be made more economical. If we'd knew whether male or female Romanian Longhorns had longer horns, we could systematically capture the sex with the longer horns. While not a particularly pleasant job, it is still the year of the pandemic and jobs are rare. Also, the MoM is a powerful institution and maybe, if you do this job, you can get a foot in the door for a proper biologist's job, so you accept. When you get started, you desperately hope, you don't have to measure too many individuals to get the answer, so a statistical power analysis to find out the reasonable number of dragons to measure is a good idea. You know that in a very old book you've read once, it has been mentioned that while horns are generally around 1m in length, females have longer horns by about 30cm. You look up that reference, but it remains vague. It does say it only measured 5 dragons and the standard deviation was 1.2. That seems a lot to you, but, alas. To work!

First, we google for R packages that run power analyses. There's a couple: `pwr()` is one that seems really nice, but at the time of writing, it's not been updated for the most recent version of R. If it has, do try it out! There's also a number of websites that directly compute the statistical power. But you want to do it in R! So on we look. Today, we're using `WebPower()`. If it's not installed on your machine, do that now. And don't ask a question online why the code doesn't run if you haven't installed the package yet - I assume you know how to install packages by now!

```r
rm(list=ls())
# we never forget this one!

require(WebPower)

?WebPower
```

Ok, that reads useful. Please click on the blue link “Index” in the help viewer and have a read through.

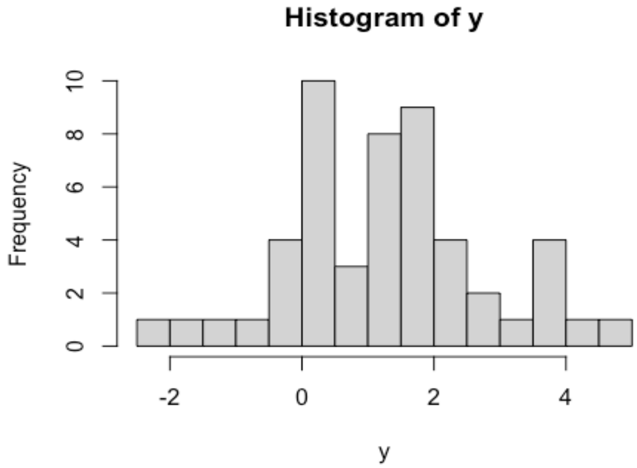
Of course, it’s always the last entry we’re interested in. Find the entry for “Statistical Power Analysis for t-Tests”, click on it, and have a read through.

The aim is to sample as many males as female dragons. Then we need our effect size d - also called *Cohen’s d*. It’s calculated as I said in the lecture as the difference between two means, divided by the standard deviation. What it does, in practice, is to provide us with a ratio - effect size to standard deviation. We take the values from the brief:

```
0.3/1.2

## [1] 0.25
```

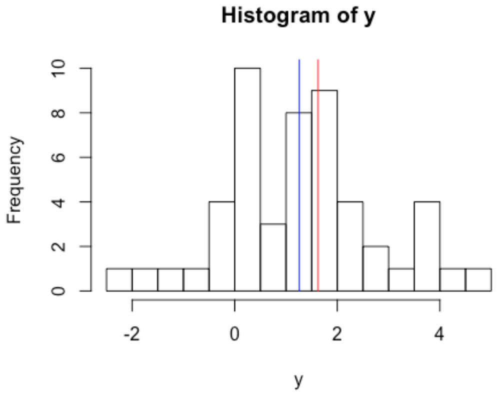
So d is 0.25. That means we want to detect an effect size that is 1/4 of a standard deviation. If we visualize that in a standard deviation density plot, it would look like the following. I simulate data with a mean of 1m, and a standard deviation of 1.3. Horns can’t be shorter than zero, so that’s our cut-off.

```
y<-rnorm(51, mean=1, sd=1.3)
x<-seq(from=0, to=5, by=0.1)

length(x)

## [1] 51

plot(hist(y, breaks=10))
```



```
mean(y)

## [1] 1.260218

sd(y)
```

```
## [1] 1.458004

segments(x0=(mean(y)), y0=(0), x1=(mean(y)), y1=40, lty=1, col="blue")

# and now 0.25 sd left of the mean (because females are larger)
segments(x0=(mean(y)+0.25*sd(y)), y0=(0), x1=(mean(y)+0.25*sd(y)), y1=40, lty
=1, col="red")
```



It becomes quite quickly obvious that that’s a fairly small effect size, given the variability of the data. You are a bit disheartened, because this is ominous - you guess you will likely need a larger sample size. But you can’t be sure unless you’ve done the math, so let’s have a go.

```
?wp.t
```

Ok, we need two sample sizes n1 and n2 (actually, that’s what we want to figure out). We’ll assume they are the same, because it’s the easiest - you can hardly predict now how many females and males you’ll capture so going with chance is your best bet. Next, your effect size is 0.25. Alpha is 0.05 - that’s set to default anyways - in 5% of all cases you won’t detect an effect that’s actually there. As for the power - you want your statistical power to be 0.80 or higher - because you’ve heard in my lecture that’s conventionally agreed upon reasonably statistical power - a 20% chance of a false negative. Type is confusing but you go with two sample because you will have two samples - males and females. You will not compare one sample against a fixed mean. Alternative -direction of the alternative hypothesis. You *think* females may be larger but the reference is cryptic, the data is poor, and the study has been done a long time ago before they knew how to properly sex dragons (it’s not as easy as one might think!). So you leave it with the default - two-sided. You have no clue what the tolerance is - it seems to have something to do with taking roots but it seems safe enough to ignore that for now. So you follow one of the examples below:

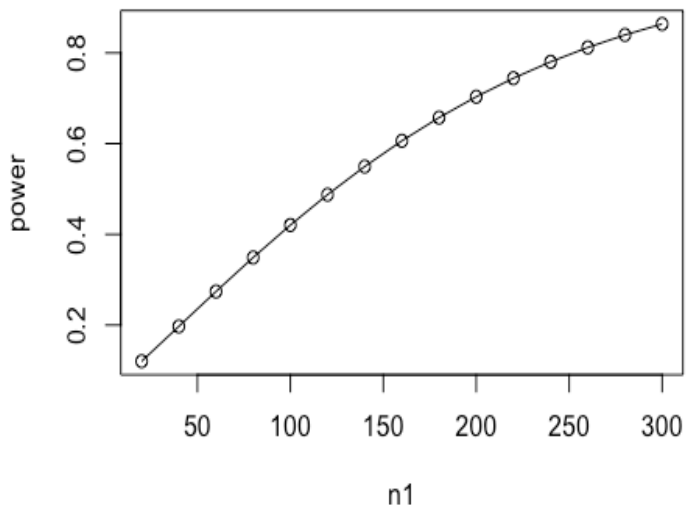
```
wp.t(d=0.25, power=0.8, type="two.sample", alternative="two.sided")

## Two-sample t-test
##
##              n    d alpha power
##     252.1275 0.25  0.05   0.8
```

```
##
## NOTE: n is number in *each* group
## URL: http://psychstat.org/ttest
```

Ah. The result is very clear, if disheartening. You need to sample 252 dragons in each group to get a clear answer. Yikes. You better go back to the MoM to ask for a raise - with that budget you'd never be able to capture that many Romanian Longhorns - if there even are that many alive! But at least you have hard data to show the MoM why you need more cash - if you sample fewer than 504 individuals it's unlikely you'll get a satisfactory answer!

As you look further through the examples in the help file, you see that it is possible to produce a power curve. Maybe that would help to convince the MoM to give you more money? If you could show them how low the power will be with a smaller sample size:

```
res.1<-wp.t(n1=seq(20,300,20), n2=seq(20,300,20), d=0.25, type="two.sample.2n
", alternative="two.sided")
res.1

## Unbalanced two-sample t-test
##
##      n1  n2     d alpha     power
##      20  20 0.25  0.05 0.1203354
##      40  40 0.25  0.05 0.1971831
##      60  60 0.25  0.05 0.2741094
##      80  80 0.25  0.05 0.3490542
##     100 100 0.25  0.05 0.4205383
##     120 120 0.25  0.05 0.4875700
##     140 140 0.25  0.05 0.5495495
##     160 160 0.25  0.05 0.6061828
##     180 180 0.25  0.05 0.6574078
##     200 200 0.25  0.05 0.7033333
##     220 220 0.25  0.05 0.7441884
##     240 240 0.25  0.05 0.7802824
##     260 260 0.25  0.05 0.8119726
##     280 280 0.25  0.05 0.8396409
##     300 300 0.25  0.05 0.8636746
##
## NOTE: n1 and n2 are number in *each* group
## URL: http://psychstat.org/ttest2n

plot(res.1, xvar='n1', yvar='power')
```



This is nice - it shows how much the power increases with increasing sample size - of one group! So, the only hope you have in getting away with a smaller sample size is if the standard deviation is somewhat smaller than 1.2m, or, if the effect size is actually larger than 0.3m. But you won't find out until you measure the dragons and as there will be a danger of false positives and negatives, there's no other way than measuring 500 Longhorns!

Exercises:

1. *You are given a dataset by your PI, where an experiment has been performed where the growth of two groups of bacterial colonies (n each group = 300) has been compared. Your PI has found an interesting difference between both groups - to an effect size of 0.11, with a p = 0.044. The result is rather ground breaking, but the PI is worried that it might be a misleading result. The PI didn't do a statistical power analysis, so they are uncertain about whether they had sufficient power. The bacterial colonies have since died, because nobody could enter the lab for months due to a pandemic. Hence, to repeat the experiment would be very expensive and time consuming. Calculate the statistical power of this test. Then discuss what the result means and how the PI should proceed.*

2. *Write up a methods section that details what you have done and justifies your decisions. Then write a results section that details your results, and use tables and/or graphs to support your points. Then write up a conclusion section, where you must make one suggestion of how, and why, the PI should proceed.*

3. *Philosophical discourse: In your groups, discuss what the consequence of Type I and Type II errors are for the body of published literatures. If one in twenty results may be a false positive or negative, then what does that mean for the whole body of literature? Discuss repercussions and potential solutions.*