ADS2 Week 2.4 Problem Sheet. Power and Sample Size

Rob Young (based on work by Xianghua Li and Wanlu Liu)

robert.young@ed.ac.uk

Question one. Two-sample sample size estimation

An investigator is planning to study the association between coffee consumption and average grade point among college seniors. The plan is to categorize students as heavy drinkers of coffee as those using 5 or more cups of coffee on a typical day as the criterion for heavy consumption. Mean grade point averages will be compared between students classified as heavy drinker versus non-heavy drinkers, using a two-sample unpaired test of means. The standard deviation in the grade point averages is assumed to be 0.42 and a meaningful difference in grade point averages (relative to coffee consumption status) is 0.25 units. How many college seniors should be enrolled in the study to ensure that the power of the test is 80% to detect a 0.25 unit difference in mean grade point? Use a two-sided test with a 5% level of significance.

This can be explored using the `power.t.test` function. The delta value is 0.25 (the unit difference in mean grade point), the standard deviation is 0.42, the alpha value is 0.05, and the required power is 0.8. The required t-test will be two-sampled, and two-sided.

```r
power.t.test(d = 0.25, sd = 0.42, sig.level = 0.05, power = 0.8,
    type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 45.28604
##           delta = 0.25
##              sd = 0.42
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

You require 46 students in each group (heavy vs non-heavy drinkers).

Question 2. relationship between statistical power, sample size, significance level and effect size.

- Take a two-sample, two-sided test, with significance level of 0.05. The standard deviation of our sample is 0.5. If our sample size is 20, what's the statistical power under an effect size (mean difference of two population) of 0.4?

When I run this, I assume a total sample size of 20 and therefore 10 in each group. You may have done this slightly differently.

```r
delta = 0.4
sample_size = 10
standard_deviation = 0.5
alpha = 0.05
```

```
power.t.test(n = sample_size, d = delta, sig.level = alpha, sd = standard_deviation,
    type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 0.4
##              sd = 0.5
##       sig.level = 0.05
##           power = 0.3949428
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The statistical power is 0.39.

- What happens to our statistical power (increases or decreases) if we increase our significance level to 0.1?

```
delta = 0.4
sample_size = 10
standard_deviation = 0.5
alpha = 0.1
power.t.test(n = sample_size, d = delta, sig.level = alpha, sd = standard_deviation,
    type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 0.4
##              sd = 0.5
##       sig.level = 0.1
##           power = 0.5303875
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The statistical power is now 0.53. It has therefore increased when we increased the significance level.

- What happens to our statistical power (increases or decreases) if we decrease our sample size to 10?

I am going to keep the parameters as the second example (significance level of 0.1). You might have done it with different values but should still see the same direction of effect.

```
delta = 0.4
sample_size = 5
standard_deviation = 0.5
alpha = 0.1
power.t.test(n = sample_size, d = delta, sig.level = alpha, sd = standard_deviation,
    type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 5
##           delta = 0.4
##               sd = 0.5
##        sig.level = 0.1
##            power = 0.31333
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

The statistical power is now 0.31. The power has been reduced when we reduced the sample size (despite keeping the standard deviation constant).

- What happens to our statistical power (increases or decreases) if we increase our effect size to 0.8?

```r
delta = 0.8
sample_size = 5
standard_deviation = 0.5
alpha = 0.1
power.t.test(n = sample_size, d = delta, sig.level = alpha, sd = standard_deviation,
    type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##                n = 5
##            delta = 0.8
##               sd = 0.5
##        sig.level = 0.1
##            power = 0.7464331
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

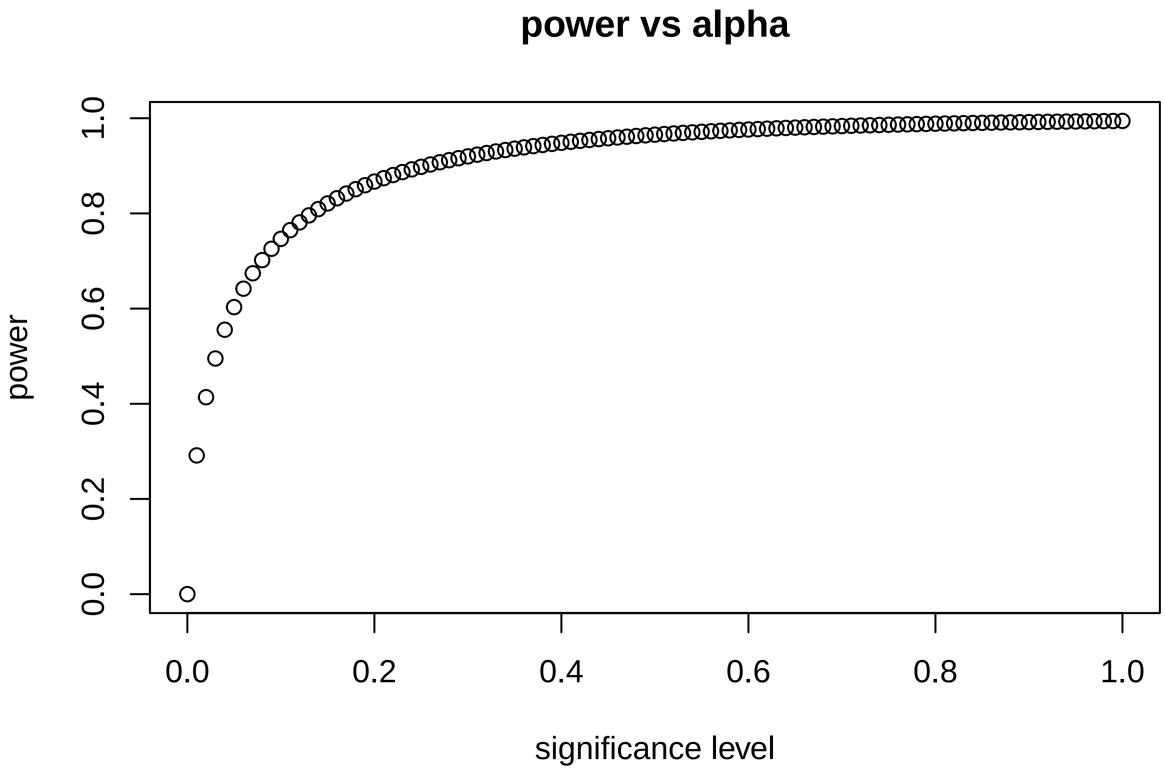
Statistical power is now 0.75. By requiring a large effect size, we have increased the probability of rejecting the null hypothesis of no difference between the groups.

- Advanced challenges (optional). If you can figure it out yourself, you are an absolute 'R master'!

Based on the example from the above, can you use simulation and R plotting to figure out the relationship between statistical power vs significance level, sample size, and effect size? Three curves with y-axis being the statistical power and x-axis being the different significance level, or sample size, or effect size (each simulate 100 data points for plotting).

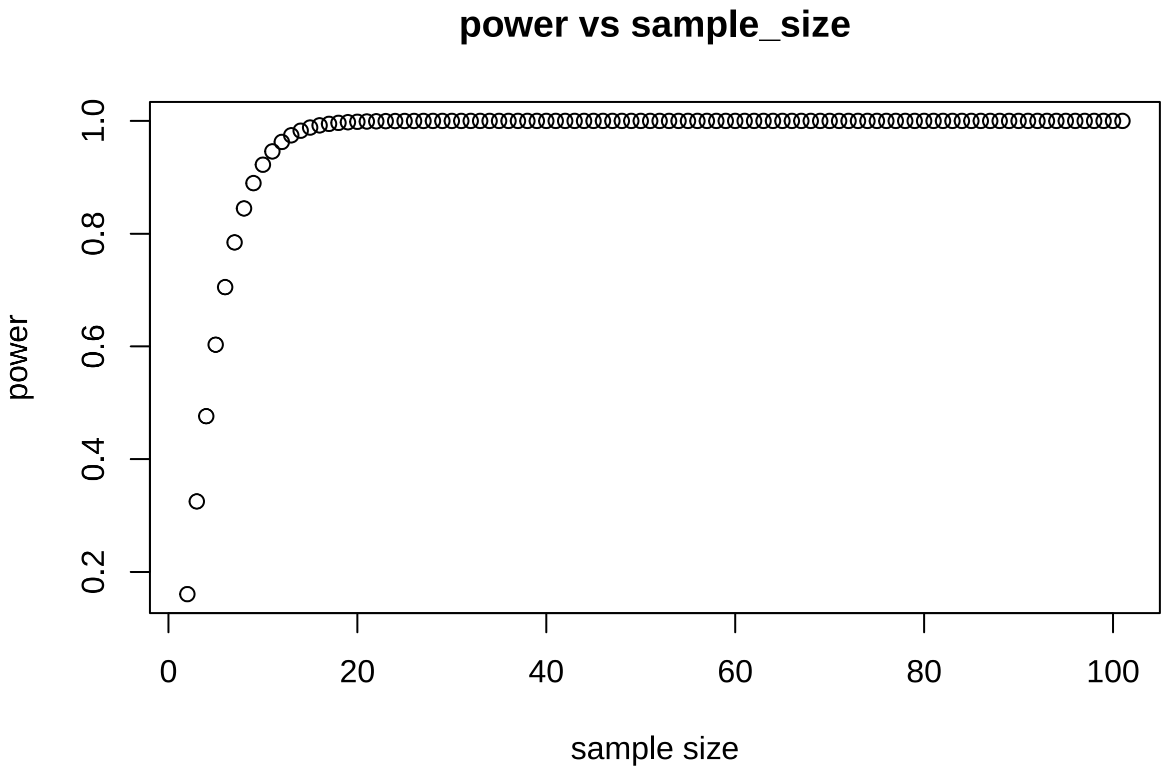
Lets first look at statistical power vs significance level.

```r
delta = 0.8
sample_size = 5
standard_deviation = 0.5
alphas = seq(0, 1, by = 0.01)
power = rep(0, length(alphas))  #initialize power vector
for (i in 1:length(alphas)) {
    power[i] = power.t.test(n = sample_size, sd = standard_deviation,
        d = delta, sig.level = alphas[i], type = "two.sample",
        alternative = "two.sided")$power
}
plot(x = alphas, y = power, xlab = "significance level", ylab = "power",
    main = "power vs alpha")
```



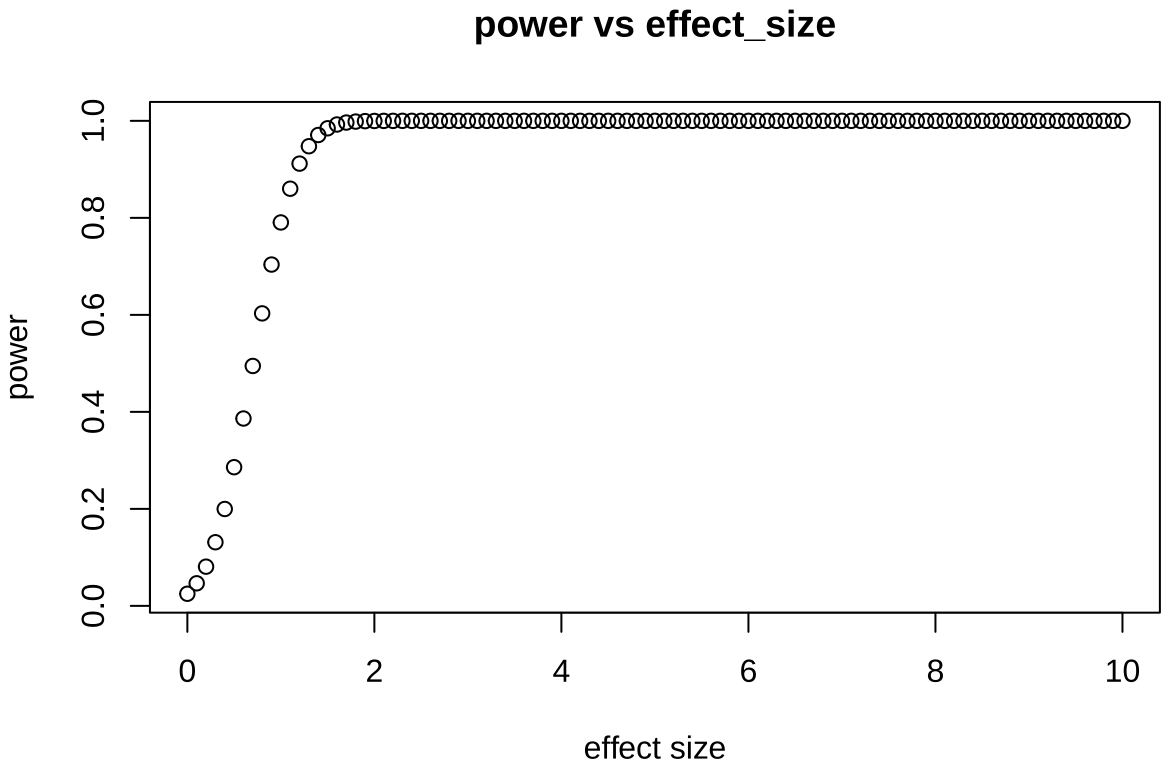
Now power vs sample size. Note I have gone back to the standard alpha threshold of 0.05.

```r
delta = 0.8
standard_deviation = 0.5
alpha = 0.05
sample_size = c(2:101)
power = rep(0, length(sample_size))  #initialize power vector
for (i in 1:length(sample_size)) {
    power[i] = power.t.test(n = sample_size[i], sd = standard_deviation,
        d = delta, sig.level = alpha, type = "two.sample", alternative = "two.sided")$power
}
plot(x = sample_size, y = power, xlab = "sample size", ylab = "power",
    main = "power vs sample_size")
```



Here is the code for power vs effect size.

```r
standard_deviation = 0.5
deltas = seq(0, 10, by = 0.1)
sample_size = 5
power = rep(0, length(deltas))  #initialize power vector
for (i in 1:length(deltas)) {
    power[i] = power.t.test(n = sample_size, sd = standard_deviation,
        d = deltas[i], sig.level = 0.05, type = "two.sample",
        alternative = "two.sided")$power
}
plot(x = deltas, y = power, xlab = "effect size", ylab = "power",
    main = "power vs effect_size")
```



Note here that increasing the alpha significance threshold continues to have an effect over most of the parameter space (ie. $0 < p < 1$). On the other hand, there seems to be little effect of sample size once you have collected at least 10 samples for each group and increasing the effect size beyond 2 has little influence on the statistica power.

How might you use these observations to inform your experimental design?

3. The relationship between sample size vs p-value

- First, for sample A, let's generate 5 random numbers from a normal distribution with mean of 10 and sd of 5. Then, for sample B, let's generate 5 random number from normal distribution with mean of 11 and sd of 5. Now we want to compare whether there is any significance difference between the mean of sample A and B, what should we do? write out the R code. Is there any significant difference for the mean of sample A and B?

```r
set.seed(13)
a = rnorm(5, mean = 10, sd = 5)
b = rnorm(5, mean = 11, sd = 5)
t.test(a, b)
```

```
##
##  Welch Two Sample t-test
##
## data:  a and b
## t = -0.10438, df = 7.6853, p-value = 0.9195
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.636179  5.151360
## sample estimates:
## mean of x mean of y
##  13.37906  13.62147
```

I have used a t-test because the data are continuous, independent, and have a normal distribution (by definition, as I used the rnorm function). With a p-value > 0.9, we can conclude that there are no significant differences for the mean of samples A and B.

- What if we now we increase the sample size to 500 (instead of 5) for sample A and B? Is there any significant difference for the mean of sample A and B? Write out R code.

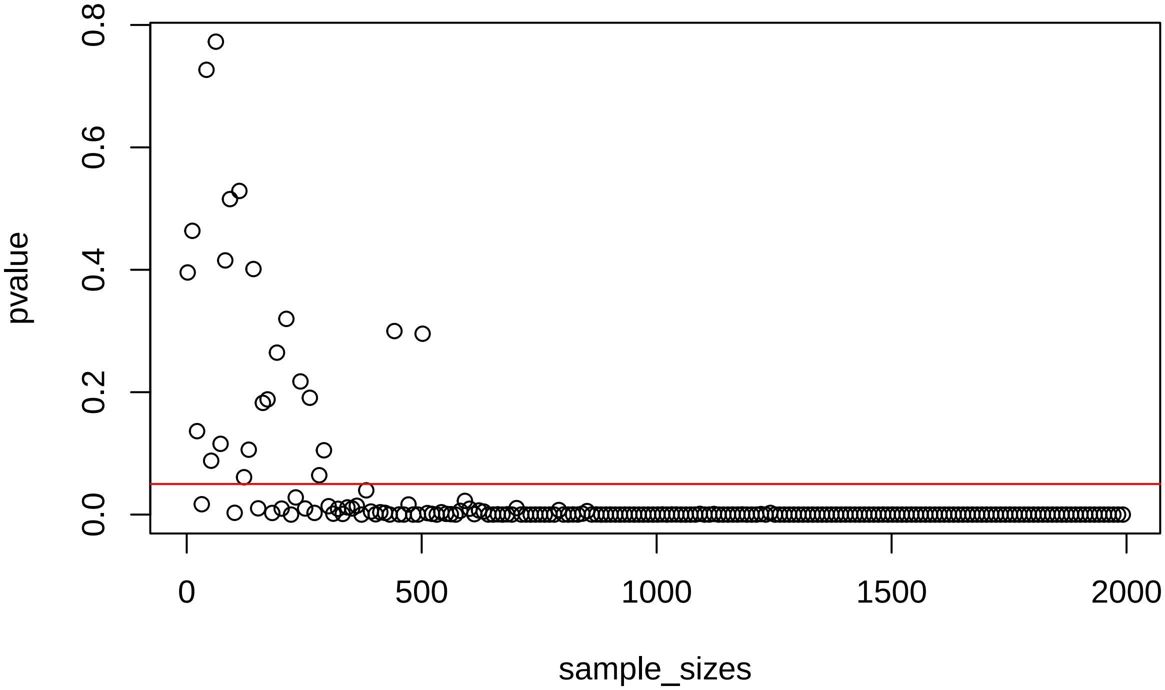
```
a = rnorm(500, mean = 10, sd = 5)
b = rnorm(500, mean = 11, sd = 5)
t.test(a, b)
```

```
##
##  Welch Two Sample t-test
##
## data:  a and b
## t = -3.3992, df = 988.45, p-value = 0.0007028
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.696671 -0.454689
## sample estimates:
## mean of x mean of y
##  9.926613 11.002293
```

Now, there is a highly significant result. The p-value < 0.05 indicates that there is a clear difference between the means of sample and B.

- Advanced challenges (optional). If you can figure it out yourself, you are an absolute 'R master'! Just like what we did earlier, can you use simulation and plotting to visualize the relationship between sample size and p-value? Plot out the curve with p-value on the y-axis and different sample number on the x-axis (with at least 100 datapoints).

```
set.seed(13)
sample_sizes = seq(2, 2000, by = 10)
a = list()  #initialize a, why I use list to initialize a here?
b = list()  #initialize b
pvalue = rep(0, length(sample_sizes))
for (i in 1:length(sample_sizes)) {
    a[[i]] = rnorm(sample_sizes[i], mean = 10, sd = 5)
    b[[i]] = rnorm(sample_sizes[i], mean = 11, sd = 5)
    pvalue[i] = t.test(a[[i]], b[[i]])$p.value
}
plot(x = sample_sizes, y = pvalue)
abline(h = 0.05, col = "red")
```



Here, I have used the t-test as previously. You can see that, overall, the p-value decreases as the sample size increases. Once I have a sample size above ~250, almost all my simulations show a significantly different mean value (t-test, $p < 0.05$) between samples A and B.