BIOM262 Winter 2018. Stats homework week 3.

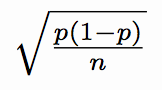
Statistical power.

1) Based on your background reading of the literature, you think that allelic variation at a particular genetic locus may help you assign individuals of an endangered mammalian species to one of two subspecies. There are two alleles at this locus and both are present in both subspecies, but in order to be useful to you, their frequencies need to differ. Previous studies have shown that allele **A** has a frequency of 52% in subspecies 1 (allele **a** is at 48%). Subspecies 2 has been much less studied, but you think that the frequency of allele **A** is larger than 52%. Your goal is to figure out whether the frequency of **A** in subspecies 2 is actually higher than in subspecies 1. In other words, you want to know whether the frequency of **A** in subspecies 2 is greater than 52%. Unfortunately, you only have enough funding to genotype 50 individuals (100 alleles) from subspecies 2. Given these restrictions, what would the actual frequency of **A** have to be in subspecies 2 at a minimum such that you would have a at least an 80% chance of being correct if you rejected your null hypothesis with a significance cutoff of 5%? Try to get within 0.005 of the minimum frequency.

Write an R script to do this. It may help you to sketch out a picture of the situation with your cutoff marked. It will also help if you work out a plan or flowchart for the steps involved in doing this before starting to program. Some useful facts:

(1) the sampling distribution for a proportion is normally distributed

(2) the standard error for a proportion *p* (the standard deviation of its sampling distribution) is:



where *n* is the sample size. In R sqrt() is the square root function.

Please answer the following:

a) What is your null distribution? What is its standard error? Where is it located (centered)? How could you generate it?

b) What is the statistic value at your significance cutoff?

c) How can you use the information from (b) to figure out where the alternative distribution would need to be located (by located I mean where it is centered) to get 80% power? What would its standard error be?

A big recommendation:

It may be easiest to start by coding how to do this for a specific hypothesized frequency of **A**, say 60%. What would be the power in that case? Once you have that code working, all you would need to do would be to try different values for this frequency until you find one that gives you 80%. [think for loop + if statement or perhaps a while loop] A useful R function is seq():

seq(a, b, by=d)

gives you a vector that goes from a to b in increments of d. For example:

> seq(4, 6, by=0.4)

[1] 4.0 4.4 4.8 5.2 5.6 6.0

Once you have working code for this problem, make it more general. Use variables in your code so that you could change the null frequency, the significance cutoff, and the desired chance of being correct (the power) at the top of your code and your script would automatically use those values. So if you use specific numbers in the guts of your code, replace them as much as possible with variables and then assign the variables values up at the top, something like:

nullValue= 0.3

significanceCutoff = 0.2

power = 0.7

That way it is easy to change your script to deal with a new situation - you just change the variable values up top and the rest of the code still works.

2) In class we discussed the permutation/scrambling/randomization method for making null distributions that represent a null model of no association between variables. Write code that uses permutation to test for an association between Height and Pulse rate in the NHANES dataset. (description here: <https://cran.r-project.org/web/packages/NHANES/NHANES.pdf>). Do not use the boot package to do this. Instead, build it out of some combination of for loops, if-then statements, sample(), and replicate(). There are different ways to code this, so if you don’t use all of these programming constructs, that’s fine as long as it works. These are both quantitative variables, so you could use correlation or regression to ask about association.

To get the dataset, install the package NHANES in R. Then load the library.

After this type:

data(NHANES)

at the prompt. This loads a data frame called NHANES into your workspace. Height and Pulse are two variables in that data frame

Note that some of the values of these variables are NA. NA means “Not Available” and it means that the value of the variable is missing for that case. You can test to see if a value is NA by using the function:

is.na()

which returns True or False depending. You will want to remove these cases before continuing.

For this problem, use regression to see if there is a significant association between pulse and height.

Please report the following:

a) What is the statistic you should use for this type of data and how will you calculate it?

b) What significance cutoff will you choose?

c) What would be interesting – a positive association, a negative association, an association in either direction? How does this affect your hypothesis test?

d) Create a null distribution using the permutation/scrambling/randomization method and report the p-value you get for your statistic

Hints:

You can estimate a linear model in R in the following way (you also did so in the data wrangling session):

linearModelResult <- lm(responseVariableName ~ explanatoryVariableName, data= dataFrameName)

This returns an lm object which is an object with several different fields. Notice that the names are unquoted. One of the fields is 'coefficients.' For a simple linear model, this gives you a vector with the values of the estimated intercept and slope. This might come in handy.

Report the p-value from your test and draw a conclusion based on your p-value and your significance cutoff.

3) Write code to estimate a bootstrap confidence interval for the association between Height and Pulse rate. As above, do not use the boot package. Report your confidence interval and interpret what it means.

4) Make sure your code runs error free from start to end in one go without me having to specify anything! It should work by me simply hitting the run button.

5) Submit both the answers and the code to me in a format similar to below:

Lastname\_Firstname\_CMM262\_answers.doc

Lastname\_Firstname\_CMM262\_code.r