

## Week 2: Null-hypothesis significance testing

### Exercise 1: The MLE versus the null ritual

A consistent failure of the significance testing approach is that it encourages scientists to conclude that there is no effect, when in fact the data suggest otherwise. This is like saying that because  $p > 0.05$ , our best guess of the population mean is actually zero, even when our lying eyes (the MLE) tell us otherwise. In this problem, you will get a feel for the zaniness of this practice.

(a) The code below will simulate sampling from a population of student heights. We are interested in whether male students are taller than female students. That is, we want to test the hypothesis that the population average male height is larger than the population average female height. So we sample from the population and compare sample means for males and females. In the true population we'll generate, males are indeed about 2 inches taller than females, on average.

The code below sets up the project. Just enter all of this at the command prompt:

```
f <- sample( c(0,1) , 1000 , replace=TRUE )
w <- rnorm( n=1000 , mean=80-f*2 , sd=6 )
sim.b <- function( pt ) {
  n <- as.integer(pt$x)
  i <- sample( 1:1000 , n )
  sf <- f[i]
  sw <- w[i]
  b <- as.numeric( mean(sw[sf==0]) - mean(sw[sf==1]) )
  #p <- t.test( sw ~ sf )$p.value
  points( n , b )
}
plot( 0 , 0 , col="white" , ylim=c(-1,5) , xlim=c(20,500) , xlab="sample size" ,
      ylab="estimate" )
true.b <- mean(w[f==0]) - mean(w[f==1])
lines( c(0,600) , c(true.b,true.b) , col="red" , lty=2 )
```

You should see an empty plot with sample size along the horizontal axis and effect size on the vertical axis. The red dashed line is the location of the true population mean difference between male and female heights.

Now to simulate samples, execute the single line below:

```
sim.b( locator(n=1) )
```

Click on the plot and R will sample from the population, using the sample size at the point you clicked. So if you click on the left side of the plot, you will use a small sample size. If you click on the right side of the plot, you will use a large sample size. Execute this single line many times, adding many points to the plot, at different sample sizes. When you start to get a sense from all that clicking, you can quickly simulate a bunch of points at random sample sizes with the line below:

```
z <- sapply( sample(20:500,200,replace=TRUE) , function(z) sim.b( list(x=z,y=0) ) )
```

Is there any trend in the estimates, as a function of sample size? Why or why not? Does the central tendency of the estimate change as a function of sample size? Does the variance in the estimate change as a function of sample size?

(b) The code below uses the same population of data, but now computes t-tests for its samples. Again you click to choose sample sizes and R will plot simulated  $p$ -values from the t-tests of male and females heights. Again ask if there is any trend, and why.

```
sim.p <- function( pt ) {
  n <- as.integer(pt$x)
  i <- sample( 1:1000 , n )
  sf <- f[i]
  sw <- w[i]
  #sub.b[j] <- as.numeric( mean(sw[sf==0]) - mean(sw[sf==1]) )
  p <- t.test( sw ~ sf )$p.value
  points( n , p )
}
plot( 0 , 0 , col="white" , ylim=c(0,1) , xlim=c(20,500) , xlab="sample size" ,
      ylab="p-value" )
lines( c(0,600) , c(0.05,0.05) , col="red" , lty=2 )

sim.p( locator(n=1) )
```

To quickly plot a bunch of random sample sizes, use:

```
z <- sapply( sample(20:500,200,replace=TRUE) , function(z) sim.p( list(x=z,y=0) ) )
```

How does the average  $p$ -value change as sample size changes, and how is this different from how the average estimate changes (or not), as sample size changes?

## Exercise 2: The rejection envelope

One symptom of accepting and rejecting hypotheses by using  $p$ -values is that there is a temptation not to publish studies that do not attain statistical significance. As a result, published studies may tend to contain effects that lie just beyond the effect size required for  $p < 0.05$ . As sample size increases, smaller effect sizes are needed to reach this threshold, and so published effect sizes may decline as sample size increases. But this bias isn't always present—many researchers publish fairly, regardless of significance. In this exercise, you'll plot such a rejection envelope—also called a “funnel graph”—and explore its consequences.

(a) First, we are going to simulated sampling births from a population in which male and female births are equally likely, so the population proportion of males (sometimes called a sex “ratio”) is 0.5. The code below simulates this sampling 100 times, across a range of sample sizes from 5 to 500.

```
# random sample sizes, between 5 and 500, with a bias for smaller samples
n <- sample( 5:500 , 100 , replace=TRUE , prob=500:5 )
# random births, sampling from a binomial distribution,
```

```
# because this is like coin tossing with discrete outcomes
r <- rbinom( length(n) , size=n , prob=0.5 ) / n
# now plot the simulated data and draw the population mean as a dashed line
plot( r ~ n , xlab="sample size" , ylab="proportion males" , ylim=c(0,1) )
lines( c(-10,600) , c(0.5,0.5) , lty=2 )
```

This is much like what you've done previously in this week's exercises. What phenomena about the relationship between sample size and our estimates do you again notice here?

**(b)** Now we're going to superimpose on the plot from (a) a rejection envelope, at the boundaries where an effect size (estimate) is far enough away from the null hypothesis of 0.5 for  $p$  to be less than 0.05.

Since the  $p$ -value is defined as the probability of the data-or-more-extreme-data, assuming the null is true, we can compute it from the tails of the cumulative binomial probability distribution. And like many common distributions, this one is built into R. The function `qbinom()` takes a probability  $P$  as an input and returns as an answer the number of successes that mark the beginning of the tail of the distribution in which the cumulative probability is the same as the supplied probability  $P$ . Here's the code to do the grunt work:

```
# generate list of sample sizes to compute rejection thresholds for
ns <- (1:50)*10
# now compute threshold r values for each sample size
th.up <- qbinom( 0.025 , ns , 0.5 , lower.tail=FALSE )/ns
th.down <- qbinom( 0.025 , ns , 0.5 , lower.tail=TRUE )/ns
# plot the boundaries
lines( ns , th.up )
lines( ns , th.down )
```

This should superimpose the rejection boundary over your previous plot, so re-run the code to do the simulation, if necessary, first.

What do you notice about the rejection envelope? Does this agree with your results from the previous exercise?

**(c)** Modify the code from parts (a) and (b) to simulate data from a population in which the true proportion of male births is 0.6. Continue to plot rejection boundaries for the null hypothesis that the true value is 0.5. What relationship do you observe between sample size and the odds of attaining a significant result?

**(d)** What would the raw (simulated) data look like, if you deleted (people didn't publish) those estimates that failed to attain significance?

**(e) BONUS** Can you modify the code to plot the thought experiment in part (d)? **HINT:** Try using indexing to extract only those samples that are significant, as defined by the rejection thresholds. Don't spend a lot of time trying to solve this coding problem, but those with a little more programming experience may enjoy the challenge.

### Exercise 3: A real example of a rejection envelope

The data contained in the file "FA meta data.csv" are 146 effect sizes and sample sizes from the Møller and Thornhill paper, "Bilateral Symmetry and Sexual Selection: A Meta-Analysis," which is also on the website this week. To read this file into R, use the line:

```
d <- read.csv( file.choose() )
```

You'll get a regular file-chooser window, and you can navigate your computer and find where you saved the data file. The data will be read into an object with the name `d`.

Each row of the data frame is a statistical test from a paper. The data frame has three columns (that we care about):

1. `r`: The correlation (effect size) from each paper.
2. `n`: The sample size from the paper.
3. `authors`: A 1/0 dummy variable marking whether the authors of the meta-analysis paper were involved in a particular study that was part of their meta-analysis.

(a) Plot the effect sizes ( $r$ ) against sample sizes ( $n$ ). Also plot a horizontal line where the null hypothesis of zero effect lies. (You might find that putting the x-axis on a log scale helps with seeing the variation in the points. Add `log="x"` to the function call, if so.)

(b) Now we'll compute the rejection envelope for these studies. These effect sizes are parametric correlations, so we just have to use the formula for the  $t$ -value of a correlation coefficient:

$$t = r \sqrt{\frac{n-2}{1-r^2}},$$

and solve it for  $r$ , as we want to know the  $r$ -value at which  $t$  reaches statistical significance. R contains all the statistical tables needed to convert between  $t$  and  $p$ , given sample size (degrees of freedom). Solving the above formula for  $r$ :

$$r = \pm \frac{t}{\sqrt{n-2+t^2}}.$$

Here's the code to do what I've explained above:

```
# a function to compute r from t and n
r.from.tn <- function(n) {
  # looks up the t value corresponding to a tail
  # of 2.5% (two-tailed 5% significance test)
  the.t <- qt(0.025,n-2)
  # return the r value, using our formula
  the.t / sqrt( n - 2 + the.t^2 )
}
```

This gives us one side of the boundary, and we just multiply this answer by  $-1$  to get the other side. So this code will compute the boundaries:

```
th1 <- sapply( 1:500 , function(z) r.from.tn(z) )  
th2 <- -th1
```

Plot these superimposed on the data, using the `lines()` function. Do you think there is any publication bias here? Does it look like there is a true effect that is far from zero? What features of the shape of the cloud of points lead you to these conclusions? That is, state what shape you expect these points to have, if there has been no publication bias.

(c) Try plotting only those studies in which the authors of the meta-analysis were not involved. Does it look like there has been publication bias in these studies? Does it look like there is a true effect that is far from zero? What features of the shape of the cloud of points lead you to these conclusions? That is, state what shape you expect these points to have, if there has been no publication bias.

(d) Now plot only those studies in which the authors of the meta-analysis *were* involved. That is, only plot those rows where `authors==1`. Does it look like there has been publication bias here? How do these points differ from those you plotted in (b) and (c)?