Week 3 Homework

Shu Guo

Sunday, February 08, 2015

## Homework 2

Load the babies dataset:

babies <- read.table("babies.txt", header = TRUE)

This is a very large dataset(1,236 cases), and we will pretend that it contains the entire population that we are interested in. As before, we will study the differences in birthweight between babies born to smoking and non-smoking mothers.

Fist, we split the into two birthweight datasets, one of birthweights to non-smoking mothers, and the other of birthweights to smoking mothers.

bwt.nonsmoke <- babies$bwt[babies$smoke == 0]  
bwt.smoke <- babies$bwt[babies$smoke == 1]

Now, we can look for the true population difference in means between smoking and non-smoking birthweights.

mean(bwt.nonsmoke)-mean(bwt.smoke)

## [1] 8.937666

sd(bwt.nonsmoke)

## [1] 17.39869

sd(bwt.smoke)

## [1] 18.09895

The population difference of mean birthweights is about 8.9 ounces. The standard deviations of the nonsmoking and smoking groups are about 17.4 and 18.1 ounces, respectively.

We will treat the babies dataset as the full population, and draw samples from it to simulate individual experiments. We will then ask whether somebody who only received the random samples would be able to draw correct conclusions about the population.

We are interested in testing whether the birth weights of babies born to non-smoking mothers are significantly different from the birth weights of babies born to smoking mothers.

Suppose we obtain two samples, each of size N, from non-smoking mothers (dat.ns) and smoking mothers (dat.s). We compute the t-value, which we call tval.

X.ns = mean(dat.ns) sd.ns = sd(dat.ns) X.s = mean(dat.s) sd.s = sd(dat.s) sd.diff = sqrt(sd.ns2/N+sd.s2/N) tval = (X.ns - X.s)/sd.diff

Question 1.1  
Compute the t-value (t-statistic) for the first 30 weights of non-smoking mothers and the first 30 weights of smoking mothers. (The first 30 as they appear in the bwt.nonsmoke and bwt.smoke). Confirm that the t-statistic calculated manually and by t.test() is the same. What is the t-value (t-statistic)?

N <- 30  
dat\_ns30 <- bwt.nonsmoke[1:30]  
X\_ns30 <- mean(dat\_ns30)  
sd\_ns30 <- sd(dat\_ns30)  
  
dat\_s30 <- bwt.smoke[1:30]  
X\_s30 <- mean(dat\_s30)  
sd\_s30 <- sd(dat\_s30)  
  
sd\_diff30 <- sqrt(sd\_ns30^2/N + sd\_s30^2/N)  
  
tval30 <- (X\_ns30 - X\_s30)/sd\_diff30  
tval30

## [1] 0.8299373

## p value generated by t.test()  
t.test(dat\_ns30, dat\_s30)$statistic

## t   
## 0.8299373

Recall that we summarize our data using a t-value because we know that in situations where the null hypothesis is true (what we mean when we say "under the null") and the sample size is relatively large, this t-value will have an approximate standard normal distribution. Because we know the distribution of the t-value under the null, we can quantitatively determine how unusual the observed t-value would be if the null hypothesis were true.

The standard procedure is to examine the probability a t-value that actually does follow the null hypothesis would have larger absolute value than the absolute value of the t-value we just observed -- this is called a two-sided test.

One way to compute this probability is to take one minus the area under the standard normal curve between -abs(tval) and abs(tval). In R, we can do this using the pnorm function, which computes the area under a normal curve from negative infinity up to the value given as its first argument:

pval = 1- pnorm(abs(tval)) + pnorm(-abs(tval))

Question 1.2

Because of the symmetry of the standard normal distribution, there is a simpler way to calculate the probability that a t-value under the null could have a larger absolute value than tval. Choose the simplified calculation from the following:

2\*pnorm(-abs(tval))

Here we will show how to generate the table from data which is in the form of a dataframe, so that you can then perform an association test to see if two columns have an enrichment (or depletion) of shared occurances.

Download the assoctest.csv file. This dataframe reflects the allele status (either AA/Aa or aa) and the case/control status for 72 individuals.

## Homework 3

### Confidence Interval

We will continue using the dataset, babies.txt, from the previous assessment in Inference I.

babies = read.table("babies.txt", header=TRUE)

We are using the baby birthweights in the column, 'bwt'.

Do NOT use the column 'weight', which is the mother's weight.

You can extract the baby birthweights from smoking and non-smoking mothers with the code:

bwt.nonsmoke = babies$bwt[babies$smoke==0]  
bwt.smoke = babies$bwt[babies$smoke==1]

t-test for Difference Between Means (Large Sample Size)

In this example, we will simulate a study where we take a sample of N=30 birthweights from each of the smoking and non-smoking groups. We will build a confidence interval for the difference between the population means using these measurements.

In the last few videos, Rafa showed us how to build a confidence interval for the mean (mu) of a single group whereas in this question, we are constructing a confidence interval for a difference between groups. This is slightly more complex when sample sizes are small, but fortunately R has a built-in function that will do it for us no matter what the sample size.

The t.test function will take as its first two arguments two samples, and return an object that summarizes the difference between their means. This summary that has several attributes, which you can access with the $ operator. For example, if you had samples s1 and s2, you could run:

mytest <- t.test(s1, s2)  
mytest$p.value  
mytest$conf.int

The last two lines would return the p-value and confidence interval for the difference in means between these two samples. See the help page ?t.test for more details.

Question 1.1

Take a random sample of N=30 measurement from each of the smoking and nonsmoking datasets. Then compute the difference in their means and construct a 95% confidence interval for the difference using t.test(). Do this 1,000 times and keep track of all the confidence intervals. One way to answer this question is to use the replicate() function in R, which will return a 2 x 1000 matrix of the lower and upper bound for each of the 1000 replications.

N <- 30  
conf <- matrix(0, 1000, 3)  
count <- 0  
popdiff = mean(bwt.nonsmoke) - mean(bwt.smoke)  
for (i in 1:1000){  
 nons <- sample(bwt.nonsmoke, N)  
 sm <- sample(bwt.smoke, N)  
 tt <- t.test(nons, sm)  
 conf[i,1] <- tt$conf.int[1]  
 conf[i,2] <- tt$conf.int[2]  
 conf[i,3] <- conf[i, 1] < popdiff & conf[i,2] > popdiff  
 }  
head(conf)

## [,1] [,2] [,3]  
## [1,] 2.8247594 21.17524 1  
## [2,] -0.3351833 18.06852 1  
## [3,] 3.0809048 21.91910 1  
## [4,] -0.3064640 17.90646 1  
## [5,] -5.3691633 13.83583 1  
## [6,] 7.8648097 24.40186 1

mean(conf[, 2]) - mean(conf[,1])

## [1] 18.22455

sum(conf[,3])

## [1] 965

What is the average length of the confidence interval?

d = read.csv("assoctest.csv")

Question 1.2

The population-level difference was

popdiff = mean(bwt.nonsmoke) - mean(bwt.smoke)

How often (what proportion of times) did the confidence intervals contain the population-level difference? That is, what proportion of times was the lower bound of the confidence interval less than popdiff and the upper bound greater than popdiff?  
0.965

In the previous video Rafa mentioned that we should report confidence intervals whenever possible because they communicate both an effect size and the statistical significance of a result. Rafa later mentions that when comparing a difference between two groups to zero, we can tell whether the difference has a p-value of less than 0.05 based on whether the confidence interval for the difference contains zero. We will explore that statement in more detail in the next two questions.

Suppose that we have drawn samples of size N=30 from the non-smoking and smoking baby populations and we want to test whether the difference between their means is significant at the alpha=0.05 level.

Recall that when we perform a t-test for the difference between two means, we calculate a t-value like the following.

dat.ns = sample(bwt.nonsmoke, 30)  
dat.s = sample(bwt.smoke, 30)  
X.ns = mean(dat.ns)  
sd.ns = sd(dat.ns)  
X.s = mean(dat.s)  
sd.s = sd(dat.s)  
sd.diff = sqrt(sd.ns^2/30 + sd.s^2/30)  
tval = (X.ns - X.s)/sd.diff

Because our sample sizes are rather large, we can then use the qnorm() function to tell whether tval corresponds to a p-value that is less than 0.05.

1.959964

This tells us that if the absolute value of tval is greater than 1.96, the p-value is less than 0.05 and the result is significant an the 0.05 level.

We can use the same numbers to construct a confidence interval for the difference in means between the smoking and nonsmoking populations. To do this, we follow Rafa's instruction in the CI 2 video

ci.upper = (X.ns-X.s) + sd.diff*1.96*  
*ci.lower = (X.ns-X.s) - sd.diff*1.96

Question 1.3

Fill in the blank: the difference in means (X.ns - X.s) must have absolute value greater than \_\_ 1.96 \_\_ times sd.diff in order for the result to be significant (at alpha=0.05).

Question 1.4

Fill in the blank: the difference in means (X.ns - X.s) must be a greater distance than **1.96**\_ times sd.diff away from 0 in order for the 95% confidence interval not to contain 0.

### Power Calculations

In the many of the analyses this week, we know that the null hypothesis is false -- for example, we knew that the difference in average baby weight in the whole population is actually around 8.9. Thus, we are concerned with how often the decision rule allows us to conclude the that the null hypothesis is actually false. Put another way, we would like to quantify the Type II error rate of the test, or the probability that we fail to reject the null hypothesis when the alternative hypothesis is true.

Unlike the Type I error rate, which we can characterize by assuming that the null hypothesis of "no difference" is true, the Type II error rate cannot be computed by assuming the alternative hypothesis alone because the alternative hypothesis alone does not specify a particular value for the difference, and thus does not nail down a specific distrbution for the t-value under the alternative.

For this reason, when we study the Type II error rate of a hypothesis testing procedure, we need to assume a particular effect size, or hypothetical size of the difference between population means, that we wish to target. We ask questions like "What is the smallest difference I could reliably distinguish from 0 given my sample size N?", or more commonly, "How big does N have to be in order to detect that the absolute value of the difference is greater than zero?" Type II error control plays a major role in designing data collection procedures before you actually see the data so that you know the test you will run has enough sensitivity or power. Power is one minus the Type II error rate, or the probability that you will reject the null hypothesis when the alternative hypothesis is true. Power and alpha

There are several aspects of a hypothesis test that affect its power for a particular effect size. Intuitively, setting a lower alpha decreases the power of the test for a given effect size because the null hypothesis will be more difficult to reject. This means that for an experiment with fixed parameters (i.e., with a predetermined sample size, recording mechanism, etc), the power of the hypothesis test trades off with its Type I error rate, no matter what effect size you target.

We can explore the tradeoff of power and Type I error concretely using the babies data. Load the babies dataset from babies.txt

babies = read.table("babies.txt", header=TRUE)  
bwt.nonsmoke = babiessmoke==0]  
bwt.smoke = babiessmoke==1]

Because we have the full population, we know what the true effect size is (about 8.93), and we can compute the power of the test for true difference between populations.

Take a random sample of N=15 measurements from each of the smoking (bwt.smoke) and nonsmoking (bwt.nonsmoke) groups. Then perform a t-test and compare the p-value to a significance level alpha. Do this 1,000 times. Decide whether or not to reject the null hypothesis based on three significance levels alpha=0.1, alpha = 0.05, alpha=0.01. For each experiment, keep track of whether you correctly rejected the null hypothesis at each of these significance levels (thus, each of the 1,000 experiments should produce 3 numbers to keep track of). For each significance level, in what proportion of the experiments did you correctly reject the null hypothesis?

N <- 15  
p\_1 <- matrix(0, 1000, 2)  
for (i in 1:1000){  
 nons <- sample(bwt.nonsmoke, N)  
 sm <- sample(bwt.smoke, N)  
 tt1 <- t.test(nons, sm, conf.level = 0.9)  
 p\_1[i, 1] <- tt1$p.value  
 p\_1[i, 2] <- tt1$p.value > 0.1  
}  
1-sum(p\_1[,2])/1000

## [1] 0.403

N <- 15  
p\_05 <- matrix(0, 1000, 2)  
for (i in 1:1000){  
 nons <- sample(bwt.nonsmoke, N)  
 sm <- sample(bwt.smoke, N)  
 tt2 <- t.test(nons, sm, conf.level = 0.95)  
 p\_05[i, 1] <- tt2$p.value  
 p\_05[i, 2] <- tt2$p.value > 0.05  
}  
1-sum(p\_05[,2])/1000

## [1] 0.274

N <- 15  
p\_01 <- matrix(0, 1000, 2)  
for (i in 1:1000){  
 nons <- sample(bwt.nonsmoke, N)  
 sm <- sample(bwt.smoke, N)  
 tt3 <- t.test(nons, sm, conf.level = 0.99)  
 p\_01[i, 1] <- tt3$p.value  
 p\_01[i, 2] <- tt3$p.value > 0.01  
}  
1-sum(p\_01[,2])/1000

## [1] 0.104

Question 3.1

What is the power at alpha=0.1?

Question 3.2

What is the power at alpha=0.05?

Question 3.3

What is the power at alpha=0.01?

Question 3.4

We will deal with the important of multiple hypothesis testing in a later course, but for now, consider this question that is very near to your heart. Suppose that one of the homework question is graded based on whether the result you reported falls within a exact 99% interval around a true value. Now suppose that 2,000 students complete the assignment, and assume that all students execute the simulation correctly. What is the expected number of student responses that would be marked wrong simply by chance?

### Association Tests Assessment

Question 2.1

Compute the Chi-square test for the association of genotype with case/control status (using the table() function and the chisq.test() function). Examine the table to see if it look enriched for association by eye. What is the X-squared statistic?

t <- table(d)  
chisq.test(t)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: t  
## X-squared = 3.3437, df = 1, p-value = 0.06746

Question 2.2

Compute the Fisher's exact test ( fisher.test() ) for the same table. What is the p-value?

fisher.test(t)

##   
## Fisher's Exact Test for Count Data  
##   
## data: t  
## p-value = 0.05194  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.940442 8.493001  
## sample estimates:  
## odds ratio   
## 2.758532