DATA557 Homework 3

Will Wright

January 26, 2019

Instructions

Submit your solutions in pdf format to the dropbox on the canvas page by 5:00PM, Wednesday January 30. You do not need to include your R code with your solutions for this assignment.

For question 1, you are to work in groups. For the other questions, you may work together to help each other solve problems, but you should do all the work, create your own solutions, and hand in your own work without copying others' work.

Question 1. (This is Q3 of Exercise 3.)

Suppose that a new experiment is being designed to determine the effect on output of temperatures higher than 100. In particular, the aim of the new experiment is to test the null hypothesis that the mean output is the same for temperature 100 and temperature 120. The researcher would like to have at least 90% power to detect a difference between these conditions in mean output equal to 75. Your job is to determine the sample sizes for each group and to decide which test statistic will be used to test the null hypothesis. Justify your answers.

For this question, you should continue to work with your group on the answer that you developed in class on Jan 23. The group member who did the original posting will receive feedback from me on canvas by Saturday morning and should relay it to the other group members. Work together either in person or electronically to develop your final solution. The group member who did the original posting should include the group's solution in their HW 3 submission, including the names of all group members. The grade assigned for the solution to this question will be applied to the grade for the HW for all group members.

```
# read file
pDat <- read.csv("../WEEK03/process.csv")
set.seed(999)

# calculate input parameters
output_diff <- 75
sd_100 <- sd(pDat$output[which(pDat$temp==100)])
n_vals <- 1:100
reps=1000

library(pwr)
library(MKmisc)

pwr.t.test(sig.level = 0.05,
           d = 75/sd_100,
           power = 0.9,
           type = "two.sample",
           alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 79.06061
##               d = 0.5187659
```

```
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(delta = 75,
             sig.level = 0.05,
             sd = sd_100,
             power = 0.90,
             alternative = "two.sided",
             type = "two.sample",
             strict = TRUE)
```

```
##
##      Two-sample t test power calculation
##
##               n = 79.06061
##           delta = 75
##              sd = 144.5739
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power.welch.t.test(delta = 75,
                   sd1 = sd_100,
                   sd2 = sd_100,
                   sig.level = 0.05,
                   power = 0.90,
                   alternative = "two.sided",
                   strict = TRUE)
```

```
##
##      Two-sample Welch t test power calculation
##
##               n = 79.06061
##           delta = 75
##             sd1 = 144.5739
##             sd2 = 144.5739
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

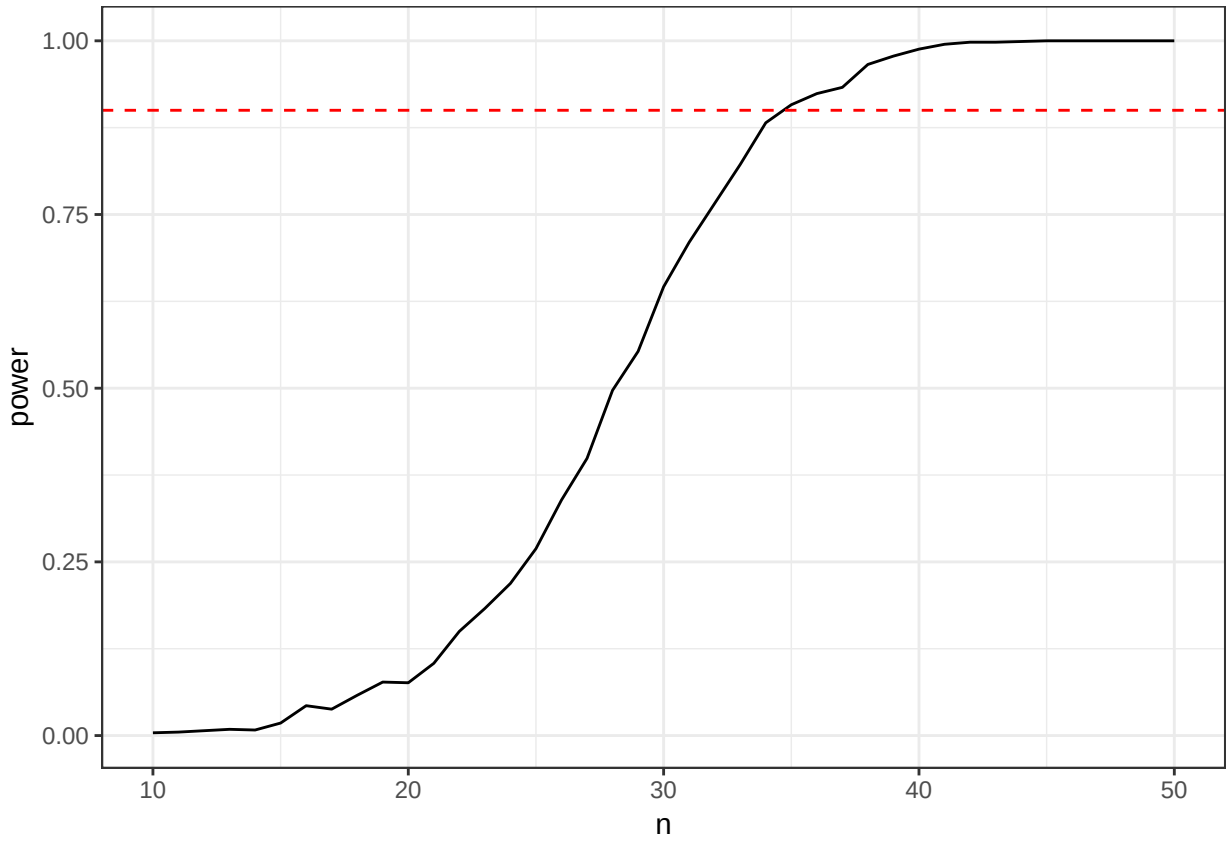
```
# simulations
powers_z <- rep(NA,length(n_vals))
powers_t <- rep(NA,length(n_vals))
powers_w <- rep(NA,length(n_vals))
for(j in 1:length(n_vals)){
  test_statistic <- rep(NA,reps)
  for(i in 1:reps){
    x <- rnorm(n_vals[j], output_diff, sd_100)
    y <- rnorm(n_vals[j], output_diff, sd_100)
```

```
    se <- sqrt(var(x)/n_vals[j]+var(y)/n_vals[j])
    test_statistic[i] <- (output_diff)/se

  }
  powers_z[j] <- mean(abs(test_statistic)>qnorm(0.975)) # for z-test
  # powers_t[j] <- mean(abs(test_statistic)>qt(0.975, df = )) # incomplete for t-test
  # powers_w[j] <- mean(abs(test_statistic)>qt(0.975)) # incomplete for welch's t-test
}
results <- data.frame(n = n_vals, power = powers_z)

g <- ggplot(results, aes(x = n, y = power))
g + geom_line() +
  scale_x_continuous(limits = c(10,50)) +
  theme_bw() +
  geom_hline(yintercept = 0.9, col = "red", linetype = "dashed")
```

```
## Warning: Removed 59 rows containing missing values (geom_path).
```




```
ideal_n <- max(results$n[which(results$power<=0.9)]) # using z-test; though this is not our recommendat
```

The results of this power analysis show that the needed sample size is 34 assuming we wanted to use the z-statistic to calculate power.

We plan to conduct a simulation study to find a sample size suited for a power of 0.9 using the Welch t-test so that our test does not rely on equal variances. Additionally, this allows our sample sizes to vary if any errors occur in sampling as well.

Then we would conduct a simulation study to assess the performance of the Welch t-test using simulated data from t distributions using the sample mean and variance of the 100 degrees data. We could also introduce variations to our sample sizes and variances to test the resilience of our experiment structure. The reliance would be evaluated by getting a balance between type one error probably (about 0.05) while maintaining the power above 0.9.

Question 2

Data set: 'fev.csv'

The data come from a study to examine, among other things, the association between smoking and lung function as measured by forced expiratory volume (FEV) in children. Variables in the data set are listed below.

id: ID number age: Age (yrs) fev: FEV (liters) ht: Height (inches) male: Sex (0=female, 1=male) smoke: Smoking Status (0=non-current smoker, 1=current smoker)

```
fevDat <- read.csv("../WEEK03/fev.csv")
smokerDat <- fevDat[which(fevDat$smoke==1),]
nonSmokerDat <- fevDat[which(fevDat$smoke==0),]
```

2.1. It is desired to conduct a test of the null hypothesis that mean FEV is the same in smokers and non-smokers. Conduct an appropriate simulation study to determine an appropriate hypothesis test procedure to use (consider only the 3 methods discussed in class). Provide a description of your simuation study in words and provide a summary of the results of your simulation study.

$H_0 : \mu_A = \mu_B$ where A are smokers and B are non-smokers $H_1 : \mu_A \neq \mu_B$

In order to determine an appropriate hypothesis test procedure, we must consider how large the sample size is,

```
# N<-5000
# set.seed(20190126)
# n1<-30
# n2<-15
# n<-n1+n2
# sim_data<-data.frame(temp=c(rep(50,n1),rep(100,n2)),output=rep(NA,n))
# sim_result<- data.frame(p.equal.var=rep(NA,N),p.Welch=rep(NA,N),p.Z=rep(NA,N))
# for(i in 1:N){
#     sim_data$output<-rnorm(n,mean=1000,sd=c(rep(80,n1),rep(20,n2)))
#     result<-with(sim_data,t.test(output[temp==50],output[temp==100],var.equal=T))
#     sim_result[i,1]<-result$p.value
#     result<-with(sim_data,t.test(output[temp==50],output[temp==100],var.equal=F))
#     sim_result[i,2]<-result$p.value
#     Z<-result$statistic
#     sim_result[i,3]<-2*(1-pnorm(Z))
# }
# apply(sim_result<0.05,2,mean)
```

2.2. Using the method chosen in Q2.1, carry out the test of the null hypothesis. Report your results along with a confidence interval.

2.3. Conduct a test of the null hypothesis that mean FEV is the same in smokers and non-smokers, among children 10 years old or older. Compare your results to the previous results for all children and provide an interpretation of any differences in results. (Note: Children younger than 10 would be unlikely to smoke and had been included in the study for other purposes besides the analysis of the association between smoking and FEV.)

Question 3

Read the paper "Why most published research findings are false." (see Week 3 folder). Do you agree with the overall conclusion of the paper that most published research findings are false? Provide your answer with your rationale in a paragraph of 200 words or less. (No simulation studies required here!)