Interpretation of p-values

Chapter 5, Lab 5: Solutions

OpenIntro Biostatistics

Suppose you read a press report in a reliable publication about two recently reported clinical trials. One describes an advance in a disease in which there has been substantial progress in the last 5 years, such as the early stage breast cancer, using a treatment regimen that is a modification of something that has been successful. The second summarizes a potentially important advance in a disease where recent studies have not led to any progress, such as Alzheimer's disease. Both studies were well-designed randomized trials with statistically significant results ($p < 0.05$), when the new treatments were compared with currently used approaches.

A friend tells you she is more skeptical about the results of the second study but believes the first result. You argue, on the other hand that the p-value for the trial is 0.04, so there must be a 96% chance that the drug is effective. Is your friend's intuition correct, or does your claim about the p-value correctly refute her claim? This lab explores how one might quantify that skepticism. It begins by reviewing the concept of type I error and power.

Introduction

Because of the aging population in the United States and other high income countries, Alzheimer's disease and its potential treatment are active areas of research. The Dementia Severity Rating Scale (DSRS)[1] is a questionnaire that is completed by a knowledgable informant (typically a spouse or other close relative), ranking in 12 domains the impairment severity of a person with Alzheimer's disease. The total scores range from 0 (no impairment) to 54 (extreme impairment). Cognitive decline over several years is measured by the annual rate of change: a patient scored for 3 consecutive years whose score increased from 7 to 14.5 has an annual rate of change of 7.5/3 = 2.5 points per year. Negative rates of change mean that an individual's score decreased, which would be an improvement from baseline.

Suppose a pharmaceutical company has developed a drug that may arrest or slow cognitive decline in Alzheimer's disease. In a hypothetical randomized trial, the company will compare the average rate of change of the DSRS for participants given a placebo with those given the new drug. Enrolled participants will be randomized to either placebo or the new drug. DSRS will be measured in both groups at randomization and 3 years later.

1. Suppose the company scientists have decided that a 1.0 difference between the treatment groups in the average rate will be sufficient to study the drug further. Newly diagnosed patients have an average score of about 20, and typically decline at the rate of 3.5 points a year. The standard deviation of the rate of decline can be assumed to be 6 points per year. How large should the study be to have 80% power to reject the null hypothesis of no difference in the rate of decline between the two groups in favor of the alternative that there is a difference if the new treatment reduces the rate of decline by 1 point per year? Assume

[1] Clark CM, Ewbank DC. Performance of the dementia severity rating scale: a caregiver questionnaire for rating severity in Alzheimer disease. Alzheimer Dis Assoc Disord. 1996;10(1):31–9.

the study will be analyzed with a two-sample, two-sided t-test, comparing the average rate of decline in the two groups. Use the R function `power.t.test` to compute the sample size.

```
change = 1.0
sd.change = 6.0
power.t.test(delta = change, sd = sd.change, power = 0.80, type = "two.sample",
             alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 566.08
##           delta = 1
##              sd = 6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The study should have 567 participants per group, or a total of 1134 patients.

2. When the two code chunks below have been completed, they will produce simulations and analyses under the null hypothesis of no difference and under an alternative hypothesis.

a. Complete the code given below to simulate and analyze one instance of the trial, assuming that participants in the both the treatment and control groups experience an average rate of decline of 3.5 points per year, i.e., that the null hypothesis of no treatment effect is the true state of nature. Assume that the standard deviation of the rate of change is 6 points per year. You may assume that the average rate of change is a normally distributed random variable. Discuss what you see.

```
set.seed(2018)
n.control = 567
n.treatment = 567
control.mean = 3.5
treatment.mean = 3.5
rate.sd = 6

control.rate = rnorm(n.control, mean = control.mean, sd = rate.sd)
treatment.rate = rnorm(n.treatment, mean = treatment.mean, sd = rate.sd)

t.test(control.rate, treatment.rate, alternative = "two.sided", mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  control.rate and treatment.rate
## t = -0.691, df = 1130, p-value = 0.49
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.95959  0.45963
## sample estimates:
## mean of x mean of y
##     3.328     3.578
```

The two observed average rates of change in the control and treatment groups, respectively, are 3.23 and 3.58, and the 95% confidence interval for the difference between the groups is (-9.96, 0.46). Since the confidence interval contains the null hypothesis value 0 and the p-value for the two-sample t-statistic is 0.49, the null hypothesis is not rejected. Even though the simulation specified equal standard deviations for the two groups, it is generally advisable to use the t-test that accommodates unequal standard deviations, since in a real study, the data analyst would not know the standard deviations for the two groups.

b. Complete the code given below to simulate and analyze one instance of the trial, assuming that participants in the treatment group decline on average 2.5 points per year and that the control group participants and control groups experience an average rate of decline of 3.5 points per year, i.e., that the null hypothesis of no treatment effect is not true. Assume that the standard deviation of the rate of change is 6 points per year and that the average rate of change per individual is a normally distributed random variable. You may assume that the average rate of change is a normally distributed random variable. Discuss what you see.

```
set.seed(2018)
n.control = 567
n.treatment = 567
control.mean = 3.5
treatment.mean = 2.5
rate.sd = 6

control.rate = rnorm(n.control, mean = control.mean, sd = rate.sd)
treatment.rate = rnorm(n.treatment, mean = treatment.mean, sd = rate.sd)

t.test(control.rate, treatment.rate, alternative = "two.sided", mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  control.rate and treatment.rate
## t = 2.07, df = 1130, p-value = 0.038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.040408 1.459629
## sample estimates:
## mean of x mean of y
##     3.328     2.578
```

The two average rates of change in the control and treatment groups, respectively, are 3.33 and 2.58. so the treatment group increased its score at a slower rate that the control; the difference between the average rates was 0.75 points per year. The 95% confidence interval for the difference between the groups is (0.04, 1.46). Since the confidence interval does not contain the null hypothesis value 0 and the p-value for the two-sample t-statistic is 0.04, the null hypothesis is rejected in favor of

the alternative.

3. This problem uses a simulation to confirm that the trial discussed in problems 1 and 2 does in fact have the correct type 1 error probability (0.05) and power (0.80). These simulations are similar to earlier lab examples.

a. Replicate the trial 10,000 times under the null hypothesis, using the same parameters for the trial as were used in Exercise 2a. How often in the 10,000 replicates did the test reject the null hypothesis. Is this result consistent with the design?

```
set.seed(2018)
num.iterations = 10000
alpha = 0.05

n.control = 567
n.treatment = 567
control.mean = 3.5
treatment.mean = 3.5
rate.sd = 6
p.vals = vector("numeric", num.iterations)

for(k in 1:num.iterations){

    control.rate = rnorm(n = n.control, mean = control.mean, sd = rate.sd)
    treatment.rate = rnorm(n = n.treatment, mean = treatment.mean, sd = rate.sd)

    p.vals[k] = t.test(control.rate, treatment.rate, alternative = "two.sided",
                       var.equal = TRUE, mu = 0, conf.level = 1 - alpha)$p.val
}

#logical vector for whether a test accepts or rejects
reject = (p.vals <= alpha)

#table of results
addmargins(table(reject))
```

```
## reject
## FALSE  TRUE   Sum
##  9513   487 10000
```

b Replicate the trial 10,000 times under the alternative hypothesis that was used for the power calculations, this time using same parameters for the trial as were used in Exercise 2b. How often in the 10,000 replicates did the test reject the null hypothesis. Is this result consistent with the design?

Replicate the trial 10,000 under the alternative hypothesis that was used for the power calculations.

```
set.seed(2018)
num.iterations = 10000
alpha = 0.05
```

```
n.control = 567
n.treatment = 567
control.mean = 3.5
treatment.mean = 2.5
rate.sd = 6
p.vals = vector("numeric", num.iterations)

for(k in 1:num.iterations){

    control.rate = rnorm(n = n.control, mean = control.mean, sd = rate.sd)
    treatment.rate = rnorm(n = n.treatment, mean = treatment.mean, sd = rate.sd)

    p.vals[k] = t.test(control.rate, treatment.rate, alternative = "two.sided",
                       var.equal = TRUE, mu = 0, conf.level = 1 - alpha)$p.val
}

#logical vector for whether a test accepts or rejects
reject = (p.vals <= alpha)

#table of results
addmargins(table(reject))
```

```
## reject
## FALSE  TRUE   Sum
##  2053  7947 10000
```

4. People who are skeptical about the results of a trial usually either question the rigor of the trial, or think that a well done trial may have yielded a false positive result and so may not be reproducible. The latter is based on an assumption that is different than the one used in a traditional design. The skeptic feels that in a large set of potential treatments to study in a disease and trials that could be done, a relatively small number will uncover an advance and the rest would be negative. That assumption might seem reasonable if most trials of potentially new drugs in a given disease area have failed to find treatment differences. Alzheimer's disease is one such area.

Interpreting the results of a trial in the context of an intuition that most trials would be expected to be negative is a broader view of the research enterprise in medicine or science than this text has discussed. But it is not difficult to explore to make that notion more precise.

Suppose one believes that there is at most 10% chance that a drug being studied for Alzheimer's disease will slow the rate of increase in the DSRS. With this view, conducting a clinical trial among the set of all trials in a disease might be regarded as a two-stage sampling exercise. Investigators choose a drug to explore, then sample from a set of hypothetical experiments in which the null hypothesis has a 90% of being true, and the alternative a 10% of being true. This view posits prior probabilities on the null and alternative hypotheses. Having seen that a trial has a positive result, the skeptic contends that the result should be interpreted acknowledging that the prior chance of the result being true is 1 in 10, and by updating the prior probability of 0.10 that the alternative is true. The updated probability for the alternative hypothesis is called a posterior probability.

These prior probabilities should not be confused with the power and type I error of a trial. The power of a trial is the probability that the null hypothesis is rejected, given that the alternative hypothesis is true; the posterior probability for an alternative is the probability that the alternative is true, given that the null hypothesis is true. As the language suggests, these notions can be formalized using Bayes' rule. This exercise instead uses a simulation to estimate a posterior probability for an alternative hypothesis.

The simulation is based on the hypothetical Alzheimer's disease trial used in the earlier exercise. The parameters p.alternative and p.null are, respectively, the prior probabilities of the alternative and null hypotheses.

Use the two-way table produced by the simulation to examine the type I error probability and power (these should match the design) and the posterior probabilities of the null and alternative hypotheses. Comment on what you find.

```
#define parameters
num.iterations = 10000
p.alternative = 0.10
p.null = 1 - p.alternative
n.control = 567
n.treatment = 567
alpha = 0.05

#create empty vectors to store results
hypothesis = vector("numeric", num.iterations)
p.vals = vector("numeric", num.iterations)

control.mean.null = 3.5
treatment.mean.null = 3.5

control.mean.alternative = 3.5
treatment.mean.alternative = 2.5

rate.sd = 6

#set the seed
set.seed(2018)

#assign state of nature
hypothesis = sample(c("null", "alternative"), size = num.iterations,
                    prob = c(1 - p.alternative, p.alternative),
                    replace = TRUE)

#simulate data and p-values
for(k in 1:num.iterations){

  if(hypothesis[k] == "null"){

    control.rate = rnorm(n.control, mean = control.mean.null, sd = rate.sd)
```

```
    treatment.rate = rnorm(n.treatment, mean = treatment.mean.null, sd = rate.sd)

    p.vals[k] = t.test(control.rate, treatment.rate, alternative = "two.sided",
                       mu = 0, conf.level = 1 - alpha)$p.val

  }

  if(hypothesis[k] == "alternative"){

    control.rate = rnorm(n = n.control, mean = control.mean.alternative, sd = rate.sd)
    treatment.rate = rnorm(n = n.treatment, mean = treatment.mean, sd = rate.sd)

    p.vals[k] = t.test(control.rate, treatment.rate, alternative = "two.sided",
                        mu = 0, conf.level = 1 - alpha)$p.val

  }

}

#logical vector for whether a test accepts or rejects
reject = (p.vals <= alpha)

#table of results
addmargins(table(hypothesis, reject))
```

```
##              reject
## hypothesis    FALSE  TRUE   Sum
##   alternative   173   752   925
##   null         8591   484  9075
##   Sum          8764  1236 10000
```

An additional problem trying different values of prior probabilities.