A Closer Look at the P-Value

Chapter 5, Lab 5: Solutions

OpenIntro Biostatistics

This lab uses simulation and the principles of conditional probability to highlight some specific misconceptions about p-values. It can also be viewed as an informal introduction to the paradigm of Bayesian inference.

The ASA Statement on Statistical Significance and P-Values

In 2016, the American Statistical Association released a formal statement clarifying principles about the proper use and interpretation of the p-value, with the aim of improving the way research is conducted in the scientific community.[1] The main points of the statement are reproduced below.

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. Many other contextual factors should be considered, such as the design of a study, the external evidence for the phenomenon under study, and the validity of the assumptions that underlie the data analysis.
4. Proper inference requires full reporting and transparency. Conducting multiple analyses and reporting only those with certain p-values compromises interpretation of the results.
5. A p-value does not measure the size of an effect or the importance of a result. Statistical significance is not equivalent to scientific, human, or economic significance.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The second point touches on an especially common misconception about p-values—that if a p-value equals, say, 0.04, the null hypothesis has only a 4% chance of being true (or that the alternative hypothesis has a 96% chance of being true). This misconception will be explored in the lab, along with an approach for quantifying external evidence for a phenomenon (as mentioned in the third point).

Introduction

Treatment for Alzheimer's disease is an active research area because of the aging population in the United States and other high-income countries. The Dementia Severity Rating Scale (DSRS) is a questionnaire completed by a knowledgeable informant (typically a spouse or other close relative) to measure the impairment severity of a person with Alzheimer's disease.[2] Scores on the DSRS range from 0 (indicating no impairment) to 54 (extreme impairment).

[1] Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

[2] Clark C.M., Ewbank D.C. (1996). Performance of the dementia severity rating scale: a caregiver questionnaire for rating severity in Alzheimer disease. *Alzheimer Disease and Associated Disorders*, *10*(1),31-9.

Cognitive decline over several years is measured by the annual rate in change in DSRS score. For example, a patient scored for three 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. A negative rate of change is indicative of improvement from baseline.

Suppose a pharmaceutical company has developed a drug to slow cognitive decline in individuals with Alzheimer's disease. In a randomized trial, the company will compare mean annual rate of change in DSRS score for participants receiving a placebo to mean annual rate of change for participants treated with the new drug. The study will enroll individuals newly diagnosed with Alzheimer's and randomize each participant to either the control group or the treatment group, with DSRS being measured for each participant at the beginning of the study and three years later.

1. The company scientists have decided that a 1.0 point difference in mean rate of change in DSRS score between the groups will be sufficient to warrant further study of the drug. Individuals newly diagnosed with Alzheimer's have an average DSRS score of about 20 and typically experience cognitive decline at the rate of +3.5 points per year. The standard deviation of rate of decline can be assumed to be 6 points per year. The study results will be analyzed with a two-sided t-test conducted at $\alpha = 0.05$.

   How many participants per treatment group will be necessary for the study to have 80% power?

   The study should have 567 participants per treatment group to have 80% power.

```
power.t.test(n = NULL, delta = 1.0, sd = 6.0,
         sig.level = 0.05, power = 0.80,
         type = "two.sample",
         alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 566.0813
##           delta = 1
##              sd = 6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

2. Suppose that the null hypothesis of no difference between treatment groups is true, and that the mean rate of change in DSRS score in both groups is 3.5 points/year, with standard deviation of 6 points. Assume that average rate of change is normally distributed.

   a) Run the following code to simulate rate of change in DSRS score for 567 individuals in the control group and 567 individuals in the treatment group, stored in the vectors `control` and `treatment`.

```
#set the parameters
control.mean = 3.5
treatment.mean = 3.5
```

```r
control.sigma = treatment.sigma = 6
control.n = treatment.n = 567

alpha = 0.05

#set seed
set.seed(2018)

#simulate data
control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = -0.69119, df = 1132, p-value = 0.4896
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9595920  0.4596292
## sample estimates:
## mean of x mean of y
##  3.328011  3.577993
```

b) Conduct a two-sided test of the null hypothesis of no difference from the simulated data.

i. Summarize your conclusions.

The p-value is 0.49; there is insufficient to reject the null hypothesis in favor of the alternative. An observed difference in means of 0.25 points/year between the groups is consistent with the null hypothesis of no difference in population mean rate of change in DSRS score.

ii. Does this represent an instance of Type I error?

A Type I error is rejecting the null hypothesis when the null hypothesis is true. Since this test is an instance of correctly failing to reject when the null hypothesis is true, it does not represent an instance of Type I error.

iii. Based on the study design, what is the probability of making a Type I error?

The probability of making a Type I error is α, which is set at 0.05 for this study.

3. Now, suppose that the alternative hypothesis of a difference between treatment groups is true, and that the mean rate of change in DSRS score in the treatment group is 2.5 points/year while mean rate of change in the control group is 3.5 points/year.

a) Run the following code to simulate rate of change in DSRS score for 567 individuals in the control group and 567 individuals in the treatment group, stored in the vectors `control` and `treatment`.

```r
#set the parameters
control.mean = 3.5
treatment.mean = 2.5
control.sigma = treatment.sigma = 6
control.n = treatment.n = 567

alpha = 0.05

#set seed
set.seed(2018)

#simulate data
control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = 2.0738, df = 1132, p-value = 0.03832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04040796 1.45962920
## sample estimates:
## mean of x mean of y
##  3.328011  2.577993
```

b) Conduct a two-sided test of the null hypothesis of no difference from the simulated data.

   i. Summarize your conclusions.

   The p-value is 0.04; there is sufficient to reject the null hypothesis in favor of the alternative at the $\alpha = 0.05$ significance level. An observed difference in means of 0.75 points/year between the groups is consistent with the alternative hypothesis that there is a difference in population mean rate of change in DSRS score.

   ii. Does this represent an instance of Type II error?

   A Type II error is failing to reject the null hypothesis when the alternative hypothesis is true. Since this test is an instance of correctly rejecting the null hypothesis when the alternative hypothesis is true, it does not represent an instance of Type II error.

   iii. Based on the study design, what is the probability of making a Type II error?

The probability of making a Type II error is $1 - 0.80 = 0.20$; the event of failing to reject the null when the alternative is true is the complement of rejecting the null when the alternative is true. Recall that power is defined as the probability of correctly rejecting the null hypothesis.

Applying Bayes' Rule to Hypothesis Testing

4. In reality, it is not possible to know whether a Type I or Type II error has been made.

   a) Suppose the scientists conduct the trial as specified previously, and the two-sided test of a null of no difference in mean rate of annual change produces a p-value of 0.49. Is it necessarily the case that the null hypothesis is true (and the alternative hypothesis is false)? Explain your answer.

   No, a p-value larger than α does not necessarily indicate that the null hypothesis is true and the alternative hypothesis is false. A p-value larger than α represents data that are relatively likely under the null distribution and relatively unlikely under the alternative distribution. However, it may be the case that this result represents an instance of Type II error: failing to reject H_0 when H_A is true.

   b) Suppose the scientists conduct the trial as specified previously, and the two-sided test of a null of no difference in mean rate of annual change produces a p-value of 0.04. Is it necessarily the case that the null hypothesis is false (and the alternative hypothesis is true)? Explain your answer.

   No, a p-value smaller than α does not necessarily indicate that the null hypothesis is false and the alternative hypothesis is true. A p-value smaller than α represents data that are relatively unlikely under the null distribution and relatively likely under the alternative distribution. However, it may be the case that this result represents an instance of Type I error: rejecting H_0 when H_0 is true.

5. Now, consider a new piece of information. Very few helpful therapies for Alzheimer's disease have been identified; four drugs are currently approved and used for symptomatic treatment, but over 100 tested as potential therapies have either been abandoned in development or failed in clinical trials.[3] Most recently, in March 2019, pharmaceutical company Biogen halted two phase-three trials of the drug aducanumab after it was determined that the drug was unlikely to provide benefit compared to a placebo.

   a) As in the previous question, suppose that testing for a difference in mean rate of annual change in DSRS score results in a p-value of 0.49. With the information about the outcomes of previous potential therapies for Alzheimer's in mind, are you inclined to believe that this outcome is an instance of Type II error? Explain your answer.

   No, it seems unlikely that this outcome is an instance of Type II error. Since many previous drugs for Alzheimer's that seemed promising initially ended up failing, it is plausible that, in the language of diagnostic testing, this "negative" outcome represents a true negative rather than a false negative (Type II error).

[3] Mehta, D., et al. (2017). Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert opinion on investigational drugs*, *26*(6), 735-739.

b) As in the previous question, suppose that testing for a difference in mean rate of annual change in DSRS score results in a p-value of 0.04. With the information about the outcomes of previous potential therapies for Alzheimer's in mind, are you inclined to believe that this outcome is an instance of Type I error? Explain your answer.

Yes, it seems likely that this outcome is an instance of Type I error. Since many previous drugs for Alzheimer's that seemed promising initially ended up failing, it is reasonable to be skeptical of this "positive" outcome and think that it seems more likely to be a false positive result (Type I error) than a true positive.

We can quantify skepticism about whether a new drug is effective by assigning probabilities to the null and alternative hypotheses. Suppose that prior to seeing the observed data, we believe that there is a 90% chance the drug is not effective and a 10% chance the drug is effective. Let

$$P(H_0) = 0.90 \text{ and } P(H_A) = 0.10.$$

Once the data are observed, update belief about the hypotheses according to Bayes' Theorem:

$$P(H_0|\text{data}) = \frac{P(\text{data}|H_0)P(H_0)}{P(\text{data}|H_0)P(H_0) + P(\text{data}|H_A)P(H_A)}$$

$$P(H_A|\text{data}) = \frac{P(\text{data}|H_A)P(H_A)}{P(\text{data}|H_0)P(H_0) + P(\text{data}|H_A)P(H_A)}$$

The probabilities for the hypotheses conditional on the data are referred to as **posterior probabilities**, in the language of Bayesian inference. These are distinct from the unconditional **prior probabilities**, $P(H_0)$ and $P(H_A)$.

6. Run the following code to conduct a simulation that estimates $P(H_0|\text{data})$ and $P(H_A|\text{data})$. Note that "data" is essentially a placeholder term for whether a particular set of observed data results in rejecting or failing to reject H_0.

   For each iteration, data are simulated according to either the null distribution or the alternative distribution. With each iteration, the code draws a new set of control and treatment values, conducts the two-sample t-test, and records the p-value. The logical vector `reject` records whether the p-value for a particular iteration was significant at α (i.e., less than or equal to α).

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

control.mean.null = treatment.mean.null = 3.5
control.mean.alternative = 3.5
treatment.mean.alternative = 2.5
control.sigma = treatment.sigma = 6

#set the seed
set.seed(2018)

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

#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 = rnorm(n.control, mean = control.mean.null, sd = control.sigma)
treatment = rnorm(n.treatment, mean = treatment.mean.null, sd = treatment.sigma)

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

  }

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

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

p.vals[k] = t.test(control, treatment, 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
```

The results of the simulation are in the form of a two-way table; each of the 10,000 t-tests (and sets of control and treatment observations) are classified based on which state of nature (i.e., hypothesis) they occurred under and whether they resulted in rejecting H_0.

a) From the results, check that the estimates of $P(H_0)$ and $P(H_A)$ are reasonably close to 0.90 and 0.10, respectively, as specified in the parameters.

   The estimates of $P(H_0)$ and $P(H_A)$ are calculated from the row totals: $925/10000 = 0.093$ and $9075/10000 = 0.908$. These marginal probabilities are reasonably close to $P(H_0)$ and $P(H_A)$ that were specified in the parameters of 0.10 and 0.90, respectively.

```
#use r as a calculator
925/10000
```

```
## [1] 0.0925
```

```
9075/10000
```

```
## [1] 0.9075
```

b) From the results, check that the estimates of power and α are reasonably close to 0.80 and 0.05, respectively, as specified in the parameters.[4]

   Recall that power is the probability of rejecting H_0 when H_A is true. In probability notation, power is $P(\text{reject } H_0|H_A)$; from the results, this is estimated as $752/925 = 0.813$, which is reasonably close to 0.80.

   Recall that α is the probability of rejecting H_0 when H_0 is true. In probability notation, α is $P(\text{rejecting } H_0|H_0)$; from the results, this is estimated as $484/9075 = 0.053$, which is reasonably close to 0.05.

```
#use r as a calculator
752/925
```

```
## [1] 0.812973
```

```
484/9075
```

```
## [1] 0.05333333
```

[4] Power is not explicitly specified in the parameters, but Question 1 determined that a sample size of 567 individuals per group is sufficient for 80% power (for the previously stated α, effect size, and standard deviation).

c) Estimate the probability that the null hypothesis is true, given that we fail to reject H_0. This is analogous to estimating the probability of a "true negative" result, in the language of diagnostic testing.

The total number of times we fail to reject H_0 is 8764; of these, 8591 represent correctly failing to reject H_0, i.e., failing to reject when the state of nature is indeed H_0. Thus, the estimated probability of a true negative is $8591/8764 = 0.98$.

```
#use r as a calculator
8591/8764
```

```
## [1] 0.9802602
```

d) Estimate the probability that the alternative hypothesis is true, given that we reject H_0. This is analogous to estimating the probability of a "true positive" result, in the language of diagnostic testing.

The total number of times we reject H_0 is 1236; of these, 752 represent correctly rejecting H_0, i.e., rejecting when the state of nature is H_A. Thus, the estimated probability of a true positive is $752/1236 = 0.608$.

```
#use r as a calculator
752/1236
```

```
## [1] 0.6084142
```

e) How would you expect the probabilities estimated in parts c) and d) to change if $P(H_A)$ were greater than 0.10, or less than 0.10?

As $P(H_A)$ increases from 0.10, the probability of a true positive result increases while the probability of a true negative result decreases; as the alternative hypothesis becomes more plausible, it becomes more reasonable to expect that rejecting H_0 is a correct rejection rather than a Type I error. Following similar logic, as $P(H_A)$ decreases from 0.10, the probability of a true positive result decreases while the probability of a true negative result increases.

In the diagnostic testing paradigm, $P(H_A)$ is analogous to $P(D)$. As the prevalence of a disease increases, the probability of a true positive test result ($P(D|T^+)$) increases and the probability of a true negative result ($P(D^C|T^-)$) decreases.

7. Address the misconception mentioned at the beginning of the lab—what is the logical flaw in concluding that if a p-value equals, say, 0.04, the null hypothesis has only a 4% chance of being true (or that the alternative hypothesis has a 96% chance of being true)?

The p-value is a conditional probability, calculated under the the assumption that the null hypothesis is true. It is not equivalent to either the marginal probability that the null hypothesis is true, or the conditional probability that the null hypothesis is true given that the test rejects the null hypothesis.