Statistical Power

Chapter 5, Lab 2: Solutions

OpenIntro Biostatistics

Topics

- Controlling Type I error
- Controlling Type II error
- Power and sample size calculations

Most studies are done to establish evidence in favor of an alternative hypothesis. The **power** of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. Power depends on the hypothesized difference between two population means ($|\mu_1 - \mu_2|$), the population standard deviation of each group (σ_1, σ_2), and the sample size of each group (n_1, n_2). Usually, a study team can only control sample size.

The power of a test can be expressed as $P(\text{reject } H_0 | H_A \text{ is true}) = 1 - \beta$, where β is the probability of making a Type II error (failing to reject H_0 when H_A is true). Recall that the significance level of a test, α, is the probability of making a Type I error (rejecting H_0 when H_0 is true).

	Result of test	
State of nature	**Reject H_0**	**Fail to reject H_0**
H_0 is true	Type I error, probability = α (false positive)	No error, probability = $1 - \alpha$ (true positive)
H_A is true	No error, probability = $1 - \beta$ (true positive)	Type II error, probability = β (false negative)

This lab uses simulation to explore how Type I and Type II error are controlled, and examines factors influencing the power of a statistical test. The last section introduces the formulas for power and sample size calculations in the two-group setting.

The material in this lab corresponds to Section 5.4 of *OpenIntro Biostatistics*.

Introduction

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure and is planning a clinical trial to test the drug's effectiveness. Participants are randomized to one of two treatments, either a currently accepted medication or the new drug. At the end of the study, a hypothesis test will be conducted to assess whether there is evidence that the new drug performs better than the standard medication.

The following sections examine simulations performed under two scenarios: 1) the null hypothesis is true, and there is no difference in population mean blood pressure between the two groups, or

2) the alternative hypothesis is true, and there is a difference in population mean blood pressure between the two groups. For clinical trial data, it is standard practice to test the two-sided alternative.

Note that for this setting, there is no actual population of individuals taking the new drug (since the drug is not yet available on the market). Regardless, the observations on the participants assigned to take the new drug are treated as if they are a random sample from a hypothetical population.

Controlling Type I error

Suppose that the null hypothesis $H_0 : \mu_{treatment} = \mu_{control}$ is true. Let the mean systolic blood pressure in both groups be 140 mm Hg, with standard deviation of 10 mm Hg; assume that blood pressures are normally distributed.

1. Run the following code to simulate blood pressure values for 50 individuals in the control group and 50 individuals in the treatment group, stored in the vectors `control` and `treatment`.

   The `rnorm()` function draws `n` random numbers from a normal distribution with a given mean and standard deviation.

```r
#set the parameters
control.mean = 140
treatment.mean = 140
control.sigma = treatment.sigma = 10
control.n = treatment.n = 50

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,
   conf.level = 1 - alpha)
```

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = -1.0019, df = 97.422, p-value = 0.3189
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.05381  1.99209
## sample estimates:
## mean of x mean of y
```

```
##  139.1965  141.2273
```

a) Calculate $\bar{x}$ and s for `control` and `treatment` to confirm that the simulated values seem plausible given the specified parameters for μ and σ. Would you expect $\bar{x}$ and s to be exactly the same as the parameter values? Why or why not?

For the control group, $\bar{x} = 139.20$, $s = 9.74$. For the treatment group, $\bar{x} = 141.23$, $s = 10.52$. The sample summary statistics are quite similar to the parameter values, as expected; since only 50 values are drawn from each distribution, it is to be expected that the sample summary statistics are not exactly equal to the parameter values.

```
#calculate x bar and s for control
mean(control)
```

```
## [1] 139.1965
```

```
sd(control)
```

```
## [1] 9.737066
```

```
#calculate x bar and s for treatment
mean(treatment)
```

```
## [1] 141.2273
```

```
sd(treatment)
```

```
## [1] 10.51855
```

b) Using `t.test()`, conduct a two-sided test of the null hypothesis from the simulated data. Summarize your conclusions.

The p-value is 0.32, which is greater than $\alpha = 0.05$. There is insufficient evidence to reject the null hypothesis that the mean blood pressure in the treatment and control groups are the same.

2. Run the following code to repeat the simulation 1,000 times. 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).

```
#set parameters
control.mean = 140
treatment.mean = 140
control.sigma = treatment.sigma = 10
control.n = treatment.n = 50

alpha = 0.05
replicates = 1000

#set seed
set.seed(2018)

#create empty list
```

```r
p.values = vector("numeric", replicates)

#run simulations
for (k in 1:replicates){
  control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
  treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

#view results
reject = (p.values <= alpha)
table(reject)
```

```
## reject
## FALSE  TRUE
##   947    53
```

a) With 50 individuals in each group, what percentage of tests result in the (incorrect) conclusion that the two population means are different?

With 50 individuals in each group, 53/1000 = 0.053 of the tests result in the incorrect conclusion that the two population means are different.

b) Does Type I error rate change with sample size? Modify the simulation code to assess sample sizes of 100, 1,000, and 10,000.

The percentage of incorrect rejections remains constant, around the α level of 0.05. Increasing sample size does not result in a clear upwards or downwards trend. Type I error does not change with sample size; Type I error is completely determined by the choice of α, where H_0 is only rejected when a p-value is smaller than α.

```
## reject
## FALSE  TRUE
##   944    56
```

```
## reject
## FALSE  TRUE
##   956    44
```

```
## reject
## FALSE  TRUE
##   951    49
```

Controlling Type II error

Now, suppose that the alternative hypothesis $H_A : \mu_{treatment} \neq \mu_{control}$ is true. Let the mean systolic blood pressure be 140 mm Hg in the control group and 138 mm Hg in the treatment group.

3. Run the code shown in the template to simulate blood pressure values for 25 individuals in the control group and 25 individuals in the treatment group.

```
#set the parameters
control.mean = 140
treatment.mean = 138
control.sigma = treatment.sigma = 10
control.n = treatment.n = 25

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,
   conf.level = 1 - alpha)
```

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = 1.1572, df = 46.218, p-value = 0.2531
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.375592  8.802923
## sample estimates:
## mean of x mean of y
##  139.8033  136.5897
```

   a) Conduct a two-sided test of the null hypothesis from the simulated data. What is the conclusion of the test?

Since $p = 0.25$, which is greater than α, there is insufficient evidence to reject the null hypothesis.

   b) Is the conclusion from part a) correct?

The conclusion from part a) is not correct, since the observations are drawn from normal distributions with different means: 140 mmHg in the control population and 138 mmHg in the treatment population.

4. *Power and Sample Size.* Run the code chunk shown in the template to repeat the simulation 1,000 times.

```
#set parameters
control.mean = 140
treatment.mean = 138
control.sigma = treatment.sigma = 10
control.n = treatment.n = 25

alpha = 0.05
replicates = 1000

#set seed
set.seed(2018)

#create empty list
p.values = vector("numeric", replicates)

#run repeated t-tests and record p-values
for (k in 1:replicates){
  control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
  treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

reject = (p.values <= alpha)
table(reject)
```

```
## reject
## FALSE  TRUE
##   885   115
```

a) With 25 individuals in each group, how many tests result in the incorrect conclusion that the two population means are not different?

With 25 individuals in each group, 115 of the tests result in the correct conclusion that the two population means are different.

b) Estimate the power of the two-sample test when each group has $n = 25$, $\sigma = 10$, and $\mu_{treatment} - \mu_{control} = -2$ mm Hg. Recall that power refers to the probability of rejecting H_0 when H_A is true.

The power of the test is estimated to be 115/1000 = 0.115.

c) How does power change with increasing sample size? Estimate the power of the test as n changes to 50, 100, 200, and 300 (leaving all other parameters the same).

As sample size increases from 50 to 300 (per group), power increases to about 15.5%, 30.8%, 50.2%, and 66.7%.

```
## reject
## FALSE  TRUE
##   845   155

## reject
## FALSE  TRUE
##   692   308

## reject
## FALSE  TRUE
##   498   502

## reject
## FALSE  TRUE
##   333   667
```

d) Run the code chunk shown in the template to create a plot of power against sample size. The code uses the command `power.t.test()` to calculate power; this command will be discussed later in the lab. Does power seem linear in relation to sample size?

Power is not linear with relation to sample size. Power increases rapidly up until around sample size of 125, then levels off.

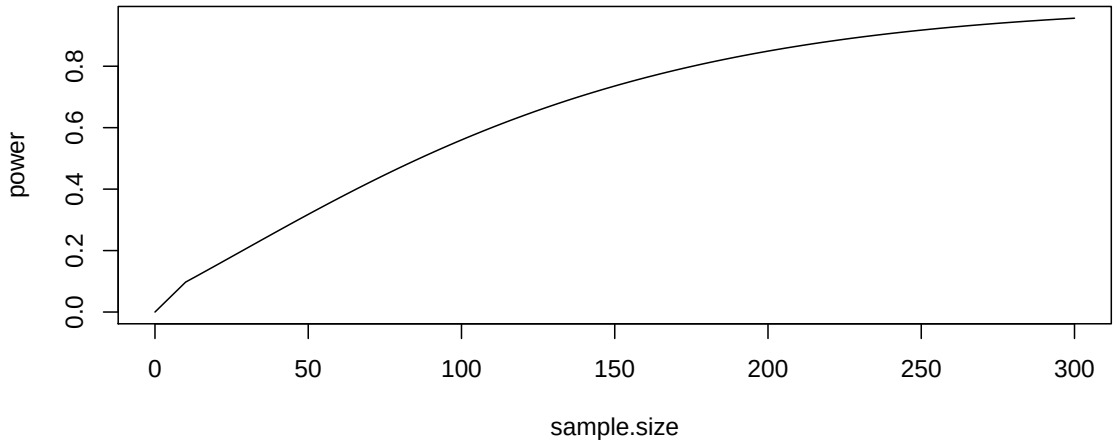
```
#set parameters
min.n = 10
max.n = 300

sample.size.range = min.n:max.n

#create empty lists
power = vector("numeric", max.n - min.n + 1)
sample.size = vector("numeric", max.n - min.n + 1)

#calculate power
for (k in sample.size.range){
  sample.size[k] = k
  power[k] = power.t.test(n = k, delta = 3, sd = treatment.sigma, type = "two.sample",
           alternative = "two.sided", strict = TRUE)$power
}

#create plot
plot(power ~ sample.size, type = "l")
```



5. *Power and Standard Deviation*. For simplicity, this simulation assumes that the standard deviation of the treatment and control groups are equal.

```r
#set parameters
control.mean = 140
treatment.mean = 138
control.sigma = treatment.sigma = 10
control.n = treatment.n = 25

alpha = 0.05
replicates = 1000

#set seed
set.seed(2018)

#create empty list
p.values = vector("numeric", replicates)

#run repeated t-tests and record p-values
for (k in 1:replicates){
  control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
  treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

reject = (p.values <= alpha)
table(reject)
```

a) Would you expect the probability of rejecting H_0 when H_A is true to increase or decrease

if there is more variation in the observations? (*Hint*: Consider the formula for the test statistic in the independent two-group setting.)

It is reasonable to expect the probability of rejecting H_0 when H_A is true to decrease if there is more variation in the observations. When there is more variation within groups, it is more difficult to discern whether there is a difference in the group means. From a mathematical perspective, high values of σ increase the denominator of the test statistic; thus for some fixed difference between two group means, the test statistic value will be lower, making it more likely that H_0 is not rejected even if H_A is true.

b) How does power change with increased standard deviation? Estimate the power of the test as the standard deviation within each group changes to 5, 10, 15, and 20 (leaving all other parameters the same).

Power decreases as standard deviation increases: from 11.5% to 8.9%, 16.2%, and 7.9%.

```
## reject
## FALSE  TRUE
##   885   115
```

```
## reject
## FALSE  TRUE
##   911    89
```

```
## reject
## FALSE  TRUE
##   921    79
```

6. *Power and Effect Size*. The population effect size refers to the difference between the population means, $\mu_{treatment} - \mu_{control}$. In the simulations so far, $\mu_{treatment} - \mu_{control} = 138 - 140 = -2$ mm Hg.

   In a realistic setting, the effect size is chosen to be the incremental value of the intervention that would justify changing current clinical recommendations from an existing intervention to a new one. The simulations so far mimic a setting in which researchers decide they are interested in detecting an effect on blood pressure that is 2 mm Hg or greater, when comparing the new drug to the old drug.

```
#set parameters
control.mean = 140
treatment.mean = 138
control.sigma = treatment.sigma = 10
control.n = treatment.n = 25

alpha = 0.05
replicates = 1000

#set seed
set.seed(2018)

#create empty list
p.values = vector("numeric", replicates)
```

```r
#run repeated t-tests and record p-values
for (k in 1:replicates){
  control = rnorm(n = control.n, mean = control.mean, sd = control.sigma)
  treatment = rnorm(n = treatment.n, mean = treatment.mean, sd = treatment.sigma)

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

reject = (p.values <= alpha)
table(reject)
```

a) If the true difference in the group means is relatively large (e.g., 5 mm Hg), as opposed to relatively small (e.g., 1 mm Hg), would you expect the probability of rejecting H_0 when H_A to be relatively large or relatively small?

If the difference between the groups is relatively large, it is easier to detect the difference; the observed values are more likely to seem different if they come from distributions with very different means.

b) How does power change with effect size? Estimate the power of the test as effect size increases; change `treatment.mean` from 138 to 137, 136, and 135 (leaving all other parameters the same).

As effect size increases, power increases. Power increased to 17.7%, 28.7%, and 39.7% as effect size increased by increments of 1 mmHg.

```
## reject
## FALSE  TRUE
##   823   177
```

```
## reject
## FALSE  TRUE
##   713   287
```

```
## reject
## FALSE  TRUE
##   603   397
```

Power and sample size calculations

Section 5.4 in the text discusses how power and sample size for a study can be calculated from first principles, using fundamental ideas behind distributions and testing. In practice, power and sample size can be calculated directly from formulas. While hand calculations can provide quick estimates in the early stages of planning a study, statistical software should be the method of choice for a formal analysis.

Calculating sample size for comparing two means

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2},$$

where n is the number of participants in each group, $\Delta = \mu_1 - \mu_2$ is the effect size, $z_{1-\alpha/2}$ is the point on a standard normal distribution with area $1-\alpha/2$ to its left, and $z_{1-\beta}$ is the point on a standard normal distribution with area $1-\beta$ to its left. The null hypothesis of the test is $H_0 : \Delta = 0$, tested against the alternative hypothesis $H_A : \Delta \neq 0$.

Calculating power for comparing two means

$$1 - \beta = P\left(Z < -z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n + \sigma_2^2/n}}\right),$$

where Z is a standard normal random variable and the study has n participants in each group.

The `power.t.test()` function in R can both compute the power of a one- or two-sample t-test (given the appropriate parameters) and determine necessary parameters (e.g., sample size) to obtain a target power. Specific instructions for using `power.t.test()` can be found in the template, the lab notes for this unit, and in the R help file.

7. A pharmaceutical company has developed a new drug to lower blood pressure and is planning a clinical trial to test its effectiveness. Individuals whose systolic blood pressures are between 140 and 180 mm Hg will be recruited for the study. Based on previous published studies, it is estimated that the patients' blood pressures will be approximately symmetrically distributed, with standard deviation of about 12 mm Hg.

   The participants will be randomly assigned to the new drug or a standard drug and at the end of the study their systolic blood pressures will be measured. The company expects to receive FDA approval for the drug if there is evidence at $\alpha = 0.05$ that in the general population of people with blood pressure in the same range, the drug lowers blood pressure, on average, by at least 3 mm Hg more than the standard drug.

   a) How large should the study be if the company wants the power of the study to be 80%?

   The study should enroll at least 253 individuals per group to have a power of 80%. Note that this value differs slightly from the one in the text due to rounding of the z-values.

```
power.t.test(n = NULL, delta = 3, sd = 12,
         sig.level = 0.05, power = .80,
```

```
         type = "two.sample",
         alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 252.1281
##           delta = 3
##              sd = 12
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

b) What would the power of the study be if 200 individuals were recruited for each group?

If 200 individuals were in each group, then the power of the study would be 70%.

```
power.t.test(n = 200, delta = 3, sd = 12,
         sig.level = 0.05, power = NULL,
         type = "two.sample",
         alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 200
##           delta = 3
##              sd = 12
##       sig.level = 0.05
##           power = 0.7033291
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

c) Does α influence power? What would the power of the test be in part b) if α increased to 0.10? What if α decreased to 0.01?

The significance level does influence power. If α increases, then power increases. When α for the test in part b) is 0.10, then the power is 80%; if $\alpha = 0.01$, then power is only 47%.

Consider that increasing α increases the chances of rejecting the null hypothesis, since an observed difference would not need to be as extreme to constitute sufficient evidence against H_0; in other words, increasing α increases the number of "positive" tests. This necessarily increases the ability of the test to detect "true positives", resulting in more cases of rejecting H_0 when H_A is true. Note that the likelihood of "false positives" also increases.

```
power.t.test(n = 200, delta = 3, sd = 12,
         sig.level = 0.10, power = NULL,
         type = "two.sample",
```

```
         alternative = "two.sided",
         strict = TRUE)
```

```
##
##      Two-sample t test power calculation
##
##               n = 200
##           delta = 3
##              sd = 12
##       sig.level = 0.1
##           power = 0.8026031
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(n = 200, delta = 3, sd = 12,
         sig.level = 0.01, power = NULL,
         type = "two.sample",
         alternative = "two.sided",
         strict = TRUE)
```

```
##
##      Two-sample t test power calculation
##
##               n = 200
##           delta = 3
##              sd = 12
##       sig.level = 0.01
##           power = 0.4656354
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```