

Stat 3202 Spring 2023 Homework 5

Nathan Johnson

2023-03-25

Due on Carmen Monday, March 27 before 11:59 pm. All uploads must be .pdf. Submissions will be accepted for 24 hours past this deadline, with a deduction of 1% per hour. Absolutely no submissions will be accepted after this grace period.

Problems will be graded for a combination of correctness and completion.

Academic Integrity

Acceptable for this assignment:

- Working together in a small group
- Working together in a Zoom chat
- Discussing or explaining your approach with students who have also already attempted a problem
- Discussing the assignment with your TA or using MSLC tutoring resources
- Asking me for help in class, office hours, or through Zoom

Not acceptable for this assignment:

- Copying answers from another student or letting another student copy your answers
- Posting solutions to online or communal forums such as a group chat, Chegg, or Stack Exchange
- Using solutions from any previous course or section of Stat 3202

Instructions

- Replace “FirstName LastName (name.n)” above with your information.
- Provide your solutions below in the spaces marked “Solution:”.
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R.
- Knit this document to **pdf** and upload both the **pdf** file and your completed **Rmd** file to Carmen
- Make sure your solutions are clean and easy-to-read by
 - formatting all plots to be appropriately sized, with appropriate axis labels.
 - only including R code that is necessary to answer the questions below.
 - only including R output that is necessary to answer the questions below (avoiding lengthy output).

- providing short written answers explaining your work, and writing in complete sentences.

Total: 20 points

Question 1

Let us take a sample from population X from a normal distribution with **unknown mean** μ_x and **unknown** variance σ_x^2 , and population Y from normal distribution with **unknown mean** μ_y and **unknown** variance σ_y^2 . Suppose that we want to find the interval estimation of differences between two population means $\mu_x - \mu_y$.

Let x_1, x_2, \dots, x_{n_x} denote the random sample of size n_x from the population X and let y_1, y_2, \dots, y_{n_y} denote the random sample of size n_y from the population Y.

Assume that the populations X and Y $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

We have used in our class the usual estimator of the common variance σ^2 is obtained by pooling the sample data to obtain the **pooled variance estimator** s_P^2 :

$$s_P^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2 + \sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_x + n_y - 2}$$

Let s_x^2 and s_y^2 be the sample variances of sample X and sample Y respectively.

a)

Show that s_P^2 is an unbiased estimator for the common variance σ^2 .

b)

Show that if $n_x = n_y$, s_P^2 is simply the average of s_x^2 and s_y^2 .

c)

Show that if $n_x \neq n_y$ and $n_x > n_y$, s_P^2 is the *weighted average* of s_x^2 and s_y^2 , with larger weight given to sample X.

d)

A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A has an average nicotine content of 3.1 milligrams with standard deviation of 0.5 milligrams, while eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviation 0.7 milligrams. Assuming that the two sets of data are independent random samples from normal populations with equal variances, construct a 95% confidence interval for $\mu_A - \mu_B$, the true difference between the mean nicotine contents of the two brands of cigarettes. You may use `ZTest()` or `t.test()` to verify your answer, but replicate the values mathematically yourself and show your work as well.

e)

Repeat part d under the assumption that the variances of each brand are *not* equal. This time, you may use `ZTest()` or `t.test()` on its own. Explain how this problem is different from part d, but no need to mathematically verify the values.


Your answers go here for question 1

Part a: Unable to find unbiased estimator. Look at pdf for work. **Part b:** $s_p^2 = \frac{s_x^2 + s_y^2}{2}$. Work on pdf. **Part c:** Work on pdf. **Part d:**

```
set.seed(1)
brandAn = 10
brandAmean = 3.1
brandAstddev = 0.5
brandBn = 8
brandBmean = 2.7
brandBstddev = 0.7


meandiff = brandAmean - brandBmean
pooledvar = ((brandAn - 1) * brandAstddev ^ 2 + (brandBn - 1) * brandBstddev ^ 2)/(brandAn+brandBn-2)
t = qt(1-0.025, brandAn+brandBn-2)
CIlower = (meandiff) - t*sqrt(pooledvar)*sqrt(1/brandAn+1/brandBn)
CIupper = (meandiff) + t*sqrt(pooledvar)*sqrt(1/brandAn+1/brandBn)
CIlower
```

```
## [1] -0.1991311
```



```
CIupper
```

```
## [1] 0.9991311
```



The lower bound is -0.1991311 and the upper bound is 0.9991311 for a 95% confidence interval of $\mu_A - \mu_B$ and assuming σ^2 is equal.

Part e:

```
set.seed(1)
brandAn = 10
brandAmean = 3.1
brandAstddev = 0.5
brandBn = 8
brandBmean = 2.7
brandBstddev = 0.7
```

```

A2 = brandAstddev ^ 2
B2 = brandBstddev ^ 2
meandiff = brandAmean - brandBmean
degfreedom = ((A2/brandAn + B2/brandBn)^2) / (((A2/brandAn)^2/(brandAn - 1)) + ((B2/brandBn)^2/(brandBn - 1)))
t = qt(1-0.025,degfreedom)
CIlower = meandiff - t*sqrt(brandAstddev^2/brandAn + brandBstddev^2/brandBn)
CIupper = meandiff + t*sqrt(brandAstddev^2/brandAn + brandBstddev^2/brandBn)
CIlower

```

```
## [1] -0.2382208
```

```
CIupper
```

```
## [1] 1.038221
```

The lower bound is -0.2382208 and the upper bound is 1.038221 for a 95% confidence interval of $\mu_A - \mu_B$ and assuming σ^2 is not equal.

Question 2

A 2005 poll by the Center for Social Research at Stony Brook University asked, “Should high school athletes who test positive for steroids or other performance-enhancing drugs be banned from high school athletic teams, or not?” Of the 830 randomly selected respondents, 631 responded, “Yes, they should be banned.” Construct a 95% confidence interval for p , the population proportion of all Americans who think such athletes should be banned. Show your work and be clear about the method you have chosen. There are many reasonable choices.

Your answers go here for question 2 First, test if sample sizes are larger than 5. ✓

```
yes = 631830*(631/830)
yesopp = 631830*(1-(631/830))
yes
```

```
## [1] 479.712
```

```
yesopp
```

```
## [1] 151.288
```

Both are above 5.

```
coverage = c()
phat = (631/830)
stddev = sqrt(((631/830)*(199/830))/830)
for(i in 1:1000) {
  sample = rnorm(830, phat, stddev)
  coverage[i] = mean(sample)
}
z = qnorm(1-0.025)
CIlower = phat - (z*stddev)
CIupper = phat + (z*stddev)
CIlower
```

```
## [1] 0.7311959
```

```
CIupper
```

```
## [1] 0.789286
```

```
phat
```

```
## [1] 0.760241
```

I chose wald type as I used \hat{p} to create my confidence interval.

Question 3

Suppose we want to assess the effectiveness of a short course on nutrition. Eight students were given a test before and after taking a short course on nutrition. The data are shown below.

Student	Before	After	Improvement
1	50	56	6
2	71	78	7
3	87	100	13
4	75	71	-4
5	82	95	13
6	91	96	5
7	63	79	16
8	80	84	4

a)

Please explain why the above sample observations are treated as matched pair (dependent) samples instead of independent samples.

b)

Let the i^{th} student's before (B) and after (A) score be denoted as $x_i^{(B)}$ and $x_i^{(A)}$. Also, let δ_i denotes the i^{th} student's score difference,

$$\delta_i = x_i^{(A)} - x_i^{(B)}$$

Suppose we are interested in an interval estimation for population mean difference μ_δ . Assuming the differences are normally distributed, please construct a 95% confidence interval for μ_δ . You may use `ZTest()` or `t.test()` to verify your answer, but replicate the values mathematically yourself and show your work as well.

Your answers go here for question 3 Part a: The After column is dependent on the Before column.

Part b:

```
before = c(50, 71, 87, 75, 82, 91, 63, 80)
after = c(56, 78, 100, 71, 95, 96, 79, 84)
diff = c(6, 7, 13, -4, 13, 5, 16, 4)
n = 8
mean = mean(diff)
var = sqrt(var(diff))/sqrt(n)
t = qt(1-0.025, 7)
CIlower = mean - t*var
CIupper = mean + t*var
CIlower
```



```
## [1] 2.156188
```

```
CIupper
```

```
## [1] 12.84381
```

```
t.test(diff)
```

```
##  
## One Sample t-test  
##  
## data: diff  
## t = 3.3187, df = 7, p-value = 0.01279  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 2.156188 12.843812  
## sample estimates:  
## mean of x  
## 7.5
```

The lower bound is 2.156188 and the upper bound is 12.84381 for a 95% confidence interval of the mean difference of the matched pair population.