STA501 Midterm Exam #1

Due: Sunday, 11:30 PM, 3/14/21

3/10/2021

1 Introduction

This is an open book and open note exam. The level of details in your solution should similar to that in the examples in the class notes. Please keep in mind that interpretation of results is as important as generation of the results. You can use either the built-in R functions or the formulas given in the class notes to complete the exams.

For confidence interval and testing hypothesis problems, you need to **justify** the sampling distributions and interpret the results. The default confidence level is 0.95 and the default significance level is 0.05.

Problem 1

In a study of physical endurance levels of male college freshmen, the following composite endurance scores based on several exercise routines were collected.

```
254, 281, 192, 260, 212, 179, 225, 179, 181, 149,
182, 210, 235, 239, 258, 166, 159, 223, 186, 190,
180, 188, 135, 233, 220, 204, 219, 211, 245, 151,
198, 190, 151, 157, 204, 238, 205, 229, 191, 200,
222, 187, 134, 193, 264, 312, 214, 227, 190, 212,
165, 194, 206, 193, 218, 198, 241, 149, 164, 225,
265, 222, 264, 249, 175, 205, 252, 210, 178, 159,
220, 201, 203, 172, 234, 198, 173, 187, 189, 237,
272, 195, 227, 230, 168, 232, 217, 249, 196, 223,
232, 191, 175, 236, 152, 258, 155, 215, 197, 210,
214, 278, 252, 283, 205, 184, 172, 228, 193, 130,
218, 213, 172, 159, 203, 212, 117, 197, 206, 198,
169, 187, 204, 180, 261, 236, 217, 205, 212, 218,
191, 124, 199, 235, 139, 231, 116, 182, 243, 217,
251, 206, 173, 236, 215, 228, 183, 204, 186, 134,
188, 195, 240, 163, 208
```

Use the above data to construct a frequency table and a histogram. Describe your findings from the histogram.

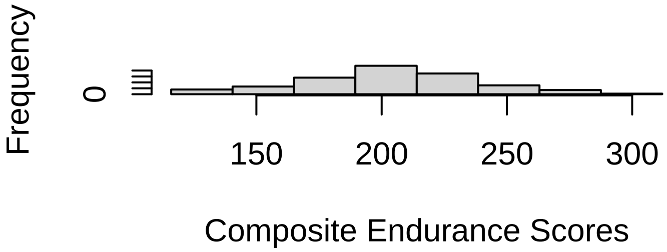
Solution We use R to construct a frequency table and convert it to a histogram. Several R functions will be used in the R code. The number of categories to be used in the frequency table is 8.

```r
dat.file = c(254, 281, 192, 260, 212, 179, 225, 179, 181, 149,
182, 210, 235, 239, 258, 166, 159, 223, 186, 190,
180, 188, 135, 233, 220, 204, 219, 211, 245, 151,
198, 190, 151, 157, 204, 238, 205, 229, 191, 200,
222, 187, 134, 193, 264, 312, 214, 227, 190, 212,
165, 194, 206, 193, 218, 198, 241, 149, 164, 225,
265, 222, 264, 249, 175, 205, 252, 210, 178, 159,
220, 201, 203, 172, 234, 198, 173, 187, 189, 237,
272, 195, 227, 230, 168, 232, 217, 249, 196, 223,
232, 191, 175, 236, 152, 258, 155, 215, 197, 210,
214, 278, 252, 283, 205, 184, 172, 228, 193, 130,
218, 213, 172, 159, 203, 212, 117, 197, 206, 198,
169, 187, 204, 180, 261, 236, 217, 205, 212, 218,
191, 124, 199, 235, 139, 231, 116, 182, 243, 217,
251, 206, 173, 236, 215, 228, 183, 204, 186, 134,
188, 195, 240, 163, 208)
##
cut.off = seq(min(dat.file), max(dat.file), length = 9)
categores = cut(dat.file, breaks = cut.off, include.lowest = TRUE)
freq = table(categores)
rel.freq = table(categores)/sum(table(categores))
cum.freq = cumsum(freq)
cum.rel.freq = cumsum(rel.freq)
freq.table = as.data.frame(cbind(freq = freq, rel.freq = rel.freq, cum.freq = cum.freq, cum.rel.freq=cu
kable(freq.table, caption ="Frequency Table of Composite Endurance Scores")
```

Table 1: Frequency Table of Composite Endurance Scores

	freq	rel.freq	cum.freq	cum.rel.freq
[116,140]	8	0.0516129	8	0.0516129
(140,165]	13	0.0838710	21	0.1354839
(165,190]	28	0.1806452	49	0.3161290
(190,214]	48	0.3096774	97	0.6258065
(214,238]	35	0.2258065	132	0.8516129
(238,263]	15	0.0967742	147	0.9483871
(263,288]	7	0.0451613	154	0.9935484
(288,312]	1	0.0064516	155	1.0000000

```
hist(dat.file, breaks = cut.off,
     main ="",
     xlab ="Composite Endurance Scores")
```



The distribution of composite endurance scores is approximately normally distributed.

Problem 2

Iron-deficiency anemia is an important nutritional health problem in the United States. A dietary assessment was performed on 51 boys 9 to 11 years of age whose families were below the poverty level. The mean daily iron intake among these boys was found to be 12.50 mg with a standard deviation of 4.75 mg. Suppose the mean daily iron intake among a large population of 9- to 11-year-old boys from all income strata is 14.44 mg. We want to test whether the mean iron intake among the low-income group is different from that of the general population. Carry out the hypothesis test using the critical-value method with an α level of .05. State the hypotheses that we can use to consider this question and summarize your findings.

Solution: The objective is to test whether the mean iron intake among the low-income group is different from that of the general population. We can then **claim** that the mean iron intake among the low-income group is different from that of the general population. Therefore, $\mu \neq 14.44$. Therefore,

$$H_o : \mu = 14.44 \leftrightarrow H_a : \mu \neq 14.44$$

Since $n = 51 > 0$, the test statistic is the standard normal random variable.

```
n = 51
xbar = 12.5
s = 4.75
mu0 = 14.44
##
TS = (xbar - mu0)/(s/sqrt(n))
## p.value = 2 times the smaller tail area
p.value = 2*min(pnorm(TS), 1-pnorm(TS))
p.value
```

```
## [1] 0.003537448
```

Since p.value $= 0.0035 < 0.05$, the null hypothesis is rejected. Therefore, we conclude that the mean iron intake among the low-income group is **different** from that of the general population.

Problem 3

A topic of recent clinical interest is the possibility of using drugs to reduce infarct size in patients who have had a myocardial infarction within the past 24 hours. Suppose we know that in untreated patients the mean infarct size is 25(ck-g-EQ/m2). Furthermore, in 8 patients treated with a drug the mean infarct size is 16 with a standard deviation of 10. Is the drug effective in **reducing** infarct size? Assuming that the infarct size is normally distributed.

Solution: The **claim** is that the drug is effective in **reducing** infarct size. That is, $\mu < 25$. This implies that

$$H_o : \mu \geq 25 \leftrightarrow H_a : \mu < 25$$

The population is normal and variance is not given. The test statistic is a t-distribution with (8-1) degrees of freedom.

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \to t_7$$

With the above information, we can use the following R to calculate the p-value.

```
n = 8
xbar = 16
s = 10
mu0 = 25
##
TS = (xbar - mu0)/(s/sqrt(n))
# This is a left-tailed test
p.value =pt(TS, df = 7)
p.value
```

```
## [1] 0.01917501
```

p-value $= 0.0192 < 0.05$, we reject the null hypothesis and conclude that $\mu < 25$.

Problem 4

Drug A was prescribed for a random sample of 12 patients complaining of insomnia. An independent random sample of 16 patients with the same complaint received drug B. The number of hours of sleep experienced during the second night after treatment began were as follows.

```
A: 3.5, 5.7, 3.4, 6.9, 17.8, 3.8, 3.0, 6.4, 6.8, 3.6, 6.9, 5.7
B: 4.5, 11.7, 10.8, 4.5, 6.3, 3.8, 6.2, 6.6, 7.1, 6.4, 4.5, 5.1, 3.2, 4.7, 4.5, 3.0
```

Construct a 95 percent confidence interval for the difference between the population means. Assume that the populations are normal and variances are unknown but equal.

Solution: The objective is to construct a confidence interval of the difference of the two population means. Since both samples are taken from normal populations with equal variance. We can use R function **t.test()** to find the confidence interval directly.

```
A =c(3.5, 5.7, 3.4, 6.9, 17.8, 3.8, 3.0, 6.4, 6.8, 3.6, 6.9, 5.7)
B =c(4.5, 11.7, 10.8, 4.5, 6.3, 3.8, 6.2, 6.6, 7.1, 6.4, 4.5, 5.1, 3.2, 4.7, 4.5, 3.0)
##
t.test(A, B,
       conf.level = 0.95,
       alternative = "two.sided")$conf.int
```

```
## [1] -2.426614  3.064114
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval is [-2.43, 3.06]. We are 95% confident that the true difference between the two population means is in the interval. Since zero is inside the interval, there is no significant difference in the treatments between the two drugs.

Problem 5

Can we conclude that the mean age at death of patients with homozygous sickle-cell disease is less than 30 years? A sample of 50 patients yielded the following ages in years:

```
15.5,  2.0, 45.1,  1.7,  0.8,  1.1, 18.2,  9.7, 28.1, 18.2,
27.6, 45.0,  1.0, 66.4,  2.0, 67.4,  2.5, 61.7, 16.2, 31.7,
 6.9, 13.5,  1.9, 31.2,  9.0,  2.6, 29.7, 13.5,  2.6, 14.4,
20.7, 30.9, 36.6,  1.1, 23.6,  0.9,  7.6, 23.5,  6.3, 40.2,
23.7,  4.8, 33.2, 27.1, 36.7,  3.2, 38.0,  3.5, 21.8,  2.4
```

Let $\alpha = 0.05$. What assumptions are necessary? Interpret the result you obtained from the data.

Solution: The **claim** is that the mean age is less than 30 years. That is, $\mu < 30$. This is a left-tailed test.

$$H_o : \mu \ge 30 \leftrightarrow H_a : \mu < 30$$

Since the sample size $n = 50 > 30$, by CLT, $\bar{x}$ is normally distributed. This implies that the distribution of the test statistic

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \to N(0, 1)$$

The p-value of this left-tailed test is defined to be

```
age=c(15.5,  2.0, 45.1,  1.7,  0.8,  1.1, 18.2,  9.7, 28.1, 18.2,
      27.6, 45.0,  1.0, 66.4,  2.0, 67.4,  2.5, 61.7, 16.2, 31.7,
       6.9, 13.5,  1.9, 31.2,  9.0,  2.6, 29.7, 13.5,  2.6, 14.4,
      20.7, 30.9, 36.6,  1.1, 23.6,  0.9,  7.6, 23.5,  6.3, 40.2,
      23.7,  4.8, 33.2, 27.1, 36.7,  3.2, 38.0,  3.5, 21.8,  2.4)
n = length(age)
xbar = mean(age)
s = sd(age)
mu0 = 30
##
TS = (xbar - mu0)/(s/sqrt(n))
## left-tailed
p.value = pnorm(TS)
p.value
```

```
## [1] 1.438406e-05
```

The p-value is close to zero. We reject the null hypothesis and conclude that the mean age is less than 30 years old.

Problem 6

Can we conclude that, on average, lymphocytes and tumor cells differ in size? The following are the cell diameters (μm) of 40 lymphocytes and 50 tumor cells obtained from biopsies of tissue from patients with melanoma.

```
## Lymphocytes
9.0,  9.4,  4.7,   4.8,  8.9,  4.9,  8.4,  5.9,  6.3,  5.7,
5.0,  3.5,  7.8,  10.4,  8.0,  8.0,  8.6,  7.0,  6.8,  7.1,
5.7,  7.6,  6.2,   7.1,  7.4,  8.7,  4.9,  7.4,  6.4,  7.1,
6.3,  8.8,  8.8,  5.2,  7.1,  5.3,   4.7,  8.4,  6.4,  8.3

## Tumor Cells
12.6, 14.6, 16.2, 23.9, 23.3, 17.1, 20.0, 21.0, 19.1, 19.4,
16.7, 15.9, 15.8, 16.0, 17.9, 13.4, 19.1, 16.6, 18.9, 18.7,
20.0, 17.8, 13.9, 22.1, 13.9, 18.3, 22.8, 13.0, 17.9, 15.2,
17.7, 15.1, 16.9, 16.4, 22.8, 19.4, 19.6, 18.4, 18.2, 20.7,
16.3, 17.7, 18.1, 24.3, 11.2, 19.5, 18.6, 16.4, 16.1, 21.5
```

What statistical method you are going to use to conduct the analysis. What are the assumptions for the method? Interpret your result appropriately.

Solution: The **claim** is that lymphocytes and tumor cells differ in size. That is, $\mu_{lymphocytes} - \mu_{tumor} \neq 0$. Then

$$H_o : \mu_{lymphocytes} - \mu_{tumor} = 0 \leftrightarrow H_a : \mu_{lymphocytes} - \mu_{tumor} \neq 0$$

Since both sample sizes are considered to be large, the test statistic is normally distributed

$$TS = \frac{(\bar{x} - \bar{y}) - (0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \to N(0, 1)$$

```
## Lymphocytes
x=c(9.0,  9.4,  4.7,   4.8,  8.9,  4.9,  8.4,  5.9,  6.3,  5.7,
5.0,  3.5,  7.8,  10.4,  8.0,  8.0,  8.6,  7.0,  6.8,  7.1,
5.7,  7.6,  6.2,   7.1,  7.4,  8.7,  4.9,  7.4,  6.4,  7.1,
6.3,  8.8,  8.8,  5.2,  7.1,  5.3,   4.7,  8.4,  6.4,  8.3)

## Tumor Cells
y=c(12.6, 14.6, 16.2, 23.9, 23.3, 17.1, 20.0, 21.0, 19.1, 19.4,
16.7, 15.9, 15.8, 16.0, 17.9, 13.4, 19.1, 16.6, 18.9, 18.7,
20.0, 17.8, 13.9, 22.1, 13.9, 18.3, 22.8, 13.0, 17.9, 15.2,
17.7, 15.1, 16.9, 16.4, 22.8, 19.4, 19.6, 18.4, 18.2, 20.7,
16.3, 17.7, 18.1, 24.3, 11.2, 19.5, 18.6, 16.4, 16.1, 21.5)
##
n1 = length(x)
xbar = mean(x)
s1 = sd(x)
n2 = length(y)
ybar = mean(y)
s2 = sd(y)
##
TS = (xbar-ybar)/sqrt(s1^2/n1 + s2^2/n2)
## two-tailed test
p.value = 2*min(pnorm(TS), 1-pnorm(TS))
p.value
```

```
## [1] 4.289584e-111
```

The p-value is almost 0. We reject the null hypothesis and conclude that lymphocytes and tumor cells differ in size. The assumption used in this solution is that both sample sizes are large.

Problem 7

To evaluate the analgesic effectiveness of a daily dose of oral methadone in patients with chronic neuropathic pain syndromes. The researchers used a visual analogue scale (0–100 mm, a higher number indicates higher pain) ratings for maximum pain intensity over the course of the day. Each subject took either 20 mg of methadone or a placebo each day for 5 days. Subjects did not know which treatment they were taking. The following table gives the mean maximum pain intensity scores for the 5 days on methadone and the 5 days on placebo.

subject ID:	1	2	3	4	5	6	7	8	9	10	11
methadone:	29.8	73.0	98.6	58.8	60.6	57.2	57.2	89.2	97.0	49.8	37.0
placebo:	57.2	69.8	98.2	62.4	67.2	70.6	67.8	95.6	98.4	63.2	63.6

Do these data provide sufficient evidence, at the .05 level of significance, to indicate that, in general, the maximum pain intensity is lower on days when methadone is taken? Perform a formal inferential procedure and interpret the result.

Solution: The **claim** is that, in general, the maximum pain intensity is lower on days when methadone is taken. in other words, $\mu_{methadone} < \mu_{placebo}$.

This is a typical paired sample problem. We will use **t.test()** to perform the test using option: `paired = TRUE`. The assumption used in the **t.test()** is that both of the populations (pre-and post- populations) are normal. The null and alternative hypothesis is given by

$$H_o : \mu_{methadone} - \mu_{placebo} \geq 0 \leftrightarrow H_a : \mu_{methadone} - \mu_{placebo} < 0$$

The following code use **t.test()** to find the p-value.

```
methadone =c(29.8, 73.0, 98.6, 58.8, 60.6, 57.2, 57.2, 89.2, 97.0, 49.8, 37.0)
placebo =  c(57.2, 69.8, 98.2, 62.4, 67.2, 70.6, 67.8, 95.6, 98.4, 63.2, 63.6)
```

```
##
t.test(methadone, placebo,    # sample 1, sample 2: order is important!
       conf.level = 0.95,     # significant level = 1 - confidence level
       alternative = "less",  # left-tailed test
       paired = TRUE          # paired t-test
       )
```

```
##
##  Paired t-test
##
## data:  methadone and placebo
## t = -3.1554, df = 10, p-value = 0.005119
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.093522
## sample estimates:
## mean of the differences
##               -9.618182
```

The p-value for testing the hypothesis is $0.0051 < 0.05$. We reject the null hypothesis at the level of 0.05. Therefore, we conclude that, in general, the maximum pain intensity is lower on days when methadone is taken.

Problem 8

A study was conducted of genetic and environmental influences on cholesterol levels. The data set used for the study was obtained from a twin registry in Sweden. Specifically, four populations of adult twins were studied: (1) monozygotic (MZ) twins reared apart, (2) MZ twins reared together, (3) dizygotic (DZ) twins reared apart, and (4) DZ twins reared together. One issue is whether it is necessary to correct **potential differences** for sex before performing more complex genetic analyses. The data in the following table were presented for total cholesterol levels for MZ twins reared apart, by sex.

	Men	Women
Mean:	253.3	271.0
sd:	44.1	44.1
size(n):	44	48

If we assume (a) serum cholesterol is normally distributed, (b) the samples are independent, and (c) the standard deviations for men and women are the same.

Using a two-sided test. State the hypotheses being tested, and implement the method. Report a p-value and interpret the result.

Solution: Based on the assumptions given above, we should use the two-sample t-test with unknown equal variances of the two underlying normal populations. However, both sample sizes are large, we can use the two-sample normal test using the CLT.

The **claim** is the `potential difference` in the total cholesterol levels between male and female groups. Or equivalently, $\mu_M - \mu_F \neq 0$. Therefore,

$$H_o : \mu_M - \mu_F = 0 \leftrightarrow H_a : \mu_M - \mu_F \neq 0$$

We perform both normal and t-tests in the following code.

```
# Mean:     253.3     271.0
# sd:        44.1      44.1
# size(n):   44        48
xbar = 253.3
```

```r
s1 = 44.1
n1 = 44
ybar = 271.0
s2 = 44.1
n2 = 48
##
TS = (xbar-ybar)/sqrt(s1^2/n1+s2^2/n2)
## Use both normal and t distribution to find the p-value
p.value.z = 2*min(pnorm(TS), 1-pnorm(TS))
p.value.t = 2*min(pt(TS, df = n1+n2-2), 1-pt(TS, df = n1+n2-2))
cbind(Z.test = p.value.z, T.test = p.value.t)
```

```
##          Z.test     T.test
## [1,] 0.05447536 0.05763708
```

As expected, the resulting p values are close to each other. Both p values are slightly bigger than the significance level 0.05. We fail to reject the null hypothesis at the level of 0.05. Therefore, we conclude that there is a significant difference in the total cholesterol levels between male and female groups.