HW7

Teryl Schmidt | tschmidt6@wisc.edu | 9072604920 | Discussion 312 | Grader: Chi-Shain Dai

11/9/2018

Problem 1

A researcher is interested in comparing the weight gain of young rabbits fed two different diets, diet I and diet II. The researcher selects 7 pairs of male litter mates (that is, 7 pairs of brothers) and randomly selects one rabbit from each pair for diet I and the other for diet II. The weight gains are recorded for each rabbit over an 8 week period. The weight gains for all rabbits (in grams) are given below:

```
Litter# |  1  |  2  |  3  |  4  |  5  |  6  |  7  |
DietI   | 368 | 293 | 401 | 314 | 384 | 404 | 267 |
DietII  | 422 | 298 | 423 | 346 | 375 | 431 | 290 |
```

a. It is claimed that the mean weight gain from diet II is 5 grams greater than the mean weight gain from diet I. Evidence against the claim will be provided when the weight gain from II exceeds the weight gain from I by more than 5 grams. Carry out a test for the claim at a 10% level by hand and then check your work in R. Interpret the results.

μ1 = Mean weight from Diet I
μ2 = Mean weight from Diet II
Test if mean weight gain from Diet II is 5 grams greater than the mean weight gain from Diet I.
Ho: μ1 = μ2
Ha: μ1 < μ2

```
DietI = c(368, 293, 401, 314,  384,  404,  267)
DietII = c(422,  298,  423,  346,  375,  431,  290)
mean(DietI)
```

```
## [1] 347.2857
```

```
mean(DietII)
```

```
## [1] 369.2857
```

$S^2_1 = \Sigma(xi - \bar{x})^2 / n$
$= \Sigma(xi)^2 / n - (\bar{x})^2$
$368^2 + 293^2 + 401^2 + 314^2 + 384^2 + 404^2 + 267^2 = 862631$
$= (862631 / 7) - (347.2857)^2 = 2625.633$

$S^2_2 = \Sigma(yi - \bar{y})^2 / n$
$= \Sigma(yi)^2 / n - (\bar{y})^2$
$422^2 + 298^2 + 423^2 + 346^2 + 375^2 + 431^2 + 290^2 = 976019$
$= (976019 / 7) - (369.2857)^2 = 3059.347$

$S^2_p = n1 * S^2_1 + n2 * S^2_2 / n1 + n2 - 2$
$7 * 2625.633 + 7 * 3059.347 / 7 + 7 - 2 = 3316.238$
$S_p = sqrt(S^2_p) = sqrt(3316.238) = 57.59$

```
t.test(DietI, DietII, var.equal = T, conf.level = 0.90, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  DietI and DietII
## t = -0.71472, df = 12, p-value = 0.2442
## alternative hypothesis: true difference in means is less than 0
## 90 percent confidence interval:
##      -Inf 19.74631
## sample estimates:
## mean of x mean of y
##  347.2857  369.2857
```

pvalue = 0.2442 > 0.05 so we reject the null.

b. State the assumptions necessary for performing the test in (a).

Assumptions: The two samples are independent, Both samples are normal or the two sample sizes are small, Both variances are unknown bu equal.

c. Construct a 90% confidence interval for the difference in mean weight gains for the two diets. Compare this confidence interval to the conclusions from the test you performed above.

$(\bar{x} - \bar{y}) +- t * s_p * \sqrt{1 / n1 + 1 / n2}$
$t = (\bar{x} - \bar{y}) / s_p * \sqrt{1 / n1 + 1 / n2}$
$= 347.2857 - 369.2857 / 57.59 * \sqrt{1 / 7 + 1 / 7}$
$= -22 / 57.59 * 0.5346 = -0.71472$
When $\alpha = 0.10$
$-t_{(n1 + n2 - 2, 0.10)} = -t_{(12, 0.10)} = -1.356$
-0.71472 > -1.356 so we accept the null at 10% level. We conclude that the mean weight gain from DietII is equal to the weight gain from DietI.

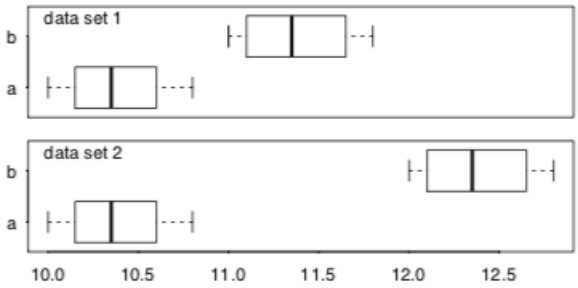
Problem 2

Data set 2 is collected to compare two treatments “a” and “b”. The observations are collected independently for each treatment. Also the samples corresponding to treatments “a” and “b” are independent. Comparison of the treatments were previously done based on data set 1 under same independence assumptions. With data set 1 (top panel) the Wilcoxon Rank Sum/Mann-Whitney test test gives a p-value p = 0.028 and the t-test gives p = 0.006.

Data set 2 (bottom panel) has the same number of values in each of treatments a and b (ie, $n_{a1} = n_{a2}$ and $n_{b1} = n_{b2}$).
For data set 2, the Wilcoxon Rank Sum/Mann-Whitney test test p-value is
☐ smaller than 0.028
☐ 0.028
☐ larger than 0.028

and the t-test p-value is
☐ smaller than 0.006
☐ 0.006
☐ larger than 0.006

Justify your choices.

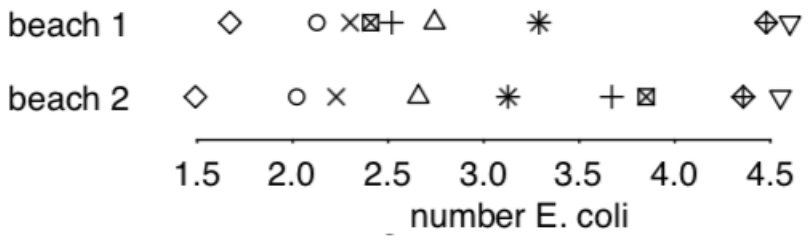


0.028 is correct.
In data set, the second sample is too far than first sample but mann whitney test only uses the rank of the data not its absolute value so Here for Wilcoxon rank sum or Mann Whitney Test the p-value is same as data set 1.

Smaller than 0.006 is correct.
but for t-test the p-value depends on the difference in sample means so here p-value is smaller than 0.006.

Problem 3

In a study of water quality in the Chicago area, concentration of E. Coli bacteria was measured in two specific locations along the lake (“Beach 1” and “Beach 2”). These measurements were taken after rainfalls, when a higher volume of water flows from the city into the lake. Rainfall dates were chosen to be at least a week apart, so that measurements taken from different dates can be considered as independent. The concentrations were measured in number of E. Coli bacteria per ml of water. They are shown below. Data taken on the same rainfall date are shown with the same symbol (with 9 rainfall dates total).



The investigator wants to know if the concentration of *E. Coli* in the water is the same at these two beach locations.

a. Determine which test might be appropriate to answer the investigators questions: check all that apply in the list below. Justify your choices.

Two independent sample t-test
Welch's t-test
Wilcoxon Rank Sum/Mann-Whitney test
Paired-sample t-test
Sign test
Wilcoxon signed rank test
Bootstrap test for one sample
Bootstrap test for two samples
There are two independent samples. But concentration of E. Coli is number of E. Coli bacteria per ml of water may not follow normal distribution and sample size 9 is small. There fore nonparametric method is preferable.

Problem 4

In an effort to link cold environment to an increase in mean blood pressure, two random samples of 5 rats each were exposed to different environments. One sample of rats was held in a normal environment of 26◦C and the other was held aat 15◦C. Blood pressures were measured for rats of both groups after 1 day and are given below:

a. If the scientists want to assume that the necessary populations are normal, what test[s] would be reasonable to run? Explain why. Identify the hypotheses of interest and then run this/these tests in R and report the test statistic, degrees of freedom, and resulting p value (it may also be useful to compute these by hand for practice, but we will not be grading you on it).

```
26◦C BP: | 214 | 194 | 221 | 198 | 212 |
15◦C BP: | 225 | 215 | 253 | 272 | 254 |
```

```
Rats26 = c(214, 194, 221, 198, 212)
Rats15 = c(225, 215, 253, 272, 254)
mean(Rats26)
```

```
## [1] 207.8
```

```
mean(Rats15)
```

```
## [1] 243.8
```

```
sd(Rats26)
```

```
## [1] 11.36662
```

```
sd(Rats15)
```

```
## [1] 23.27445
```

Confidence interval, Z test, T test.
Ho: μ >= 0
Ha: μ < 0
The pvalue < 0.05 so we reject the null hypothesis.

```
t.test(Rats26, Rats15, paired = F, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Rats26 and Rats15
## t = -3.1078, df = 5.8054, p-value = 0.01092
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -13.35509
## sample estimates:
## mean of x mean of y
##     207.8     243.8
```

b. If the researcher does not want to assume that the relevant populations are normally distributed, what 3 tests could they perform?

Bootstrap Confidence interval, Wilcoxn ranked sum test, T test.

c. Perform each of the three tests listed above (using set.seed(1) and B=10000 for any that require computer simulation)

1. identify the assumptions

Bootstrap: Independence
Wilcoxn: Independence
T test: Independence

2. report the observed test statistic and

```
B = 10000
#Bootstrap for Two Sample Test
boottwo <- function(dat1, dat2, B) {
  bootstat <- numeric(B)
  truediff <- mean(dat1) - mean(dat2)
  n1 <- length(dat1)
  n2 <- length(dat2)
  for (i in 1:B) {
    samp1 <- sample(dat1, size = n1, replace = T)
    samp2 <- sample(dat2, size = n2, replace = T)
    bootmean1 <- mean(samp1)
    bootmean2 <- mean(samp2)
    bootvar1 <- var(samp1)
    bootvar2 <- var(samp2)
    bootstat[i] <- (bootmean1 - bootmean2 - truediff)/sqrt((bootvar1/n1) + (bootvar2/n2))
  }
  return(bootstat)
}


set.seed(1)
tobs = -3.1078
RatBoot = boottwo(Rats26, Rats15, B)
(m.low<-sum(RatBoot<= tobs))  #205
```

```
## [1] 205
```

```
(m.up<-sum(RatBoot >= tobs))   #9795
```

```
## [1] 9795
```

```
(pval<-m.low/B)  #0.0205
```

```
## [1] 0.0205
```

```
wilcox.test(Rats26, Rats15, paired = F, alternative = "less")
```

```
##
##  Wilcoxon rank sum test
##
## data:  Rats26 and Rats15
## W = 1, p-value = 0.007937
## alternative hypothesis: true location shift is less than 0
```

```
t.test(Rats26, Rats15, paired = F, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Rats26 and Rats15
## t = -3.1078, df = 5.8054, p-value = 0.01092
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -13.35509
## sample estimates:
## mean of x mean of y
##     207.8     243.8
```

3. the resulting p value.

Bootstrap pvalue = 0.0205 which is less than 0.05 so we reject the null hypothesis.
Wilcoxn pvalue = 0.007937 which is less than 0.05 so we reject the null hypothesis.
T test pvalue = 0.01092 which is less than 0.05 so we reject the null hypothesis.

d. Based on your findings from the tests in parts a and b, what conclusion would you draw and what recommendations would you give to the scientist?

We have enough evidence to conclude that rats exposed to 15C have a higher blood pressure than rats exposed to 26C.

Problem 5

A reporter for Time Magazine was interested in residents' levels of worry about being the victim of crime in their neighborhood. They performed a telephone poll of 500 adult Americans, 140 from urban areas, 160 from suburban, and 200 from rural areas. The number of adults who reported worring about being a victim of crime was urban:83, suburban: 92, and rural: 86.

a. Perform a hypothesis test at the 5% level of significance to determine if there is evidence of a difference in the proportion of urban and suburban residents who worry about being the vicim of crime? (Be sure to state your hypotheses, assumptions, and show your computations.)

Ho: $P_1 = P_2$
Ha: $P_1 =/= P_2$
P = (p1 + p2) / (n1 + n2)
P = (83 + 92) / (140 + 160) = 0.5833
SE = sqrt(p * (1 - p) * [(1 / n1) + (1 / n2)])
SE = sqrt(0.5833 * (1 - 0.5833) * [(1 / 140) + (1 / 160)]) = 0.057
Z = ((p1 / n1) - (p2 / n2)) / SE
Z = ((83 / 140) - (92 / 160)) / 0.057 = 0.31328
(1 - 0.31328) = 0.68672 because two tailed Check the Z table for 0.68 - 0.69 which is 0.754 Since the pvalue is 0.754, which is greater than 0.05. We don't have enough evidence to reject the null.

b. Create a 95% confidence interval for the difference in proportion of rural and urban residents who worry about being a victim of crime.(Be sure to state your assumptions and show your computations.)

Assumptions: ????? CI = (p1 - p2) +- $Z_{0.05 / 2}$ * sqrt(p1 * (1 - p1) / n1 + p2 * (1 - p2) / n2)
p1 = 83 / 140 = 0.5928571
p2 = 92 / 160 = 0.575
CI = (0.5928571 - 0.575) +- 1.96 * sqrt(0.5928571 * (1 - 0.5928571) / 140 + 0.575 * (1 - 0.575) / 160)
CI = (0.5928571 - 0.575) +- 1.96 * 0.05702166
CI = 0.0178571 +- 0.1117625
CI = (-0.0939054, 0.1296196)