Advanced Data Analysis
Haoyang Chen | hc2812 | Assignment 2

1. Determine whether there is a significant difference in the mean weights of chicks fed *soybean* vs. those fed *casein* using each of the following procedures

a). A parametric procedure

- Assumption: The distributions of two samples have the same variance; the samples are independently random selected from the distributions respectively
- Validity: The test is valid when the distributions are normal with common variance.
- Remedial Procedure: Could test whether the variances are same first, if not same, use F test or T test without equal variance

Test Result: *p-value* = 0.002869 indicates that the mean of soybean's weight is significantly different with the casein's weight.

```
> t.test(soybean, casein, var.equal = TRUE)

     Two Sample t-test

data:  soybean and casein
t = -3.3199, df = 24, p-value = 0.002869
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -125.12024  -29.18928
sample estimates:
mean of x mean of y
 246.4286  323.5833
```

b). A non-parametric procedure

- Assumption: No assumption underlying the distribution
- Validity: The test is valid especially when the sample size is not large enough
- Remedial Procedure: Could draw a density plot first, if the distribution is normal, could use parametric methods

Test Result: p-value = 0.005919 indicates that the mean of soybean's weight is significantly different with the casein's weight.

```
     Wilcoxon rank sum test with continuity correction

data:  soybean and casein
W = 30, p-value = 0.005919
alternative hypothesis: true location shift is not equal to 0
```

c). A re-sampling procedure

- Assumption: No assumption underlying the distribution
- Validity: The test is valid when the sample size is not large
- Remedial Procedure: None

Test Result: *p-value = 0.0051* indicates that the mean of soybean's weight is significantly different with the casein's weight.

```
> bootstrap.test(soybean, casein, 10000)
[1] 0.0051
```

2. Using the data for chicks fed casein and those on sunflower, compute the following, based on a suitable bootstrap method.

a). A 95% confidence interval for the difference in median weight for the two groups

- Assumption: No assumption underlying the distribution
- Validity: The test is valid when the sample size is not large
- Remedial Procedure: None

Result:

```
> MedianDiff.ConfidenceInterval(casein, sunflower, 100000)
 2.5% 97.5%
-58.0  51.5
```

b). A 95% bootstrap confidence interval for the ratio of the variances casein fed to sunflower fed chicks

- Assumption: No assumption underlying the distribution
- Validity: The test is valid when the sample size is not large
- Remedial Procedure: None

Result:

```
> ratio.ConfidenceInterval(casein, sunflower, 10000)
      2.5%      97.5%
 0.5355768 10.5640564
```

c). 95% confidence interval for the ratio of the variances of casein fed to sunflower fed chicks under normal assumption.

- Assumption: The ration of the variances is under normal distribution
- Validity: The test is valid when the ration is under normal distribution
- Remedial Procedure: None

Result: 95% confidence interval is (2.529837, 2.664038)

```
> NormalRatio.ConfidenceInterval(casein, sunflower, 10000)
[1] 2.529837 2.664038
attr(,"conf.level")
[1] 0.95
```

3. Assume that if the weight of a chick is below 256, that chick is classified under “LOW WEIGHT”. For chicks fed meatmeal vs. those fed soybean.

a). Determine whether there is a significant difference in the proportions of the chicks classified under “LOW WEIGHT”.
Result: p-value = 0.1511 indicates that there is not a significant difference for the proportion of “LOW WEIGHT” in the two groups.

```
> prop.test(x, n)

     2-sample test for equality of proportions with continuity
     correction

data:  x out of n
X-squared = 2.0607, df = 1, p-value = 0.1511
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.81498148  0.07472174
sample estimates:
   prop 1    prop 2
0.2727273 0.6428571
```

b). Construct a 95% confidence interval for the difference in the proportions of the chicks classified under “LOW WEIGHT”.

Result: 95% confidence interval of the difference of proportions for two groups is
(-0.81498148, 0.07472174)

Code:

```
# This is assignment 1
data("chickwts")

# 1 test significant diff
soybean <- chickwts[chickwts$feed == 'soybean',][['weight']]
casein <- chickwts[chickwts$feed == 'casein',][['weight']]

# parametric procedure
t.test(soybean, casein, var.equal = TRUE)

# Non-parametric procedure
```

```
wilcox.test(soybean, casein)

# re-sampling procedure
bootstrap.test <- function(x, y, B){
    x.mean <- mean(x)
    y.mean <- mean(y)
    x.var <- var(x)
    y.var <- var(y)
    n <- length(x)
    m <- length(y)
    z.obs <- abs((x.mean - y.mean) / sqrt(x.var / n + y.var / m))
    y <- y + x.mean - y.mean
    z <- c()
    for(i in 1:B){
        x.star <- sample(x, n, replace = TRUE)
        y.star <- sample(y, m, replace = TRUE)
        x.star.mean <- mean(x.star)
        y.star.mean <- mean(y.star)
        x.star.var <- var(x.star)
        y.star.var <- var(y.star)
        z[i] <- abs((x.star.mean - y.star.mean) / sqrt(x.star.var / n +
y.star.var / m))
    }
    return(sum(z > z.obs) / B)
}

bootstrap.test(soybean, casein, 10000)


# 2 confidenct interval
sunflower <- chickwts[chickwts$feed == 'sunflower',][['weight']]

# a. 95% confidence interval for median

MedianDiff.ConfidenceInterval <- function(x, y, B){
    median.diff <- c()
    for(i in 1: B){
        x.sample <- sample(x, length(x), replace = T)
        y.sample <- sample(y, length(y), replace = T)
        median.diff[i] <- median(x.sample) - median(y.sample)
    }
    print(quantile(median.diff, c(.025, .975)))
}

MedianDiff.ConfidenceInterval(casein, sunflower, 100000)



# b. 95% CI for ratio of the variance
ratio.ConfidenceInterval <- function(x, y, B){
    ratio <- c()
    for(i in 1: B){
        x.sample <- sample(x, length(x), replace = T)
        y.sample <- sample(y, length(y), replace = T)
        ratio[i] <- var(x.sample) / var(y.sample)
    }
    print(quantile(ratio, c(.025, .975)))
}
```

```
ratio.ConfidenceInterval(casein, sunflower, 10000)


# c. 95% CI for ratio of the variance under normal assumption
NormalRatio.ConfidenceInterval <- function(x, y, B){
    ratio <- c()
    for(i in 1: B){
        x.sample <- sample(x, length(x), replace = T)
        y.sample <- sample(y, length(y), replace = T)
        ratio[i] <- var(x.sample) / var(y.sample)
    }
    print(t.test(ratio)$conf.int)
}

NormalRatio.ConfidenceInterval(casein, sunflower, 10000)


# 3 significant diff

meatmeal <- chickwts[chickwts$feed == 'meatmeal',][['weight']]
meatmeal.LowWeight <- sum(meatmeal < 256)
soybean.LowWeight <- sum(soybean < 256)
x <- c(meatmeal.LowWeight, soybean.LowWeight)
n <- c(length(meatmeal), length(soybean))
prop.test(x, n)
```