**Question 1**

(a)

> runs=read.table("baseball.txt",header=T)

> runs

Team Runs League

1 Rockies 10.4 National

2 Phillies 10.3 National

3 Cincinnati 10.0 National

4 Diamondbacks 10.0 National

5 Braves 9.7 National

6 Dodgers 9.7 National

7 Cubs 9.6 National

8 Giants 9.5 National

9 Nationals 9.5 National

10 Brewers 9.4 National

11 Pirates 9.1 National

12 Marlins 9.0 National

13 Cardinals 9.0 National

14 Astros 8.9 National

15 Mets 8.9 National

16 Padres 7.9 National

17 Royals 11.4 American

18 WhiteSox 10.6 American

19 Rangers 10.4 American

20 RedSox 10.3 American

21 Orioles 10.2 American

22 BlueJays 10.0 American

23 Yankees 9.9 American

24 Indians 9.9 American

25 DevilRays 9.7 American

26 Tigers 9.1 American

27 Twins 9.0 American

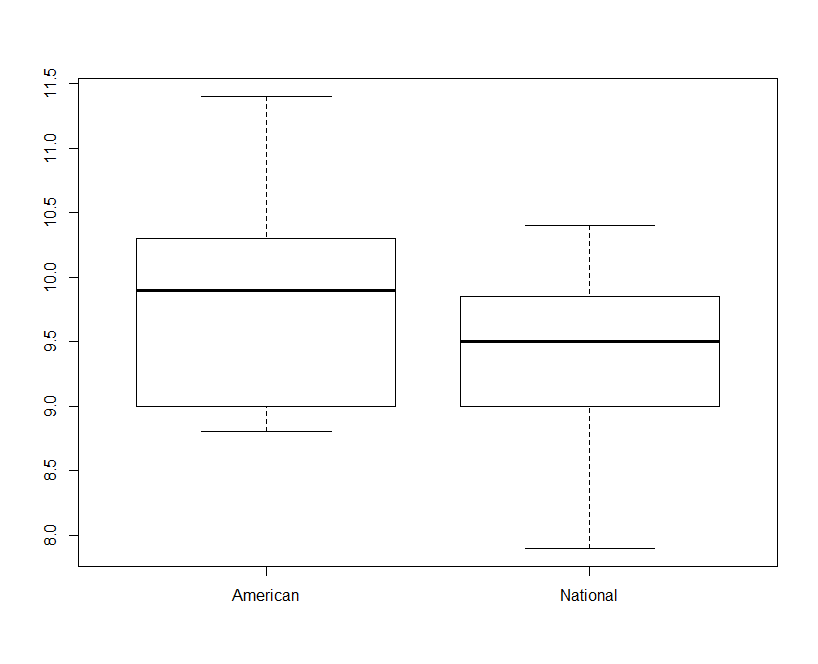
28 Mariners 9.0 American

29 Athletics 8.9 American

30 Angels 8.8 American

(b)

> boxplot(Runs~League,data=runs)



From the boxplot, we can see that the median of the National League is smaller than that of the American League.

(c)

> t.test(Runs~League,data=runs,alternative="greater")

Welch Two Sample t-test

data: Runs by League

t = 1.4298, df = 25.231, p-value = 0.08252

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.0716207 Inf

sample estimates:

mean in group American mean in group National

9.80000 9.43125

Since the P-value=0.08252 > 0.05, therefore we cannot reject the null hypothesis ( i.e., there is no difference in means).

**Question 2**

(a)

*CODE:*

>data baseball;

> infile '/home/shu.tu/baseball.txt' firstobs=2;

> input team $ runs leagues $;

>

>proc print;

*RESULTS:*

| **Obs** | **team** | **runs** | **leagues** |
| --- | --- | --- | --- |
| **1** | Rockies | 10.4 | National |
| **2** | Phillies | 10.3 | National |
| **3** | Cincinna | 10.0 | National |
| **4** | Diamondb | 10.0 | National |
| **5** | Braves | 9.7 | National |
| **6** | Dodgers | 9.7 | National |
| **7** | Cubs | 9.6 | National |
| **8** | Giants | 9.5 | National |
| **9** | National | 9.5 | National |
| **10** | Brewers | 9.4 | National |
| **11** | Pirates | 9.1 | National |
| **12** | Marlins | 9.0 | National |
| **13** | Cardinal | 9.0 | National |
| **14** | Astros | 8.9 | National |
| **15** | Mets | 8.9 | National |
| **16** | Padres | 7.9 | National |
| **17** | Royals | 11.4 | American |
| **18** | WhiteSox | 10.6 | American |
| **19** | Rangers | 10.4 | American |
| **20** | RedSox | 10.3 | American |
| **21** | Orioles | 10.2 | American |
| **22** | BlueJays | 10.0 | American |
| **23** | Yankees | 9.9 | American |
| **24** | Indians | 9.9 | American |
| **25** | DevilRay | 9.7 | American |
| **26** | Tigers | 9.1 | American |
| **27** | Twins | 9.0 | American |
| **28** | Mariners | 9.0 | American |
| **29** | Athletic | 8.9 | American |
| **30** | Angels | 8.8 | American |

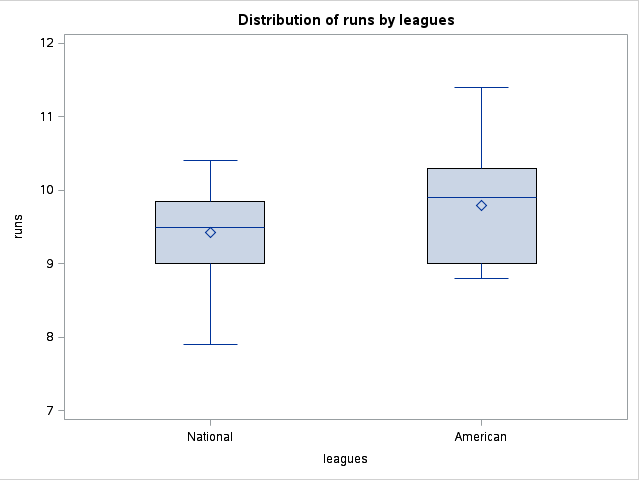
(b)

*CODE:*

>proc boxplot;

> plot runs\*leagues/boxstyle=schematic;

*RESULTS:*



(c)

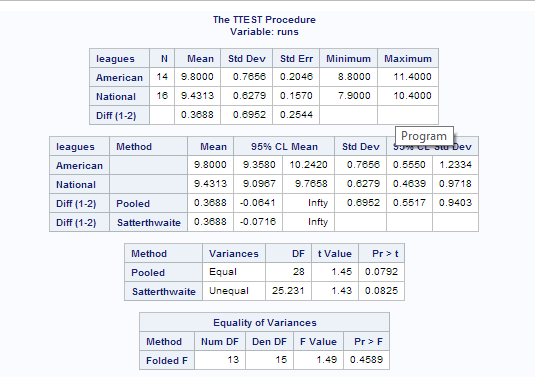
*CODE:*

>proc ttest side=u;

> var runs;

> class leagues;

*RESULTS:*



Since the P-value in R is 0.8252, which is the same as SAS had, therefore, R did the Satterthwaite test. In SAS, the P-values for both methods are bigger than 0.05, hence, we cannot reject the null hypothesis (i.e., there is no difference in means) at α=0.05.

**Question3**

(a)

> snacks=read.table("snacks.txt",header=T)

> snacks

Weights

1 29.2

2 28.5

3 28.7

4 28.9

5 29.1

6 29.5

(b)

> attach(snacks)

> t.test(Weights,mu=29.2)

One Sample t-test

data: Weights

t = -1.4739, df = 5, p-value = 0.2005

alternative hypothesis: true mean is not equal to 29.2

95 percent confidence interval:

28.60544 29.36123

sample estimates:

mean of x

28.98333

Since the P-value=0.2005 > 0.05, therefore we cannot reject the null hypothesis (i.e., the population mean is equal to 29.2) at α=0.05;

Hence, the data are consistent with a mean of 29.2.

Yes, the value 29.2 is inside the 95% confidence interval, since the confidence interval is (28.60544, 29.36123); We accept the null hypothesis, hence 29.2 must be inside the confidence interval. Therefore these conclusions are consistent with each other.

(c)

Since the P-value=0.2005 > 0.02, so we cannot reject the null hypothesis at α=0.2, i.e., 29.2 should be inside the 80% confidence interval, even though an 80% confidence interval is shorter than a 95% one.

> t.test(Weights, mu=29.2, conf.level=0.80)

One Sample t-test

data: Weights

t = -1.4739, df = 5, p-value = 0.2005

alternative hypothesis: true mean is not equal to 29.2

80 percent confidence interval:

28.76637 29.20030

sample estimates:

mean of x

28.98333

The 80% confidence interval is (28.76637, 29.20030), and apparently the value 29.2 is inside the confidence interval.

**Question4**

(a)

*CODE:*

>data snacks;

> infile '/home/shu.tu/snacks.txt' firstobs=2;

> input weights;

>

>proc print;

*RESULTS:*

| **Obs** | **weights** |
| --- | --- |
| **1** | 29.2 |
| **2** | 28.5 |
| **3** | 28.7 |
| **4** | 28.9 |
| **5** | 29.1 |
| **6** | 29.5 |

Yes, the results agree with R's.

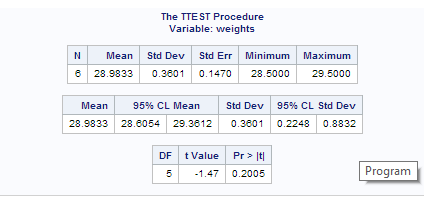
(b)

*CODE:*

>proc ttest h0=29.2;

> var weights;

*RESULTS:*



The P-value=0.2005, and the 95% confidence interval is (28.6054, 29.3612). Everything is the same as in R.

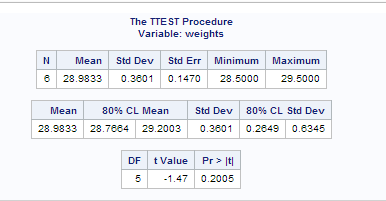
(c)

CODE:

>proc ttest h0=29.2 alpha=0.2;

> var weights;

RESULT:



The 80% confidence interval is (28.7664, 29.2003), which is the same as in R.

**Question 5**

(a)

> power.t.test(n=20, delta=130-125, sd=10, type="one.sample",alternative="two.sided")

One-sample t test power calculation

n = 20

delta = 5

sd = 10

sig.level = 0.05

power = 0.5644829

alternative = two.sided

The power is 0.56.

(b)

> power.t.test(powe=0.75, delta=130-125, sd=10, type="one.sample",alternative="two.sided")

One-sample t test power calculation

n = 29.73565

delta = 5

sd = 10

sig.level = 0.05

power = 0.75

alternative = two.sided

The sample size needed is 30.

(c)

In part (a), power=0.56 when n=20, and in part (b), the power increased to 0.75 and the sample size rose as well to 30. This result makes sense, because the power depends on the sample size, the lager the sample size, the more power we obtain.

(d)

*CODE:*

>proc power;

> onesamplemeans

> test=t

> sides=2

> mean=5

> stddev=10

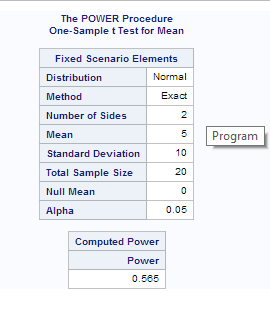
> ntotal=20

> power=.;

>

>run;

*RESULTS:*



The power is 0.565, same as in R.

*CODE:*

>proc power;

> onesamplemeans

> test=t

> sides=2

> mean=5

> stddev=10

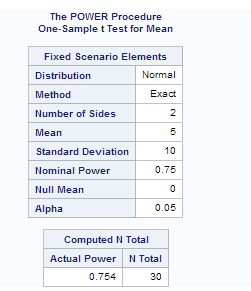
> ntotal=.

> power=0.75;

>

>run;

*RESULTS:*



The sample size equals to 30, as in R.

Therefore, the answers are consistent.

**Question 6**

(a)

> install.packages("BSDA")

> library(BSDA)

Loading required package: e1071

Loading required package: lattice

Attaching package: 態SDA?

The following object is masked from 憄ackage:datasets?

Orange

(I don't know why there is Chinese there....)

> head(Tablrock)

hour X03 tmp vdc wd ws amb dew so2 no no2 nox co co2 gas air

1 01:00 22 23.1 1.47 203 0.2 17.9 15.8 0.0 0.0 0.4 0.4 0.6 0.6 -1.5 -0.06

2 02:00 21 23.2 1.47 22 0.4 17.4 15.4 0.0 0.0 0.2 0.2 0.6 0.6 -1.5 -0.06

3 03:00 17 23.4 1.47 90 0.1 17.0 14.9 -0.1 0.1 0.1 0.2 0.6 0.6 -1.5 -0.06

4 04:00 13 23.6 1.47 50 1.3 16.2 14.4 -0.1 0.0 -0.1 -0.1 0.6 0.6 -1.5 -0.06

5 05:00 12 23.7 1.47 39 2.5 15.6 14.3 -0.1 0.1 -0.1 0.0 0.6 0.6 -1.5 -0.06

6 06:00 17 23.9 1.47 0 0.0 16.1 14.1 -0.1 0.1 0.0 0.1 0.6 0.6 -1.5 -0.06

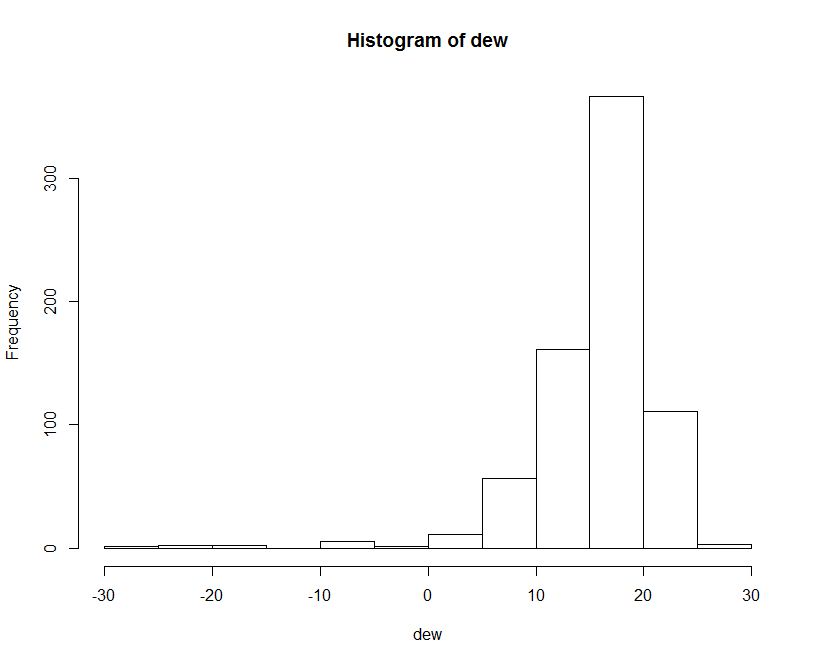
(b)

> attach(Tablrock)

The following object is masked from package:datasets:

co2

> hist(dew)



The shape is left-skewed. And there are some outliers in the data.

(c)

Because of the left skewness and the outliers in the data, median is better than mean. And since the sign test tests for median, while t-test tests the mean, therefore, a sign test should be more appropriate here.

(d)

> SIGN.test(dew, md=16)

One-sample Sign-Test

data: dew

s = 440, p-value = 9.494e-10

alternative hypothesis: true median is not equal to 16

95 percent confidence interval:

17.4 18.0

sample estimates:

median of x

17.7

Conf.Level L.E.pt U.E.pt

Lower Achieved CI 0.9476 17.4 18

Interpolated CI 0.9500 17.4 18

Upper Achieved CI 0.9561 17.4 18

Since the P-value is 9.494e-10 , which is very small, so we reject the null hypothesis, and we conclude that the median is not 16.

(e)

> t.test(dew, mu=16)

One Sample t-test

data: dew

t = 0.0515, df = 718, p-value = 0.9589

alternative hypothesis: true mean is not equal to 16

95 percent confidence interval:

15.59749 16.42420

sample estimates:

mean of x

16.01085

The P-value is 0.9589, which is big, therefore we cannot reject the null hypothesis. And we conclude that the mean is equal to 16.

(f)

The 95% confidence interval from the sign test is (17.4, 18.0) and the 95% confidence interval from the t-test is (15.59749, 16.42420). These two confidence intervals do not even overlap, therefore they are noticeably different.

**Question7**

(a)

*CODE:*

>data tablrock;

> infile '/home/shu.tu/tablrock.txt' firstobs=2;

> input hour x03 tmp vdc wd ws amb dew so2 no no2 nox co co2 gas air ;

>

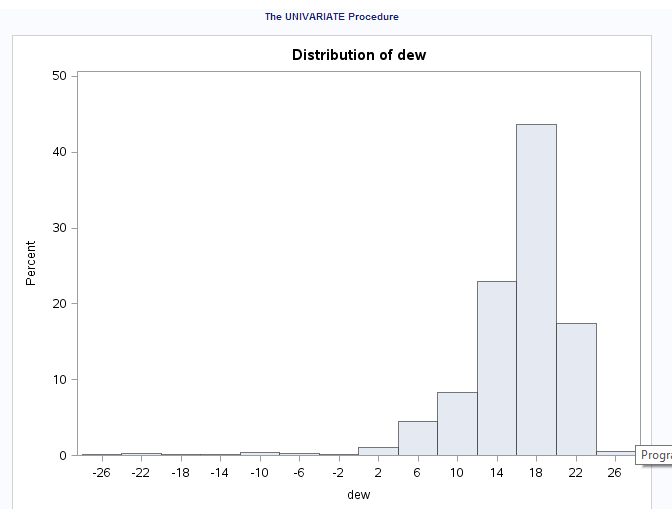
>

>proc univariate noprint;

> var dew;

> histogram;

*RESULTS:*



It is the same shape as with R: left-skewed with outliers.

(b)

*CODE:*

>data tablrock;

> infile '/home/shu.tu/tablrock.txt' firstobs=2;

> input hour x03 tmp vdc wd ws amb dew so2 no no2 nox co co2 gas air ;

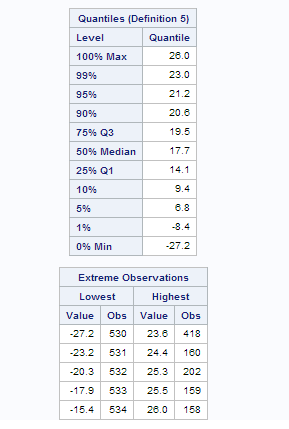
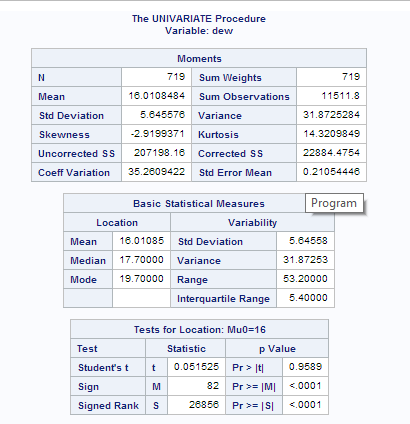
>

>

>proc univariate mu0=16;

> var dew;

*RESULTS:*



The P-value for sign test is less than 0.001, and the P-value for t-test is 0.9589. Both consistent with R.

**Question 8**

(a)

> skull=read.table("skull.txt", header=T)

> skull

X2000BCE X4000BCE

1 141 131

2 141 125

3 135 131

4 133 119

5 131 136

6 140 138

7 139 139

8 140 125

9 138 131

10 132 134

11 134 129

12 135 134

13 133 126

14 136 132

15 134 141

16 131 131

17 129 135

18 136 132

19 131 139

20 139 132

21 144 126

22 141 135

23 130 134

24 133 128

25 138 130

26 131 138

27 136 128

28 132 127

29 135 131

30 141 124

(b)

> skull2=stack(skull)

> skull2

values ind

1 141 X2000BCE

2 141 X2000BCE

3 135 X2000BCE

4 133 X2000BCE

5 131 X2000BCE

6 140 X2000BCE

7 139 X2000BCE

8 140 X2000BCE

9 138 X2000BCE

10 132 X2000BCE

11 134 X2000BCE

12 135 X2000BCE

13 133 X2000BCE

14 136 X2000BCE

15 134 X2000BCE

16 131 X2000BCE

17 129 X2000BCE

18 136 X2000BCE

19 131 X2000BCE

20 139 X2000BCE

21 144 X2000BCE

22 141 X2000BCE

23 130 X2000BCE

24 133 X2000BCE

25 138 X2000BCE

26 131 X2000BCE

27 136 X2000BCE

28 132 X2000BCE

29 135 X2000BCE

30 141 X2000BCE

31 131 X4000BCE

32 125 X4000BCE

33 131 X4000BCE

34 119 X4000BCE

35 136 X4000BCE

36 138 X4000BCE

37 139 X4000BCE

38 125 X4000BCE

39 131 X4000BCE

40 134 X4000BCE

41 129 X4000BCE

42 134 X4000BCE

43 126 X4000BCE

44 132 X4000BCE

45 141 X4000BCE

46 131 X4000BCE

47 135 X4000BCE

48 132 X4000BCE

49 139 X4000BCE

50 132 X4000BCE

51 126 X4000BCE

52 135 X4000BCE

53 134 X4000BCE

54 128 X4000BCE

55 130 X4000BCE

56 138 X4000BCE

57 128 X4000BCE

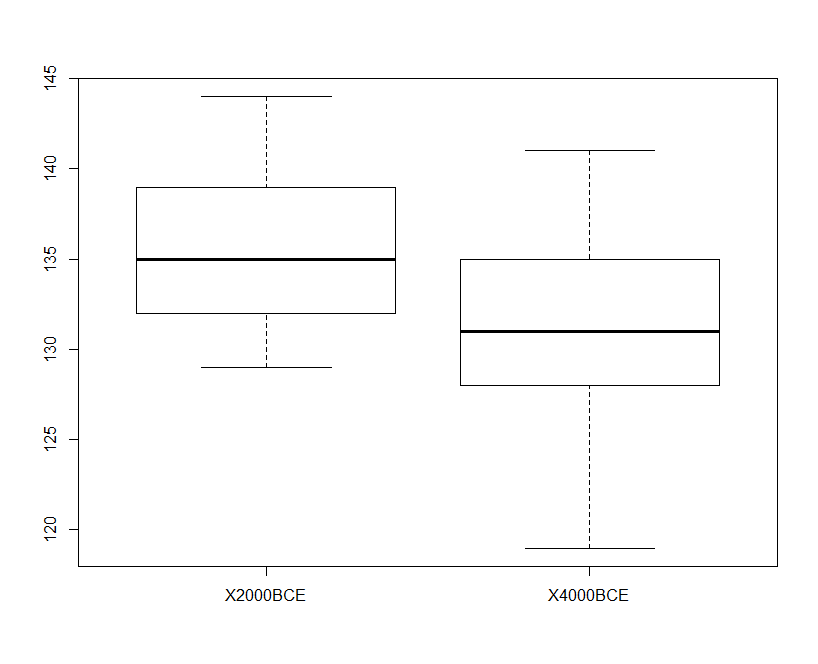
58 127 X4000BCE

59 131 X4000BCE

60 124 X4000BCE

(c)

> boxplot(values~ind, data=skull2)



The distribution for X2000BCE is slightly skewed to the right, and the distribution for X4000BCE is slightly skewed to the left, and neither of them has outliers, therefore I do not think this would impact the t-test.

(d)

> mean2=aggregate(values~ind, data=skull2, mean)

> mean2

ind values

1 X2000BCE 135.6333

2 X4000BCE 131.3667

(e)

> meandiff=mean2[1,2]-mean2[2,2]

> meandiff

[1] 4.266667

(f)

i.

> attach(skull2)

> myshuffle=sample(ind,length(ind))

> myshuffle

[1] X2000BCE X2000BCE X4000BCE X2000BCE X2000BCE X4000BCE X2000BCE X4000BCE X4000BCE X4000BCE X4000BCE X4000BCE X2000BCE

[14] X2000BCE X2000BCE X2000BCE X4000BCE X4000BCE X4000BCE X2000BCE X4000BCE X4000BCE X4000BCE X2000BCE X2000BCE X2000BCE

[27] X4000BCE X2000BCE X2000BCE X2000BCE X2000BCE X4000BCE X4000BCE X4000BCE X2000BCE X4000BCE X2000BCE X4000BCE X2000BCE

[40] X4000BCE X4000BCE X2000BCE X4000BCE X2000BCE X2000BCE X4000BCE X2000BCE X4000BCE X4000BCE X2000BCE X4000BCE X2000BCE

[53] X2000BCE X4000BCE X4000BCE X2000BCE X2000BCE X4000BCE X4000BCE X2000BCE

Levels: X2000BCE X4000BCE

ii.

> themeans=aggregate(values~myshuffle, data=skull2, mean)

> themeans

myshuffle values

1 X2000BCE 134.6

2 X4000BCE 132.4

iii.

> meandiff2=themeans[1,2]-themeans[2,2]

> meandiff2

[1] 2.2

The mean difference is 2.2, which is less than the observed difference in part (d).

(g)

> nism=10000

> ans=numeric(nsim)

> for (i in 1:10000)

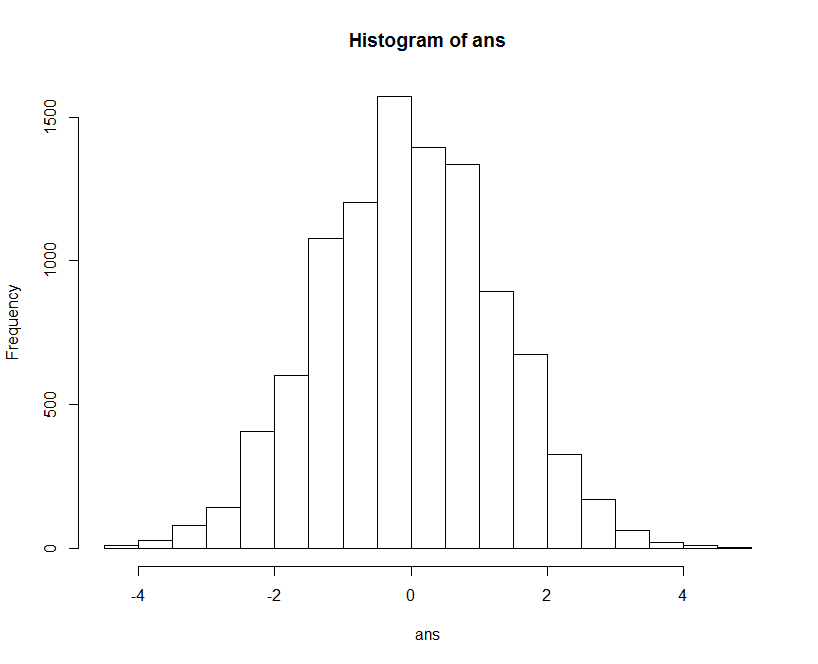
+ {

+ myshuffle=sample(ind, length(ind))

+ themeans=aggregate(values~myshuffle, data=skull2, mean)

+ ans[i]=themeans[1,2]-themeans[2,2]}

> hist(ans)



(h)

The observed difference in part (e) is 4.266667, which is on the right-most end(tail) of the distribution. therefore, mean difference of 4.26667 or bigger appears unlikely.

(i)

> isbig=table(ans>=4.266667)

> isbig

FALSE TRUE

9996 4

The P-value = 4/10000 = 0.0004, which is very small, therefore, we reject the null hypothesis, and conclude that the observed difference between skull breadths is significant.