Quantitative Research Methods:R_code_week 7

sn0wfree

11/10/2016

Contents

Quantitative Research Methods:R_Code_week 7

coding=UTF8

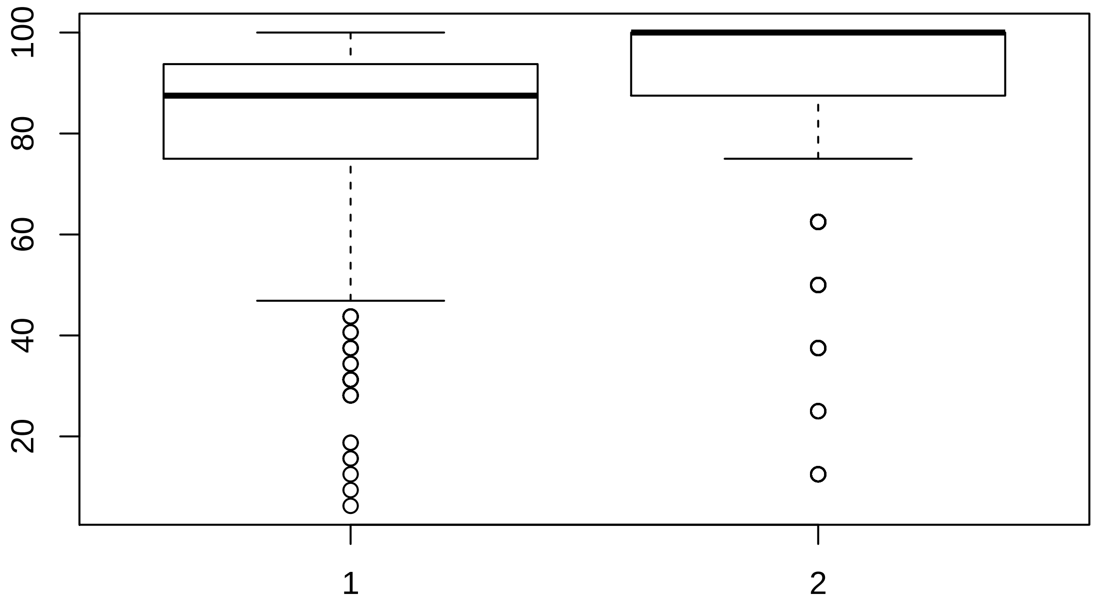
Import the data

```
#*coding==UTF8*
#dataset=attendSLIM.xls
#attendSLIM=read.table(pipe("pbpaste"),header=TRUE)
attendSLIM=read.csv("/Users/sn0wfree/Desktop/attendSLIM.csv",header=TRUE)
head(attendSLIM)
```

```
##   ACT priorGPA termGPA attendPC  hwPC year final
## 1  23     2.64    3.19   84.375 100.0    2    28
## 2  25     3.52    2.73   68.750  87.5    3    26
## 3  24     2.46    3.00   93.750  87.5    3    30
## 4  20     2.61    2.04   96.875 100.0    2    27
## 5  23     3.32    3.68  100.000 100.0    2    34
## 6  26     2.93    3.23   90.625 100.0    2    25
```

```
#par(mfrow=c(4,4)) #outframe set to 4*4
```

```
boxplot(attendSLIM$attendPC,attendSLIM$hwPC)
```



t-test and the related tests

t-test use to test the small sample(n<30), it is a method to test the different of the mean between two small sampl-size data. it use the T-dsitribution to estimate/predict the probability of difference among two samples, and evaluate the whether exist the significant difference between two mean.

two-tail test

```
t.test(attendSLIM$attendPC, attendSLIM$hwPC) #normal t-test
```

```
##
##  Welch Two Sample t-test
##
## data:  attendSLIM$attendPC and attendSLIM$hwPC
## t = -6.267, df = 1329.3, p-value = 4.964e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.138751 -4.258155
## sample estimates:
## mean of x mean of y
##  81.70956  87.90801
```

```
t.test(attendSLIM$attendPC, attendSLIM$hwPC, var.equal = TRUE) #t-test with var as equal # deprecated
```

```
##
##  Two Sample t-test
##
## data:  attendSLIM$attendPC and attendSLIM$hwPC
## t = -6.2704, df = 1352, p-value = 4.836e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -8.137673 -4.259233
## sample estimates:
## mean of x mean of y
##  81.70956  87.90801
```

```
t.test(attendSLIM$attendPC,attendSLIM$hwPC, paired = TRUE)#pair two sample to test whether the means ar
```

```
##
##  Paired t-test
##
## data:  attendSLIM$attendPC and attendSLIM$hwPC
## t = -9.5874, df = 673, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.971387 -4.601313
## sample estimates:
## mean of the differences
##                -5.78635
```

one-tail test

mu show the true mean

```
t.test(attendSLIM$attendPC) #test mu =0:mu show the ture mean in this one sample t-test
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC
## t = 124.99, df = 679, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  80.42600 82.99312
## sample estimates:
## mean of x
##  81.70956
```

```
t.test(attendSLIM$attendPC, mu =85) #test mu=85:mu show the ture mean in this one sample t-test
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC
## t = -5.0334, df = 679, p-value = 6.179e-07
## alternative hypothesis: true mean is not equal to 85
## 95 percent confidence interval:
##  80.42600 82.99312
## sample estimates:
## mean of x
##  81.70956
```

```r
t.test(attendSLIM$attendPC-85) # equivalent of t.test(attendSLIM$attendPC, mu =85)
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC - 85
## t = -5.0334, df = 679, p-value = 6.179e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.574001 -2.006881
## sample estimates:
## mean of x
## -3.290441
```

```r
t.test(attendSLIM$attendPC, mu =85, alternative = "less")# one tail test
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC
## t = -5.0334, df = 679, p-value = 3.09e-07
## alternative hypothesis: true mean is less than 85
## 95 percent confidence interval:
##      -Inf 82.78631
## sample estimates:
## mean of x
##  81.70956
```

```r
t.test(attendSLIM$attendPC, mu =82.99312)
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC
## t = -1.9635, df = 679, p-value = 0.05
## alternative hypothesis: true mean is not equal to 82.99312
## 95 percent confidence interval:
##  80.42600 82.99312
## sample estimates:
## mean of x
##  81.70956
```

```r
t.test(attendSLIM$attendPC, mu =81.70956)
```

```
##
##  One Sample t-test
##
## data:  attendSLIM$attendPC
## t = -1.7996e-06, df = 679, p-value = 1
## alternative hypothesis: true mean is not equal to 81.70956
## 95 percent confidence interval:
```

```
##  80.42600 82.99312
## sample estimates:
## mean of x
##  81.70956
```

Redo with another test: wilcox.test

the different of t-test and Wilcox.test−>t-test and wilcox.test<−Warning: this chinese website the data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

```r
wilcox.test(attendSLIM$attendPC,attendSLIM$hwPC, paired = 1) #1 means TRUE here
```

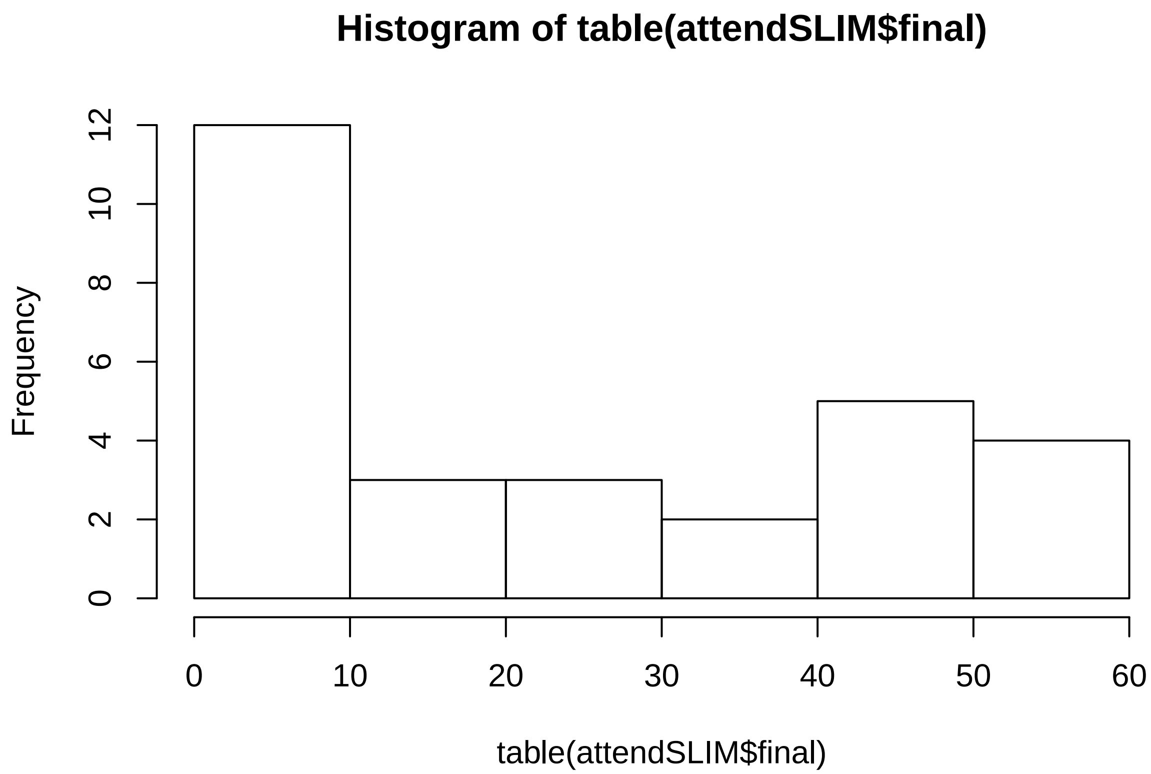
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  attendSLIM$attendPC and attendSLIM$hwPC
## V = 42922, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(attendSLIM$attendPC, mu =85)
```

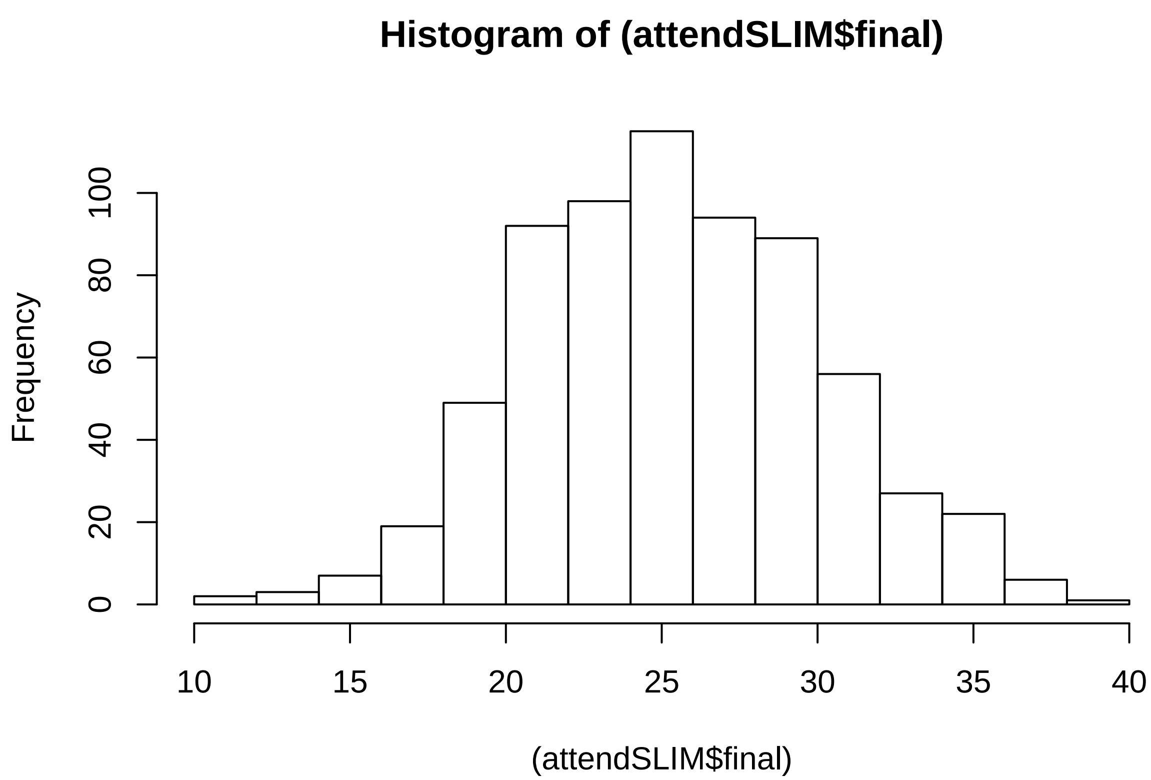
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  attendSLIM$attendPC
## V = 110120, p-value = 0.2698
## alternative hypothesis: true location is not equal to 85
```

plot residuals

```r
hist(table(attendSLIM$final))
```




```
hist((attendSLIM$final))#?
```



```r
table(attendSLIM$final>=25)
```

```
##
## FALSE  TRUE
##   270   410
```

```r
table(attendSLIM$final>=25,attendSLIM$year)
```

```
##
##           1   2   3
##   FALSE  79 146  45
##   TRUE   79 246  85
```

```r
chisq.test(table(attendSLIM$final>=25,attendSLIM$year))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(attendSLIM$final >= 25, attendSLIM$year)
## X-squared = 9.3925, df = 2, p-value = 0.009129
```

example tables and data for rank

table original

original|blue|black|black blue |sum ——-|—-|—-|———-|—- M | 20 | 30 | 70 | 120 F | 10 | 30 | 50 | 90 ——-|—-|—-|———-|—– sum | 30 |60 |120 |210

table expected

Expected	blue	black	black blue	sum
M	17.14285714	34.28571429	68.57142857	120
F	12.85714286	25.71428571	51.42857143	90
——-	—–	—–	———-	—–
sum	30	60	120	210

use $x^2 => (\sum(o-e)^2)$ to test different
degree of freedom:(row-1)(column-1)

and Hypothesis: h1: pass rate different between year and h0: pass rate same between year and

Regression

lm for linear model 1st: the simple regression on final to act

For exmaple $final = a + b * ACT$ hypothesis: H0: ACT can not predict the final:$b = 0$ H1: ACT can predict the final:$b! = 0$

```r
m1<-lm(attendSLIM$final~attendSLIM$ACT)
summary(m1)
```

```
##
## Call:
## lm(formula = attendSLIM$final ~ attendSLIM$ACT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1551  -2.8021  -0.1299   3.3072  12.3827
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.91953    1.10062   13.56   <2e-16 ***
## attendSLIM$ACT  0.48741    0.04832   10.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.395 on 678 degrees of freedom
## Multiple R-squared:  0.1305, Adjusted R-squared:  0.1292
## F-statistic: 101.8 on 1 and 678 DF,  p-value: < 2.2e-16
```

Correlation test

```r
cor.test(attendSLIM$final,attendSLIM$ACT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  attendSLIM$final and attendSLIM$ACT
## t = 10.088, df = 678, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2940498 0.4248936
## sample estimates:
##       cor
## 0.3612486
```

Draw the graph