Midterm Coursework

RyanYip

# Part 1

## 1a Descriptive statistics

## [1] "The Mean,Std Deviation of Democrat is :401.00,125.18"

## [1] "The Mean,Std Deviation of Republicant is :343.00,49.70"

The *mean* represents the central trend of the data, and the *standard deviation* represents the degree of fluctuation of the data.

|  |  |  |
| --- | --- | --- |
|  | **Mean** | **Std Deviation** |
| **Registered Democrats** | 401.00 | 125.18 |
| **Registered Republicans** | 343.00 | 49.70 |

## 1b Difference in means and confidence interval

According to the data above,I can conclude that this is a mean hypothesis testing Problem for two independent samples.And we use t-test method for this problem.

In statistics, we can define the corresponding null hypothesis (H0) as follow:

* H0: Mean of *Democrat* is equal Mean of *Republicant*
* H1: Mean of *Democrat* is not equal Mean of *Republicant*

And according to the t-test method,we will get the difference in means ,standard error of the difference in means and the confidence intervals of the difference in means.

Accoring the above analysis,we can know

* The difference in means between *registered Democrats* and *registered Republicans* is 58
* The standard error of the difference in means is 23.44
* The 95% confidence intervals of the difference in means is (11.16,104.84)
* The 99% confidence intervals of the difference in means is (-4.24,120.24)

## 1c Summary

According to the 1.b analysis,we have 95% confidence to we reject the null hypothesis,that is the number of people who support *Democrat* and support *Republicant* are not equal.But according to 99% confidence interval,we can not reject the null hypothesis.

And we will test which is more use the t-test again bellow.

t.test(Democrat,Republicant,alternative = "greater",conf.level = 0.95)

##   
## Welch Two Sample t-test  
##   
## data: Democrat and Republicant  
## t = 2.4739, df = 41.844, p-value = 0.008754  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 18.56342 Inf  
## sample estimates:  
## mean of x mean of y   
## 401 343

Based on the result of t-test with alternative hypothesis(Democrat > Republicant),the p-value is less than 0.05,so reject the null hypothesis,that is,in statistics means,we have 95% confidence to believe that the Democrathas more support.

# Part 2

### Read Data

## 'data.frame': 376 obs. of 9 variables:  
## $ code : int 416 417 267 92 98 423 418 268 219 160 ...  
## $ country : int 2 2 1 1 1 2 2 1 1 1 ...  
## $ pop\_density : int 2 1 2 1 1 1 1 2 2 1 ...  
## $ pmdeaths\_total : num 20484 15328 14930 16548 15268 ...  
## $ pmdeaths\_female: num 16368 13123 11442 13522 12273 ...  
## $ pmdeaths\_male : num 24600 17533 18418 19574 18262 ...  
## $ mean\_income : num 36950 37008 28537 31561 31396 ...  
## $ edu\_level3 : num 0.0964 0.1016 0.1118 0.1152 0.1199 ...  
## $ edu\_level4 : num 0.332 0.27 0.22 0.228 0.232 ...  
## - attr(\*, "na.action")=Class 'omit' Named int [1:2] 113 155  
## .. ..- attr(\*, "names")= chr [1:2] "113" "155"

## 2a Descriptive Statistics

We will use the mean ,median,min and max to measure the central tendency of variable,use the variance as measures of the dispersion following

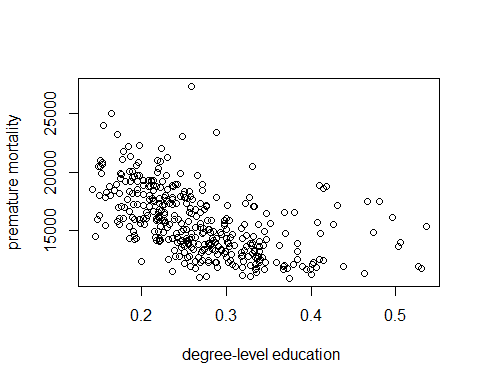
## [1] "The variable edu\_level3 --- mean:0.120286,median:0.119151,min:0.075525,max:0.192466,std:0.017279"

## [1] "The variable edu\_level4 --- mean:0.267991,median:0.256720,min:0.142148,max:0.535734,std:0.073995"

## [1] "The variable pop\_density --- mean:1.622340,median:1.500000,min:1.000000,max:3.000000,std:0.693537"

According to the variable describe and the visualization,the variances in *edu\_level3* and *edu\_level4* is small,which mean distribution is relatively stable.And the mean and median are almost similiar,which indicate they both have central trendancy.In variable *pop\_density*,we can the data mainly concentrate in 1 and 2,and the variance is relativey large.Above plot only show them in visualization.

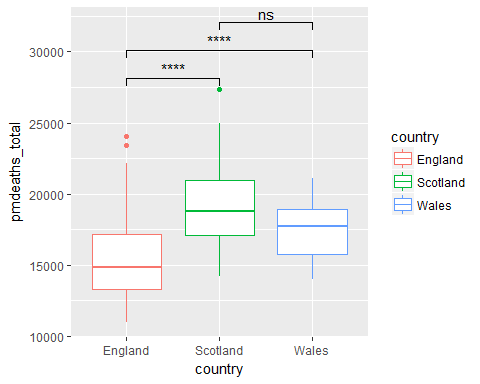
## 2b Visualization

* 2b.1) Scatterplot 

According to the scatterplot,we can know that *degree-level education* has a negative linear correlation with *premature mortality*,which means that when *degree-level education* increase,and premature mortality will decrease.And the correlation between they is -0.5198572,which can comfirm this.

* box plot

## Loading required package: magrittr

 The boxplot tells that the *premature mortality* in Scotland is the most highest,and Wales secondly,the England is the most smallest.And difference in Eangland and Scotland,England and Wales is sigficant.

## 2c Difference in Means

## The mean difference between premature mortality among men and women in Great Britain is : 6069.719

## Conduct t-test:

##   
## Welch Two Sample t-test  
##   
## data: pmdata$pmdeaths\_male and pmdata$pmdeaths\_female  
## t = 28.245, df = 647.34, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 5715.739 Inf  
## sample estimates:  
## mean of x mean of y   
## 18763.22 12693.50

In the test,I use the alternative hypothesis is : male is greater than femele.According to the t-test result,we know that the p-value is smaller than 0.05,which means that we has 95% confidence to believe thaht *premature mortality* among men and women in Great Britain is different in statistics.In fact ,it is consistent with commen sense,in many surveys, men’s lives are indeed shorter than women’s.

## 2d Linear Regression

##   
## Call:  
## lm(formula = pmdeaths\_total ~ mean\_income, data = pmdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5589.1 -1838.1 -389.1 1382.8 11010.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.107e+04 5.147e+02 40.94 <2e-16 \*\*\*  
## mean\_income -1.590e-01 1.484e-02 -10.71 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2475 on 374 degrees of freedom  
## Multiple R-squared: 0.2348, Adjusted R-squared: 0.2327   
## F-statistic: 114.8 on 1 and 374 DF, p-value: < 2.2e-16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 2.107e+04 | 515 | 40.94 | <2e-16 |
| mean\_income | -1.60 | 0.0015 | -10.71 | <2e-16 |

Accoring to the model summary,the coefficient of *mean\_income* is -0.16,and the test is sinificant,which indicate that mean\_income has a negative effect on *pmdeaths\_total*,when *mean\_income* add one unit,*pmdeaths\_total* will substract about 0.16 unit.And it is credible in statistics based on test.

The coefficient of intercept is 2.107e+04,the significant test for intercept indicates it is necessary to add in model.The coefficient of intercept tells us when the predictor(*mean\_income*) is equal to zero,the *pmdeaths\_total* will be a constant,and is equal to 2.107e+04

The *R square* of mdoel is about 0.23 ,it is a litte smaller in statistics,which means that the predictor *mean\_income* can explain 23% message in *pmdeaths\_total*,and it represent the goodness of fit for model.So,only use *mean\_income* as predictor is not enough to explain the model I think.

# Part 3

## 3a Multiple Linear Regression

Accoring to the findings of two linear regression models from table 1,we know that,in model 1 and model 2, they has common predictor *salary*.In model 1, the coefficient of *salary* is -0.56,p-value is 0.10,which means *salary* has negative effect on percentage of legislators.When salary add one unit,the *percentage of legislators* substract 0.56 unit.But this variable is not significant in model in statistics.As we know,the salary actually effect the *percentage of legislators* in substantive life,may be it is not consistent?.And in model 2 ,the coefficient of *salary* is -0.61,p-value is 0.13,the explaition is same as in model 1.

In model 2 result,only *pct\_union,pct\_black,pct\_urban* are significant,the estimated coefficient are relatively 0.12,-0.06,-0.03,which indicates that,when *pct\_union* add one unit,*percentage of legislators* will add 0.12 unit;when *pct\_black* add one unit,*percentage of legislators* will add -0.06 unit;when *pct\_urban* add one,*percentage of legislators* will add -0.03.It is consistent that *percentage of workers* belonging to a labour union has a positive effect on *percentage of legislators* in real life;due to racial discrimination and the gap between rich and poor in real life, *percentage of state residents living in urban areas* and *percent of state residents living below the poverty line* actually have negative effect on *percentage of legislators*.All of these significant variable’s hypothesis is established.

The *R square* in model 1 is about 0.18,in model 2 is about 0.35,which mean that salary in model only can explain 18% message on *percentage of legislators*.And a number of variables in model can explain 35% message on *percentage of legislators*.And model 2 has better goodness of fit of model than model 1.And I think model 2 can continue to improve and perform better,because in this model,still have many insinificant variale.

# Appendix section

Code:

# Part 1

### Dataset

Democrat<-c(170, 448, 369, 182, 394, 206, 258, 433, 503, 426, 409, 421, 355, 516, 226, 535, 489,

337, 464, 508, 325, 521, 533, 533, 237, 476, 312, 493, 411, 464, 192, 689, 398)

Republicant<-c(377, 373, 284, 249, 402, 336, 377, 263, 352, 341, 290, 365, 416, 299, 334, 353, 363,346, 285, 371, 387, 333, 431, 349, 307, 294, 374, 254, 366, 416, 274, 343, 415)

## 1a Descriptive statistics

sprintf("The Mean,Std Deviation of Democrat is :%f,%f",mean(Democrat),sd(Democrat))

sprintf("The Mean,Std Deviation of Republicant is :%f,%f",mean(Republicant),sd(Republicant))

## 1b Difference in means and confidence interval

Md=mean(Democrat)

Mr=mean(Republicant)

nd=length(Democrat)

nr=length(Republicant)

df=nr+nd-2

S\_square=(sum((Democrat-Md)^2)+sum((Republicant-Mr)^2))/(nd+nr-2)

Mean\_diff=Md-Mr

std\_mean\_diff=sqrt(S\_square/nd+S\_square/nr)

q\_alpha\_95=qt(1-0.05/2,df=df)

q\_alpha\_99=qt(1-0.01/2,df=df)

lower\_95=Mean\_diff-q\_alpha\_95\*std\_mean\_diff

upper\_95=Mean\_diff+q\_alpha\_95\*std\_mean\_diff

lower\_99=Mean\_diff-q\_alpha\_99\*std\_mean\_diff

upper\_99=Mean\_diff+q\_alpha\_99\*std\_mean\_diff

## 1c Summary

t.test(Democrat,Republicant,alternative = "greater",conf.level = 0.95)

# Part 2

### Read Data

pmdata <- read.csv("./pmgb2012\_2014.csv",header = TRUE)

pmdata<-na.omit(pmdata) # remove NA value if has

pmdata<-pmdata[!duplicated(pmdata),] # remove duplicate data if has

str(pmdata)

edu\_level3<-pmdata$edu\_level3

edu\_level4<-pmdata$edu\_level4

pop\_density<-pmdata$pop\_density

describe<-function(variable,name){

Mean<-mean(variable)

Med<-median(variable)

Range<-range(variable)

Variance<-var(variable)

sprintf("The variable %s --- mean:%f,median:%f,min:%f,max:%f,variance:%f",name,

Mean,Med,Range[1],Range[2],Variance)

}

describe(edu\_level3,name="edu\_level3")

describe(edu\_level4,name="edu\_level4")

describe(pop\_density,name="pop\_density")

plot(density(edu\_level3),col='red',main="edu\_level3")

boxplot(edu\_level3,main="edu\_level3")

plot(density(edu\_level4),col='red',main="edu\_level4")

boxplot(edu\_level4,main="edu\_level4")

prop.table(table(pop\_density))

boxplot(pop\_density,main="pop\_density")

## 2b Visualization

### Scatterplot

plot(x=pmdata$edu\_level4,y=pmdata$pmdeaths\_total,xlab=" degree-level education",ylab ="premature mortality" )

### box plot

library(ggplot2)

library(ggpubr)

country=pmdata$country

country=ifelse(country==1,"England",ifelse(country==2,"Scotland","Wales"))

df=data.frame(pmdeaths\_total=pmdata$pmdeaths\_total,country=as.factor(country))

ggplot(data=df,aes(x=country,y=pmdeaths\_total,color=country))+geom\_boxplot()+

stat\_compare\_means(comparisons = list(c("England","Scotland"),c("England","Wales"),c("Scotland","Wales")),

aes(label=..p.signif..))

## 2c Difference in Means

mean\_delta=mean(pmdata$pmdeaths\_male)-mean(pmdata$pmdeaths\_female)

cat("The mean difference between premature mortality among men and women in Great Britain is :",mean\_delta,"\n")

cat("Conduct t-test: \n")

t.test(pmdata$pmdeaths\_male,pmdata$pmdeaths\_female,conf.level = 0.95,alternative = "greater")

## 2d Linear Regression

model<-lm(pmdeaths\_total~mean\_income,data = pmdata)

summary(model)