STAT 477/577 - Technology Guide

Module 2 - Section 2

Comparing Two or More Population Proportions

Below is an explanation of the R commands and functions needed to conduct the analyses on population proportions and multinomial distributions.

- **Inference for Difference in Two Proportions (Part A)**

  We will begin by reading in the doctor's survey data and set the baseline category for both variables to `No`.

```
survey.data<- read.csv(file.choose(), header = T)
```

```
survey.data$Receive.Letter<- factor(survey.data$Receive.Letter,
         levels = c("Yes", "No"))

survey.data$Return.Survey<- factor(survey.data$Return.Survey,
         levels = c("Yes", "No"))
```

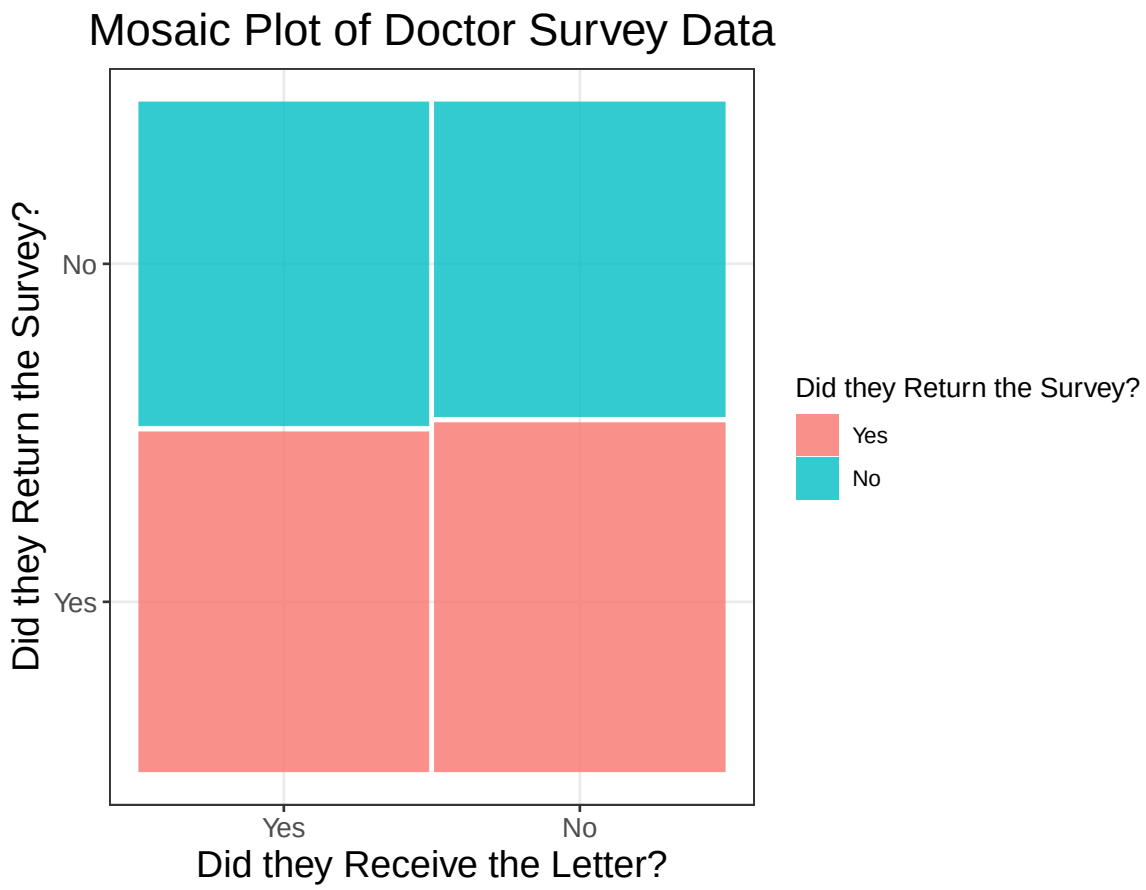
  We will then obtain a contingency table for the two variables along with a mosaic plot.

```
survey.table<- table(survey.data$Receive.Letter,
                     survey.data$Return.Survey)
survey.table
```

```
##
##        Yes   No
##   Yes 2570 2448
##   No  2645 2384
```

```
  ggplot(data = survey.data)+
    geom_mosaic(aes(x = product(Return.Survey, Receive.Letter),
                    fill = Return.Survey),
                na.rm = TRUE, divider =  mosaic("h"))+
    theme_bw()+
    theme(plot.title = element_text(hjust=0.5, size = rel(1.6)),
          axis.title.y = element_text(size = rel(1.4)),
          axis.title.x = element_text(size = rel(1.4)),
          axis.text.x = element_text(size = rel(1.2)),
          axis.text.y = element_text(size = rel(1.2)),
          strip.text.y = element_text(size = rel(1.2)))+
    labs(x = "Did they Receive the Letter?",
```

```
        y = "Did they Return the Survey?",
        fill = "Did they Return the Survey?",
        title = "Mosaic Plot of Doctor Survey Data")
```



The number of surveys returned from the letter and non-letter groups are in the first column of the contingency table. We also need to obtain the total number of observations in each group, which is the row margin of the contingency table.

```
groupsize<- margin.table(survey.table, 1)
```

To obtain a hypothesis test for the equality of two proportions in R, the function is `prop.test`. You will need to provide the number of successes `y` and sample size `n` for the two groups, along with an alternative (either `two.sided`, `greater`, or `less`). Here is the general command:

```
prop.test(y, n, alternative = "two.sided", correct = F)
```

For example, for the study of the effectiveness of a letter of introduction on response rates for a survey of doctors, the command is:

```
prop.test(survey.table[,1], groupsize,
          alternative = "two.sided", correct = F)


##
##  2-sample test for equality of proportions without continuity correction
##
## data:  survey.table[, 1] out of groupsize
## X-squared = 1.9143, df = 1, p-value = 0.1665
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.033330958  0.005744447
## sample estimates:
##    prop 1    prop 2
## 0.5121562 0.5259495
```

The value of the z test statistic from class is the square root of the `X-squared` value given in the output. The sign of the z test statistic is the same as the sign of the quantity $\hat{p}_1 - \hat{p}_2$.

Confidence intervals for the difference in two proportions can be handled in a similar manner. First, we will read in the data and set the baseline categories for both variables to `No`.

```
surgery.data<- read.csv(file.choose(), header = T)
```

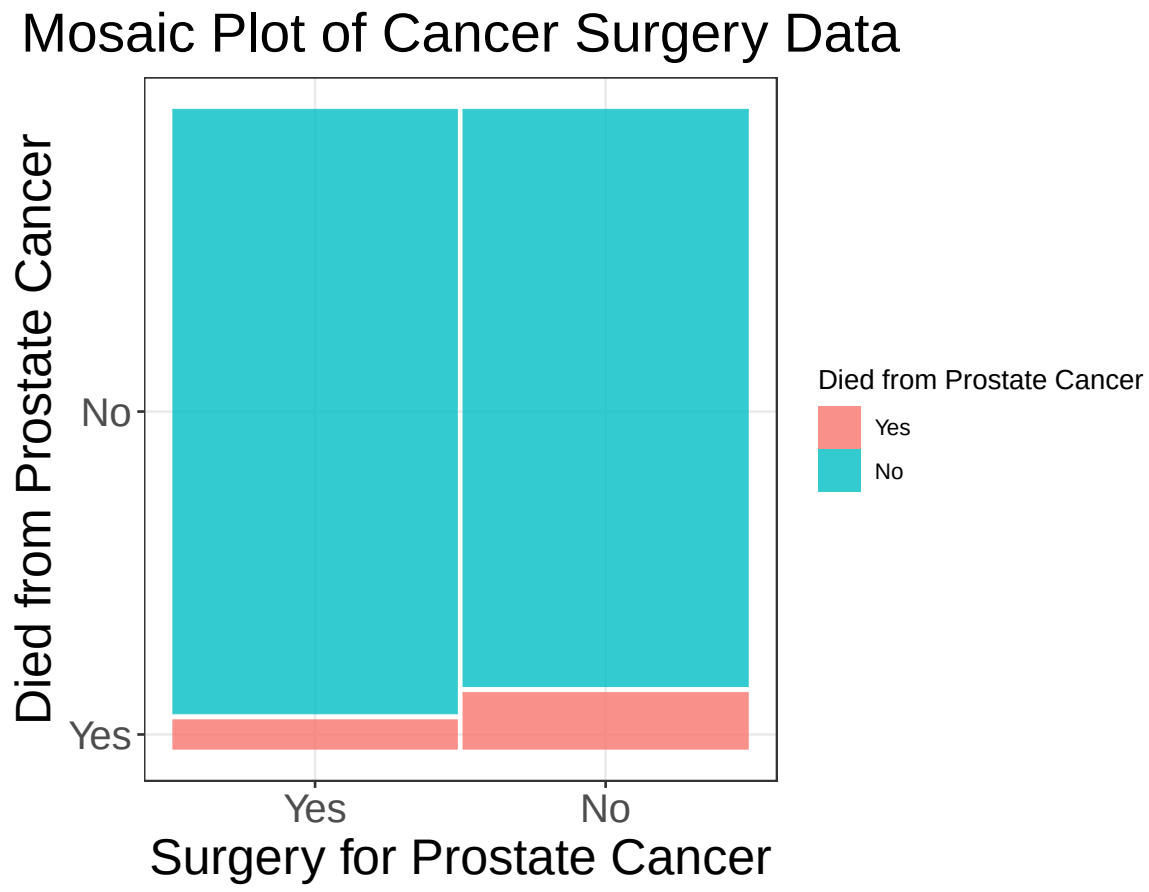
```
surgery.data$Surgery<- factor(surgery.data$Surgery,
                                      levels = c("Yes", "No"))
surgery.data$Died<- factor(surgery.data$Died,
                                     levels = c("Yes", "No"))
```

We will then obtain the contingency table for the two variables along with a mosaic plot.

```
surgery.table<- table(surgery.data$Surgery,
                      surgery.data$Died)
surgery.table


##
##       Yes  No
##   Yes  16 331
##   No   31 317
```

```
ggplot(data = surgery.data)+
  geom_mosaic(aes(x = product(Died, Surgery), fill = Died),
                na.rm = TRUE, divider =  mosaic("h"))+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5, size = rel(2)),
        axis.title.y = element_text(size = rel(1.8)),
        axis.title.x = element_text(size = rel(1.8)),
        axis.text.x = element_text(size = rel(1.8)),
        axis.text.y = element_text(size = rel(1.8)),
        strip.text.y = element_text(size = rel(1.8)))+
  labs(x = "Surgery for Prostate Cancer",
       y = "Died from Prostate Cancer",
       fill = "Died from Prostate Cancer",
       title = "Mosaic Plot of Cancer Surgery Data")
```



The total number of observations in each group is:

```
groupsize<- margin.table(surgery.table, 1)
```

We will then use the function `prop.test` to calculate the confidence interval in the difference in the proportion of patients who died from prostate cancer between those who had surgery and those that did not have surgery as:

```
prop.test(surgery.table[,1], groupsize, alternative = "two.sided",
          conf.level = 0.95, correct = F)


##
##  2-sample test for equality of proportions without continuity correction
##
## data:  surgery.table[, 1] out of groupsize
## X-squared = 5.0883, df = 1, p-value = 0.02409
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.080154987 -0.005786913
## sample estimates:
##     prop 1     prop 2
## 0.04610951 0.08908046
```

- **Sample Size Calculations (Part A)**

  There are no built-in functions for calculating the sample size necessary to obtain a certain power level for a hypothesis test or a certain margin of error for a confidence interval. I wrote two functions, `n2prop.test` and `n2prop.ci`, that will perform the sample size calculations from the lecture notes.

  To determine the sample size needed from each group to obtain a certain power for a hypothesis test, the general command is:

  `n2prop.test(p1, p2, alternative, alpha, power)`

  where `p1` and `p2` are the estimated values of the population proportions from each group, `alternative` is the direction of the alternative hypothesis (either `two.sided`, `greater`, `less`), `alpha` is the Type I error rate and `power` is the desired power of the hypothesis test for the given difference between `p1` and `p2`.

  For the doctors survey example, the sample size calculation from the lecture notes is:

```
n2prop.test(p1 = 0.55, p2 = 0.5, alternative = "two.sided",
            alpha = 0.05, power = 0.9)


## [1] 2095
```

  To determine the sample size needed from each group to obtain a certain margin of error for a confidence interval, the general command is:

  `n2prop.ci(p1, p2, m, conf.level)`

where `p1` and `p2` are the estimated values of the sample proportions from each group, `m` is the desired margin of error for the confidence interval and `conf.level` is the confidence level for the interval as a decimal.

For the prostate cancer surgery example, the sample size calculation from the lecture notes is:

```
n2prop.ci(p1 = 0.05, p2 = 0.09, m = 0.03, conf.level = 0.95)

## [1] 553
```

- **Inference for Equality of Multiple Proportions (Part B)**

  The hypothesis test for the equality of multiple proportions works in much the same way as the analysis above. First, we read in the data from the file **diodedata.csv** and set the baseline category for the response variable `Status` to be `Non-Conforming` and set the group variable `Lot` to be categorical.

```
diode.data<- read.csv(file.choose(), header = T)
```

```
diode.data$Status<- factor(diode.data$Status,
                    levels = c("Non-Conforming", "Conforming"))
diode.data$Lot<- as.factor(diode.data$Lot)
```

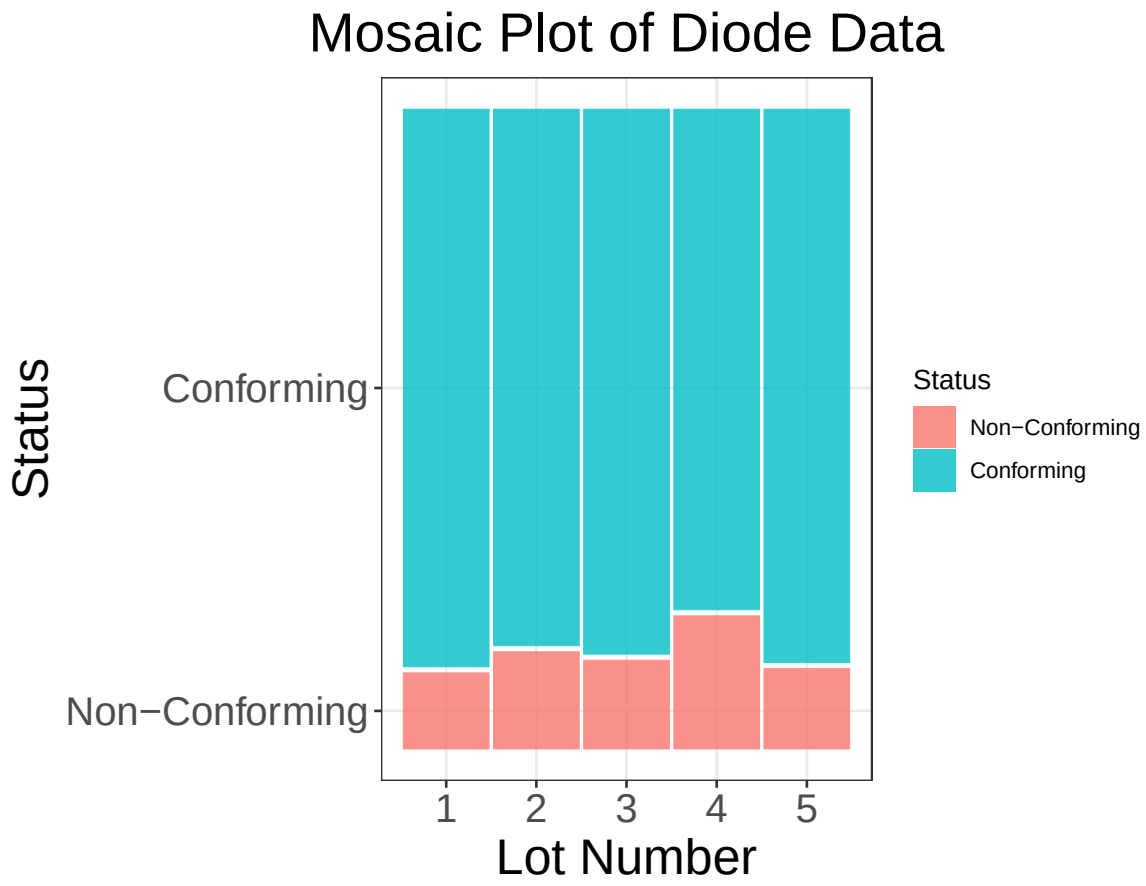
  Then we will use R to obtain the contingency table for the two variables and graph the mosaic plot.

```
diode.table<- table(diode.data$Lot, diode.data$Status)
diode.table

##
##     Non-Conforming Conforming
##   1             36        264
##   2             46        254
##   3             42        258
##   4             63        237
##   5             38        262
```

```
ggplot(data = diode.data)+
  geom_mosaic(aes(x = product(Status, Lot), fill = Status),
                na.rm = TRUE, divider =  mosaic("h"))+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5, size = rel(2)),
```

```
        axis.title.y = element_text(size = rel(1.8)),
        axis.title.x = element_text(size = rel(1.8)),
        axis.text.x = element_text(size = rel(1.8)),
        axis.text.y = element_text(size = rel(1.8)),
        strip.text.y = element_text(size = rel(1.8)))+
  labs(x = "Lot Number",
       y = "Status",
       fill = "Status",
       title = "Mosaic Plot of Diode Data")
```



The total number of observations in each lot is:

```
lotsize<- margin.table(diode.table, 1)
```

We can conduct the test of equality using the function `prop.test`. For the test for the equality of proportions for the diodes data, the command is:

```
prop.test(diode.table[,1], lotsize)

##
```

```
##  5-sample test for equality of proportions without continuity correction
##
## data:  diode.table[, 1] out of lotsize
## X-squared = 12.131, df = 4, p-value = 0.01641
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5
## 0.1200000 0.1533333 0.1400000 0.2100000 0.1266667
```

You can use the function `pairwise.prop.test` to study the pairwise comparisons of the proportions between the groups. The general command is:

`pairwise.prop.test(y, n, p.adjust.method = "BH")`

To control for multiple comparisons, the Benjamini & Hochberg or "BH" method is used. This method controls for the false discovery rate - the expected proportion of false discoveries (wrong decisions) among the tests which reject the null hypotheses.

The command for the diode data from the lecture notes is:

```
pairwise.prop.test(diode.table[,1], lotsize,
                   p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Pairwise comparison of proportions
##
## data:  diode.table[, 1] out of lotsize
##
##   1     2     3     4
## 2 0.570 -     -     -
## 3 0.777 0.810 -     -
## 4 0.042 0.226 0.105 -
## 5 0.901 0.684 0.810 0.044
##
## P value adjustment method: BH
```

- **Inference for Equality of Multiple Multinomial Responses (Part C)**

  To analyze the smoking survey data, we will first need to read in the data from the file **smokingsex.csv**.

```
smoke.data<- read.csv(file.choose(), header = T)
```

The categories for the variable `Smoke` are NonSmoker, PastSmoker, CurrentSmoker. Given the definitions of these categories, it would make sense to present them in their logical order.

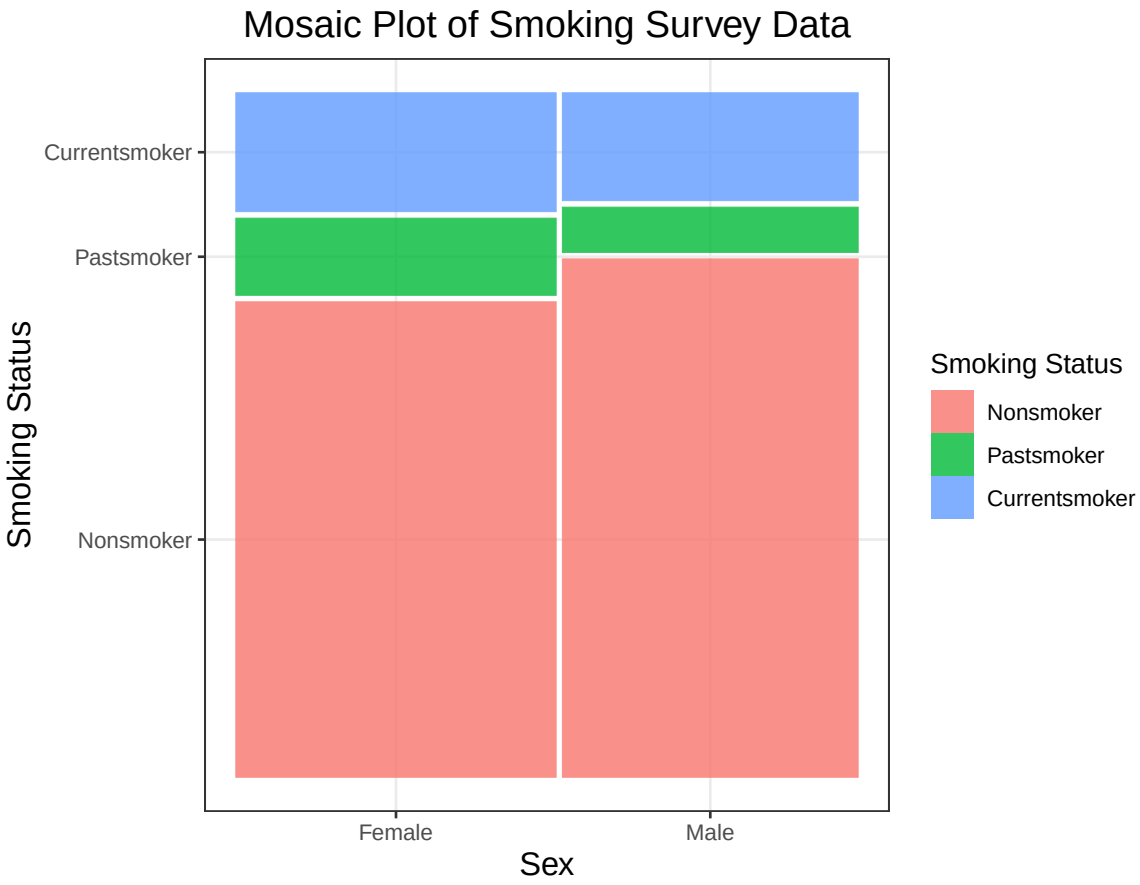
```
smoke.data$Smoke<- factor(smoke.data$Smoke, levels =
                            c("Nonsmoker", "Pastsmoker", "Currentsmoker"))
```

Then we can use R to calculate the contingency table for the two variables and graph the mosaic plot.

```
smoke.table<- table(smoke.data$Sex, smoke.data$Smoke)
smoke.table
```

```
##
##          Nonsmoker Pastsmoker Currentsmoker
##   Female       148         24            37
##   Male         149         13            31
```

```
ggplot(data = smoke.data)+
  geom_mosaic(aes(x = product(Smoke, Sex), fill = Smoke),
                na.rm = TRUE, divider =  mosaic("h"))+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5, size = rel(1.4)),
        axis.title.y = element_text(size = rel(1.2)),
        axis.title.x = element_text(size = rel(1.2)),
        axis.text.x = element_text(size = rel(1)),
        axis.text.y = element_text(size = rel(1)),
        strip.text.y = element_text(size = rel(1.2)))+
  labs(x = "Sex",
       y = "Smoking Status",
       fill = "Smoking Status",
       title = "Mosaic Plot of Smoking Survey Data")
```



When the response variable has more than two categories and we are comparing the multinomial distributions between groups, we will need to use the function `chisq.test` to obtain the hypothesis test for the equality of the multinomial distributions.

```
smoke.test<- chisq.test(smoke.table)
smoke.test

##
##  Pearson's Chi-squared test
##
## data:  smoke.table
## X-squared = 3.1713, df = 2, p-value = 0.2048
```

By saving the test to the variable `smoke.test`, we can obtain more information about the test, including the estimated expected values for each cell and the contribution of each cell to the test statistic.

```
smoke.test$expected

##
```

```
##          Nonsmoker Pastsmoker Currentsmoker
##   Female  154.4104   19.23632      35.35323
##   Male    142.5896   17.76368      32.64677
```

```
(smoke.test$residual)^2
```

```
##
##           Nonsmoker Pastsmoker Currentsmoker
##   Female 0.26613381 1.17967804    0.07670695
##   Male   0.28819671 1.27747518    0.08306608
```