Study the Choice of Contraceptive Methods in Women

Youqing Xiang

December 5, 2015

Part 1 - Introduction

Contraceptive method choice is always an interesting topic, especially after women getting married. Contraceptive methods give the women choice whether or when they want to get pregnant and women get health and social beneftis from planned pregnancies. However, how women choose what kind of contraceptive methods is very complicated. This project is focusing on studying the relationship between the contraceptive method choice of a married woman in Indonesia and her demographic and socio-economic characteristics and then trying to build a model to predict her choice.

Part 2 - Data

Data collection Here is the link for the data source. This is a subset data of the 1987 National Indonesia Contraceptive Prevalence Survey.

Cases There are 1473 cases. The samples are married women who were either not pregnant or do not know if they were at the time of interview.

Variables **The explanatory variables are:**

- Wife's age: numerical data type
- Wife's education: categorical data type; 1=low and 2,3,4=high
- Husband's education: categorical data type; 1=low and 2,3,4=high
- Number of children ever born: numerical data type
- Wife's religion: binary data type; 0=Non-Islam, 1=Islam
- Wife's now working: binary data; type 0=Yes, 1=No
- Husband's occupation: categorical data type; 1=low and 2,3,4=high
- Standard-of-living index: categorical data type; 1=low and 2,3,4=high
- Meia exposure: binary data type; 0=Good, 1=Not good

The response variable is: contraceptive method used and the data type is categorical; 1=No-use, 2=Long-term, 3=Short-term

Type of study The type of study is an observational study. In order to understand the relationships between explanatory variables and reponse variable, I study the dataset collected by sampling population in three different ways: explanatory data analysis with both summary statistics and data visulization, inference analysis and applying for Machine learning knowledge.

Scope of inference - generalizability This study is interested in what factors contribute to the choice of married women's contraceptive methods in Indonesia. Since the data were sampled by national survey and there were 1473 cases, it is reasonable to generalize the conclusion to all the married women in Indonesia. However, we should realize that the data was collected in 1987 and should be very cautious if we want to generalize the conclusion up to today's married women in Indonesia. We can get useful information from this study, but it is not reasonable to generalize the conclusion from this study to all the married women around the whole world.

Scope of inference - causality Observation study is used to look for the associations between the variables of interest. We generally don't make causual conclusions based on observational data. If we want to investigate the possibility of a causal connection, we should conduct an experiment instead of observation study.[1]

Part 3 - Exploratory data analysis

```
colnames(cmc_data) <- c('Wife_age',
                        'Wife_education',
                        'Husband_education',
                        'Number_children',
                        'Wife_religion',
                        'Wife_working',
                        'Husband_occupation',
                        'Standard_living_index',
                        'Media_exposure',
                        'Contraceptive_method')
cmc_data$Wife_education <- as.factor(cmc_data$Wife_education)
cmc_data$Husband_education <- as.factor(cmc_data$Husband_education)
cmc_data$Wife_religion <- as.factor(cmc_data$Wife_religion)
levels(cmc_data$Wife_religion) <- c('Non-Islam', 'Islam')
cmc_data$Wife_working <- as.factor(cmc_data$Wife_working)
levels(cmc_data$Wife_working) <- c('Yes', 'No')
cmc_data$Husband_occupation <- as.factor(cmc_data$Husband_occupation)
cmc_data$Standard_living_index <- as.factor(cmc_data$Standard_living_index)
cmc_data$Media_exposure <- as.factor(cmc_data$Media_exposure)
levels(cmc_data$Media_exposure) <- c('Good', 'Not good')
cmc_data$Contraceptive_method <- as.factor(cmc_data$Contraceptive_method)
levels(cmc_data$Contraceptive_method) <- c('No-use', 'Long-term', 'Short-term')
```

Data Transformation

```
summary(cmc_data)
```

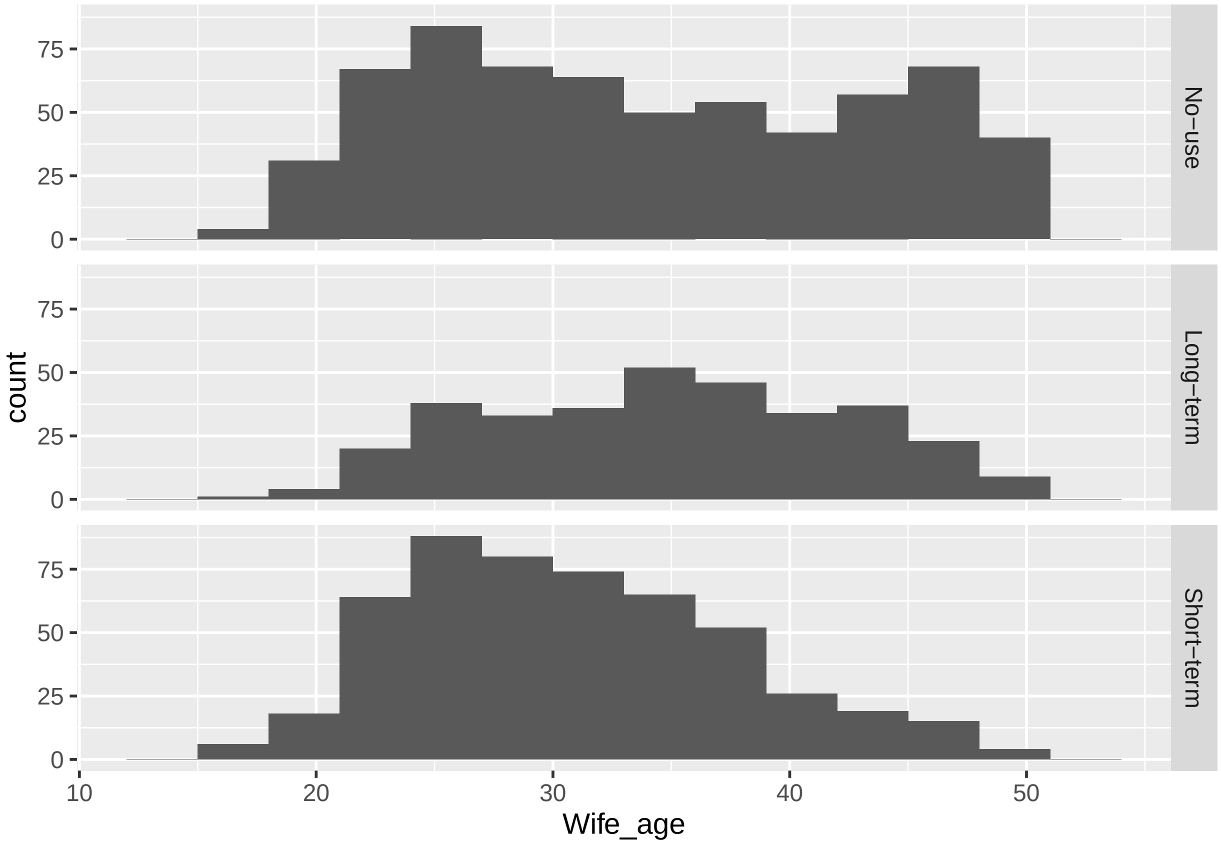
Summary of data set

```
##     Wife_age     Wife_education Husband_education Number_children
##  Min.   :16.00   1:152          1: 44             Min.   : 0.000
##  1st Qu.:26.00   2:334          2:178             1st Qu.: 1.000
##  Median :32.00   3:410          3:352             Median : 3.000
##  Mean   :32.54   4:577          4:899             Mean   : 3.261
##  3rd Qu.:39.00                                    3rd Qu.: 4.000
##  Max.   :49.00                                    Max.   :16.000
##    Wife_religion  Wife_working Husband_occupation Standard_living_index
##  Non-Islam: 220   Yes: 369     1:436              1:129
##  Islam    :1253   No :1104     2:425              2:229
##                                3:585              3:431
##                                4: 27              4:684
```

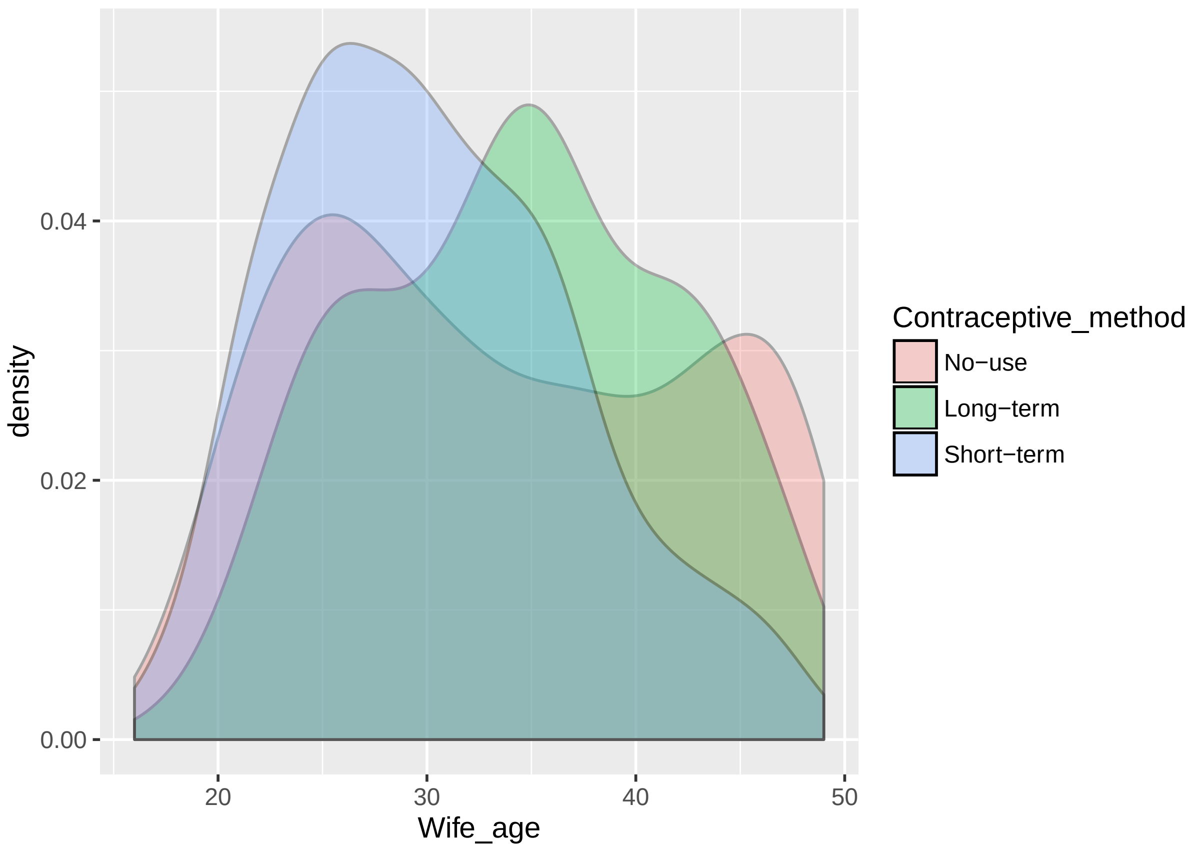
```
##
##
##   Media_exposure Contraceptive_method
##  Good    :1364   No-use    :629
##  Not good: 109   Long-term :333
##                  Short-term:511
##
##
##
```

Data visualization   1. The relationship between wife's age and contraceptive method used

```r
library(ggplot2)
ggplot(data=cmc_data,aes(x=Wife_age)) + geom_histogram(binwidth=3) +
  facet_grid(Contraceptive_method ~ .)
```

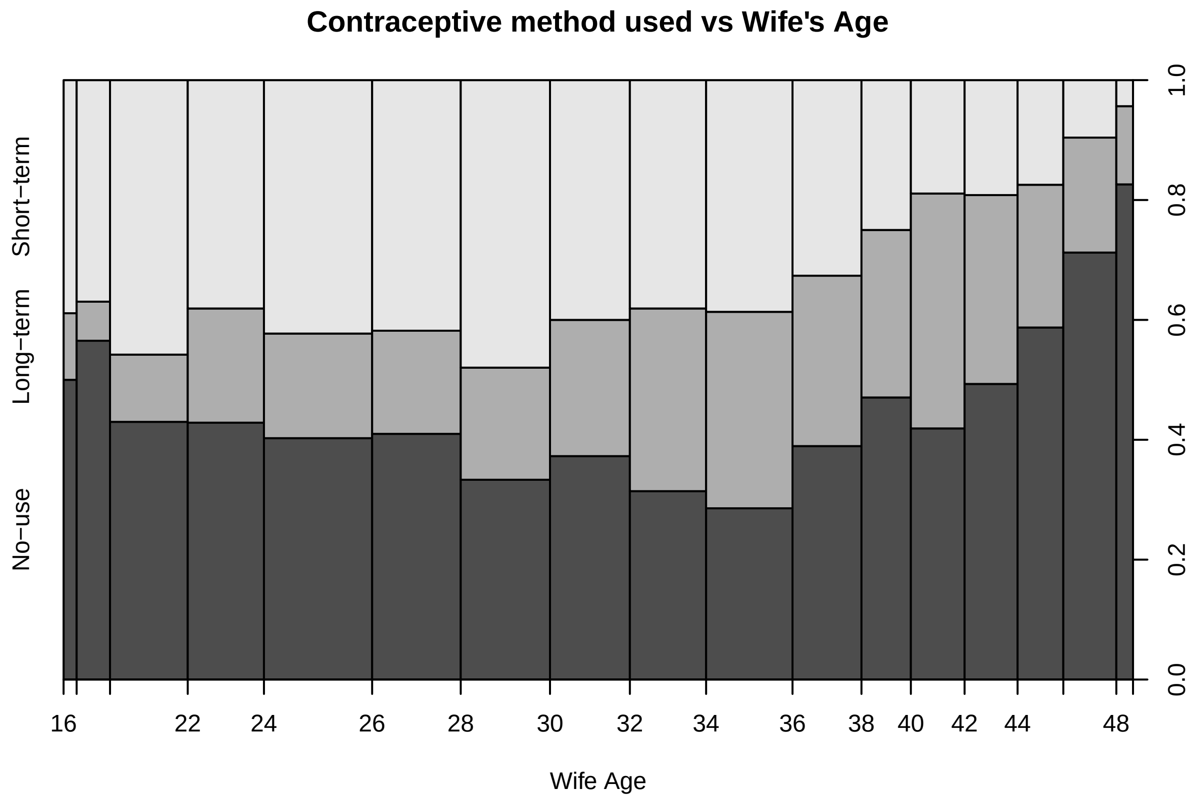


```r
ggplot(data=cmc_data,aes(x=Wife_age,fill=Contraceptive_method)) + geom_density(alpha=0.3)
```



From the above histogram and density plots, we can see that the distributions of wife's age are different among different contraceptive method used groups. For no-use contraceptive method group, the distribution is bimodal with two peaks (one is around middle 20s and the other is around middle 40s); for short-term contraceptive method, the distribution is unimodal and slight right skewed; for long-term contraceptive method, the distribution is multimodal.

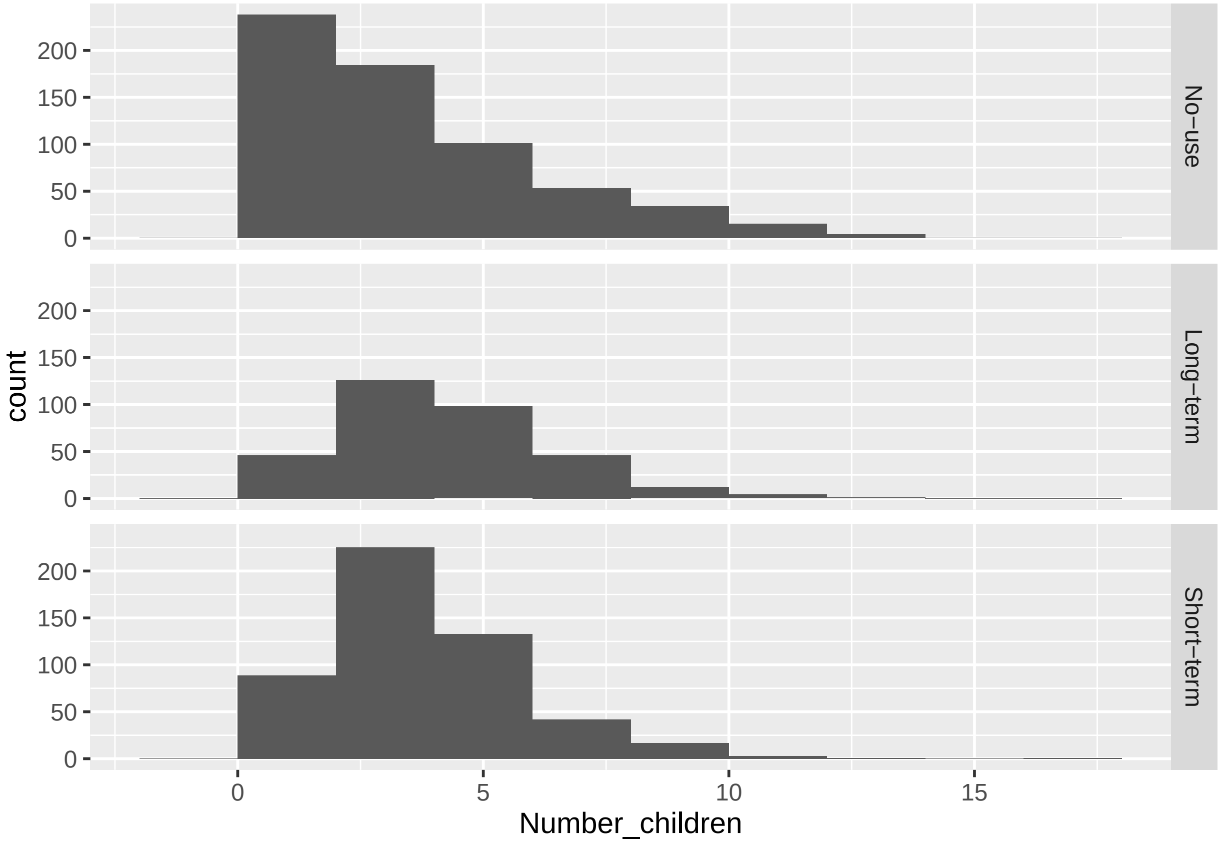
```
par(cex=0.75)
plot(Contraceptive_method ~ Wife_age, data=cmc_data,ylab='',xlab='Wife Age',
     main="Contraceptive method used vs Wife's Age")
```



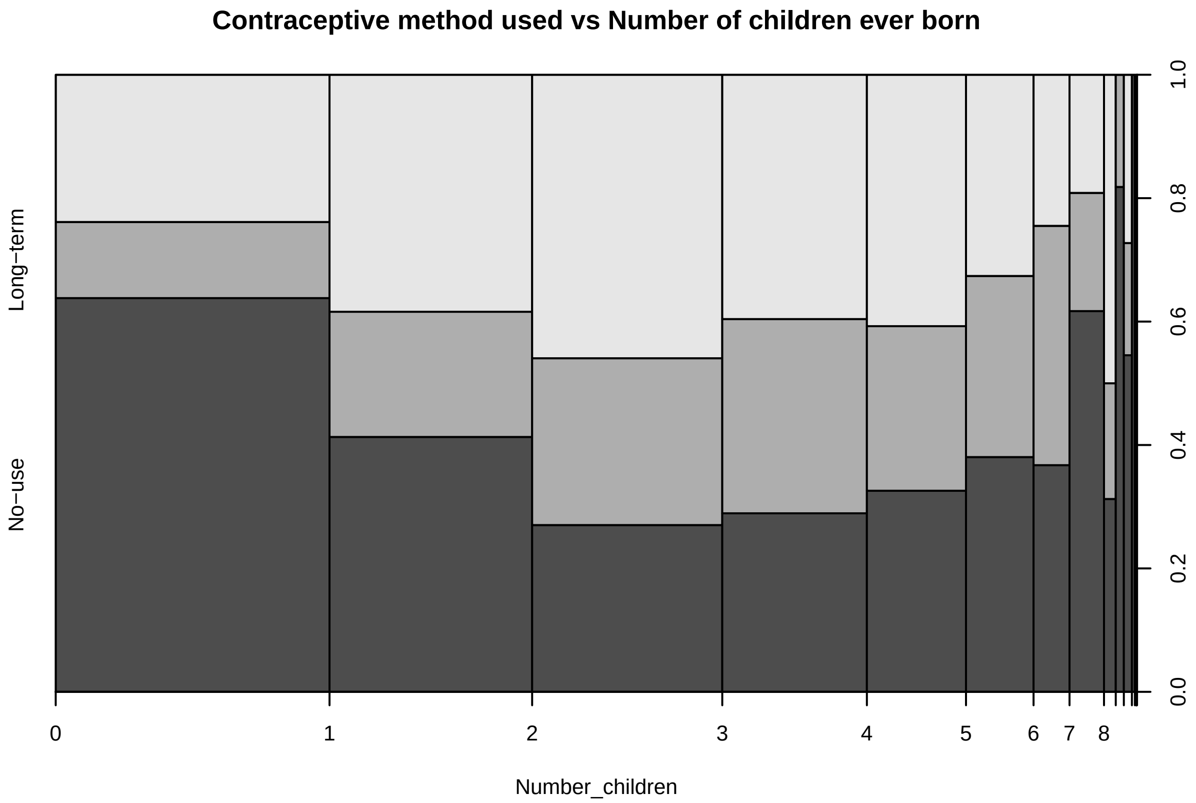
From the above plot, we can see: after a certain age (around age 36) less women report the use of long-term and short-term contraceptive methods, and more women report the no-use contraceptive method.

2. The relationship between number of children ever born and contraceptive method used

```
par(cex=0.70)
ggplot(data=cmc_data,aes(x=Number_children)) + geom_histogram(binwidth=2) + facet_grid(Contraceptive_me
```



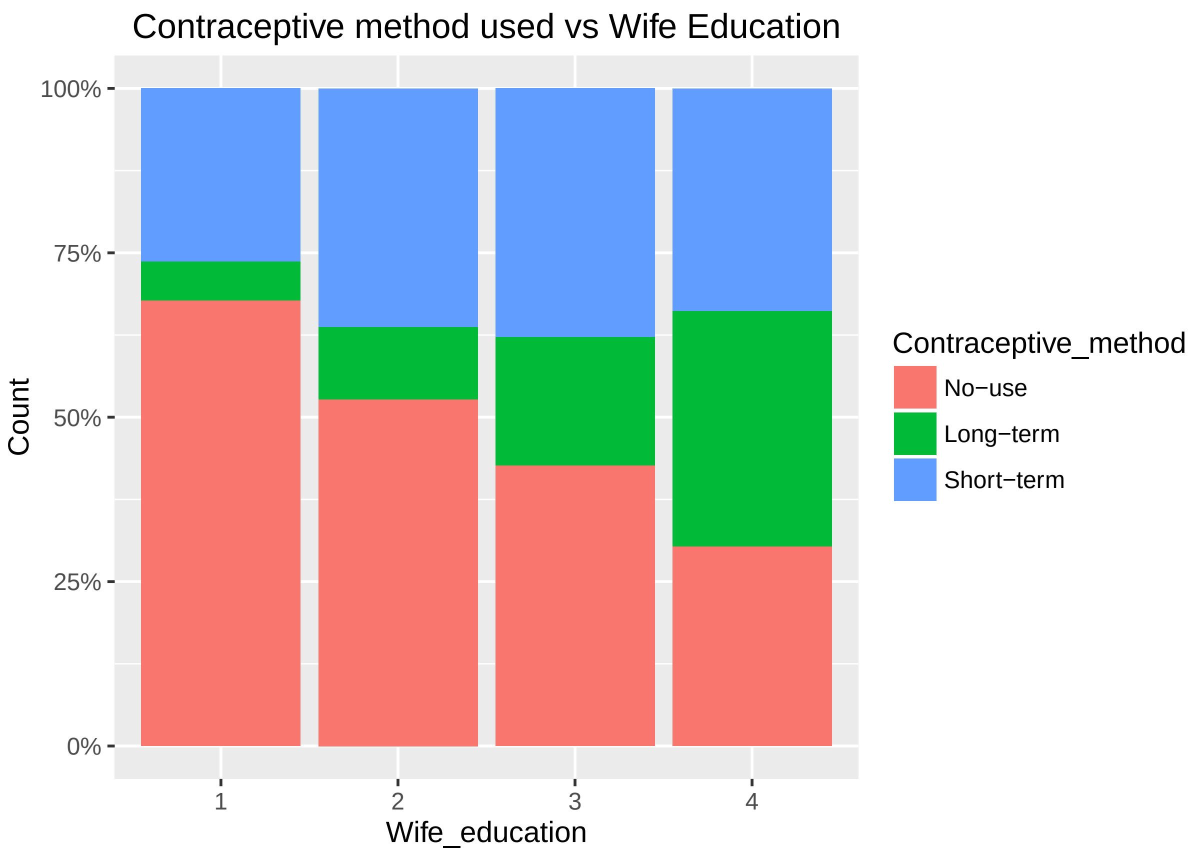
```
plot(Contraceptive_method ~ Number_children, data=cmc_data,ylab='',
     xlab='Number_children',
     main='Contraceptive method used vs Number of children ever born')
```



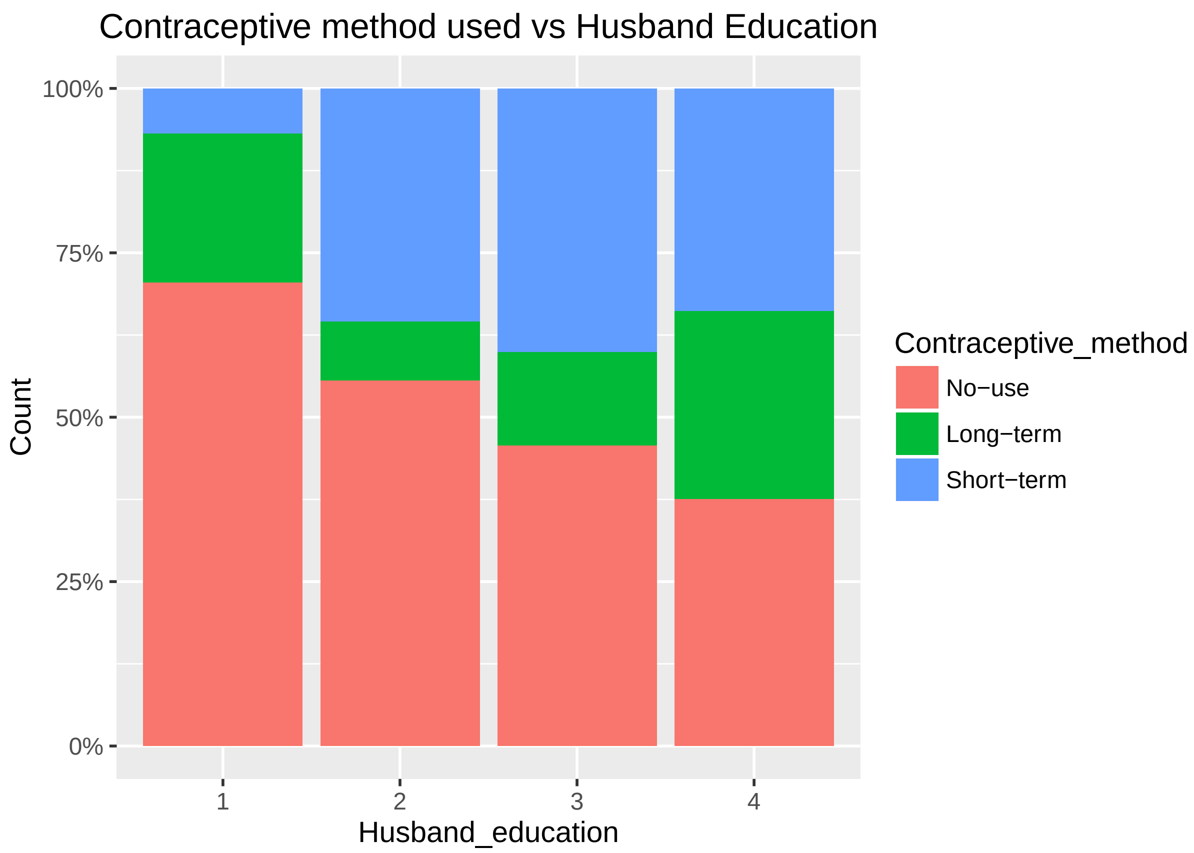
From the above plots we can see that the distribution of number of children ever born among different contraceptive method used groups are very different. For no-use contraceptive method group, the distribution is unimodal but extremely left skewed; for short-term and long-term contraceptive method groups, the distribution is unimodal and very close to normal. And when the number of children ever born is between 0 and 3, with the increasing of the number of children, women report the increasing use of short-term and long-term contraceptive methods and the decreasing no-use contraceptive methods. However, when the number of children ever born exceeds 3, the trend almost goes the oppositve way: women report the decreasing use of short-term and long-term contraceptive methods and the increasing no-use contraceptive methods.

3. The relationship between categorical variables and contraceptive method used

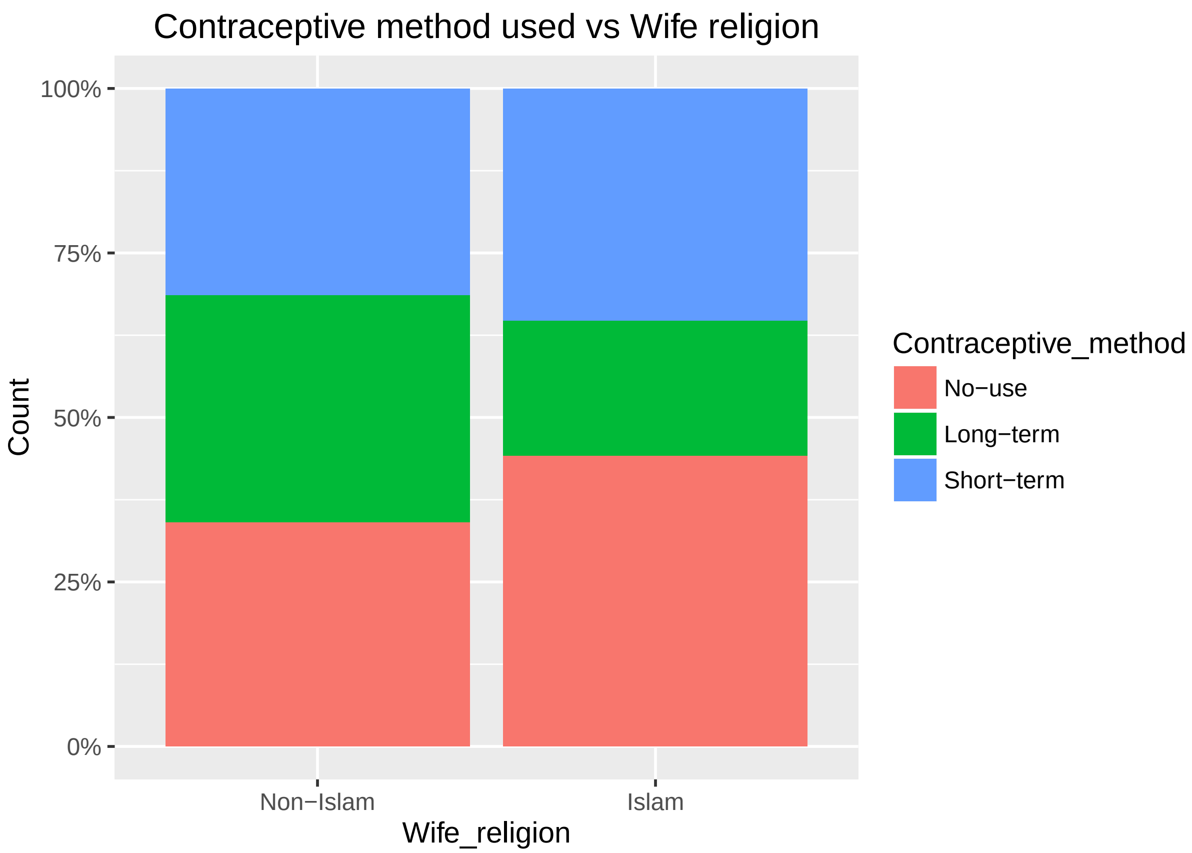
```
library(scales)
counts_1 <- table(cmc_data$Wife_education,cmc_data$Contraceptive_method)
count_1 <- as.data.frame(counts_1)
names(count_1) <- c('Wife_education','Contraceptive_method','Count')
ggplot(data=count_1,aes(y=Count,x=Wife_education,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Wife Education')
```



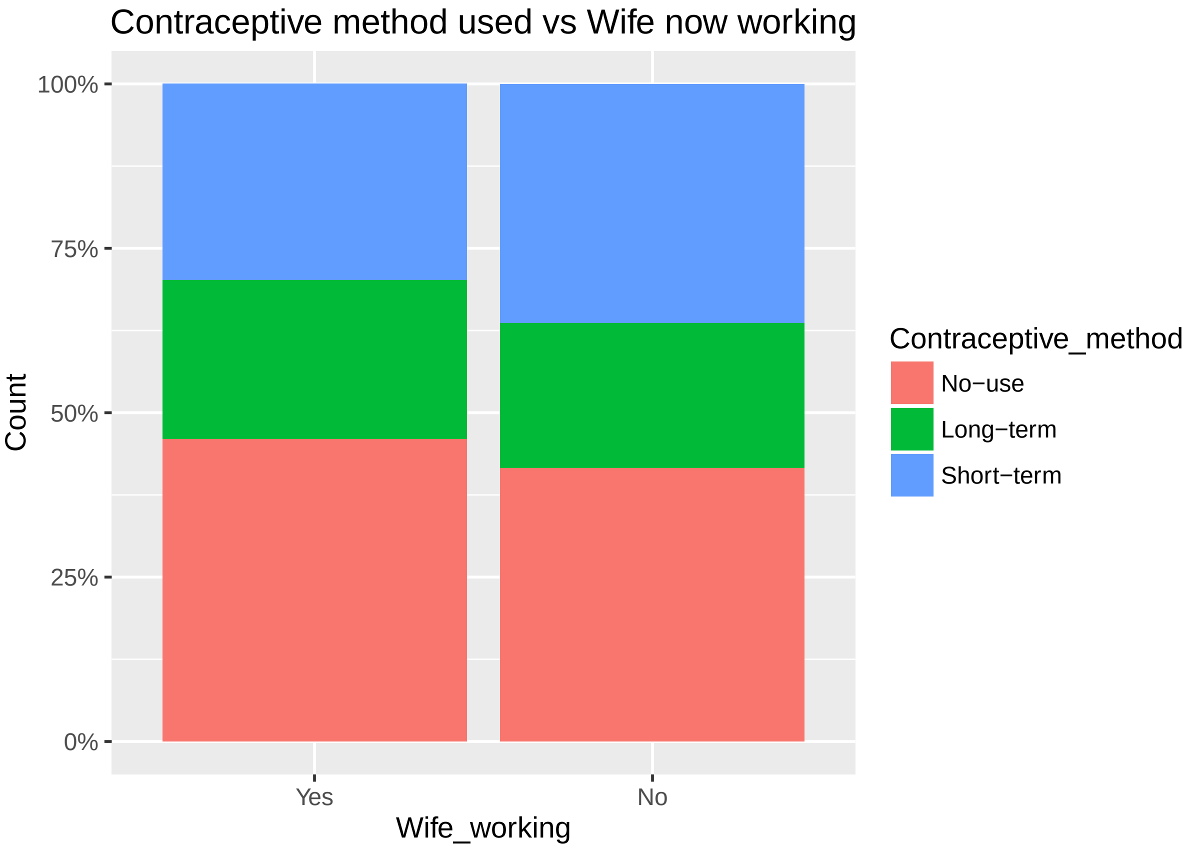
```
counts_2 <- table(cmc_data$Husband_education,cmc_data$Contraceptive_method)
count_2 <- as.data.frame(counts_2)
names(count_2) <- c('Husband_education','Contraceptive_method','Count')
ggplot(data=count_2,aes(y=Count,x=Husband_education,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Husband Education')
```



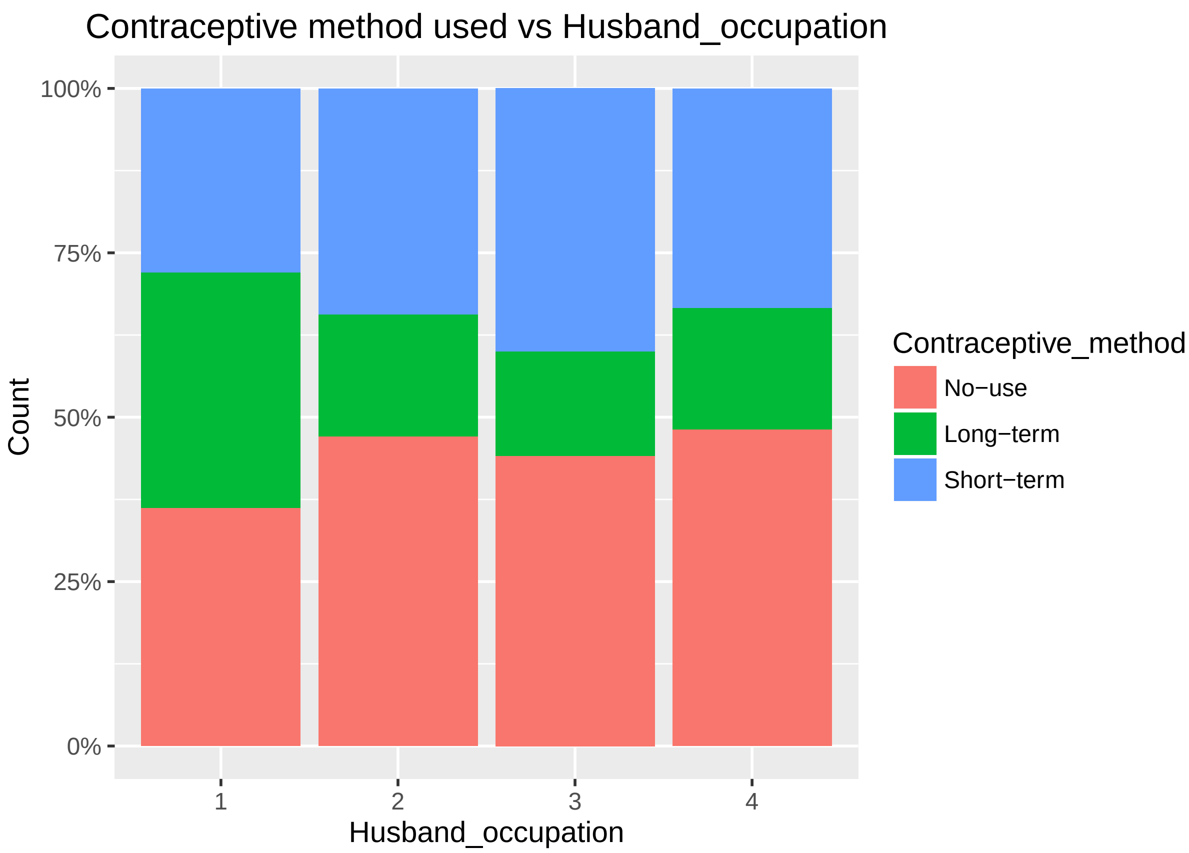
```
counts_3 <- table(cmc_data$Wife_religion,cmc_data$Contraceptive_method)
count_3 <- as.data.frame(counts_3)
names(count_3) <- c('Wife_religion','Contraceptive_method','Count')
ggplot(data=count_3,aes(y=Count,x=Wife_religion,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Wife religion')
```



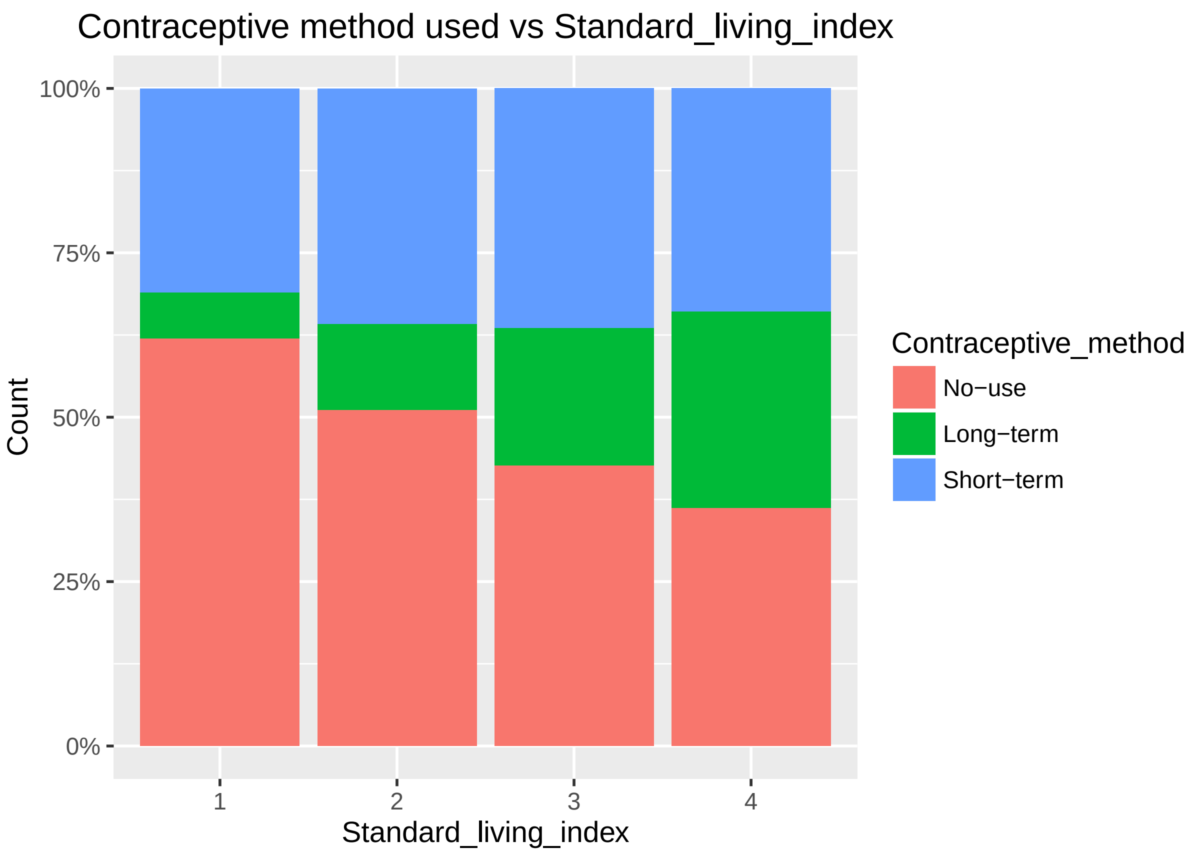
```
counts_4 <- table(cmc_data$Wife_working,cmc_data$Contraceptive_method)
count_4 <- as.data.frame(counts_4)
names(count_4) <- c('Wife_working','Contraceptive_method','Count')
ggplot(data=count_4,aes(y=Count,x=Wife_working,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Wife now working')
```



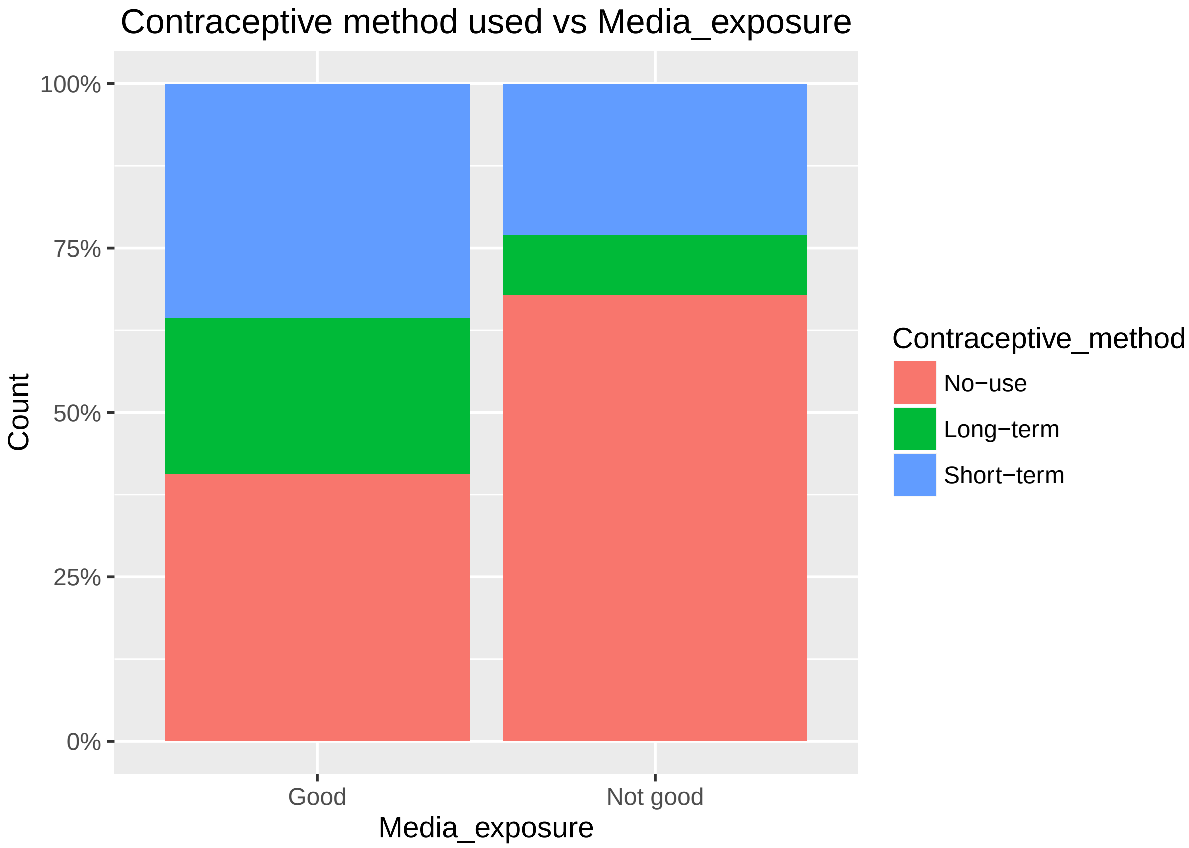
```
counts_5 <- table(cmc_data$Husband_occupation,cmc_data$Contraceptive_method)
count_5 <- as.data.frame(counts_5)
names(count_5) <- c('Husband_occupation','Contraceptive_method','Count')
ggplot(data=count_5,aes(y=Count,x=Husband_occupation,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Husband_occupation')
```



```
counts_6 <- table(cmc_data$Standard_living_index,cmc_data$Contraceptive_method)
count_6 <- as.data.frame(counts_6)
names(count_6) <- c('Standard_living_index','Contraceptive_method','Count')
ggplot(data=count_6,aes(y=Count,x=Standard_living_index,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Standard_living_index')
```



```
counts_7 <- table(cmc_data$Media_exposure,cmc_data$Contraceptive_method)
count_7 <- as.data.frame(counts_7)
names(count_7) <- c('Media_exposure','Contraceptive_method','Count')
ggplot(data=count_7,aes(y=Count,x=Media_exposure,fill=Contraceptive_method)) +
  geom_bar(position='fill',stat='identity') +
  scale_y_continuous(labels=percent) +
  ggtitle('Contraceptive method used vs Media_exposure')
```

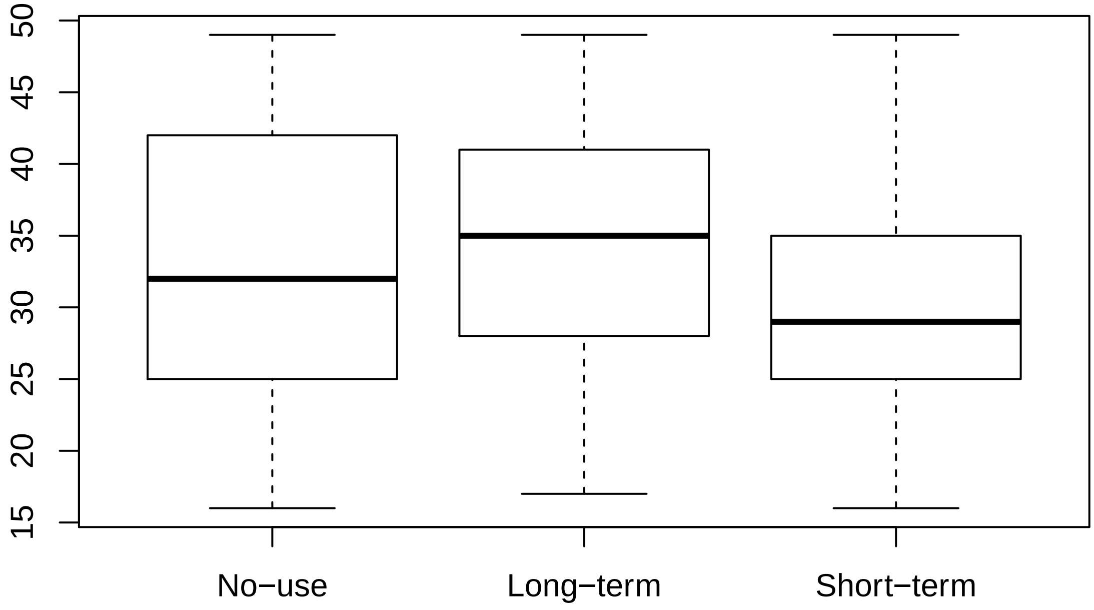


From the above plots, we see there are the associations between contraceptive methods used and some categorical variables, such as wife education, standard living index and median exposure. With the increasing of wife education levels and standard living index, women report the increasing use of short-term and long-term contraceptive methods and the decreasing no-use contraceptive methods; women with good media exposure report more use of short-term and long-term contraceptive methods, less no-use contraceptive method, compared with women with not good media exposure.

Part 4 - Inference

Wife's age and contraceptive method used 1. Study question

```
boxplot(cmc_data$Wife_age ~ cmc_data$Contraceptive_method)
```



```
by(cmc_data$Wife_age, cmc_data$Contraceptive_method, mean)
```

```
## cmc_data$Contraceptive_method: No-use
## [1] 33.42448
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Long-term
## [1] 34.38438
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Short-term
## [1] 30.24462
```

We see the averages of wife age are different among different contraceptive methods used groups. Here I want evaluate whether this kind of difference is statistically significant.

2. Checking conditions

We ususally consider to use ANOVA test to check the means across multiple groups. But before we do that, we need to check the following conditions.

(1) Independence Assumption

- Groups must be independent of each other
- Data within each group must be independent and nearly normal
- Randomization Condition

(2) Equal Variance Assumption

```r
by(cmc_data$Wife_age, cmc_data$Contraceptive_method, var)
```

```
## cmc_data$Contraceptive_method: No-use
## [1] 83.24469
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Long-term
## [1] 55.5747
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Short-term
## [1] 48.21652
```

```r
by(cmc_data$Wife_age, cmc_data$Contraceptive_method, shapiro.test)
```

```
## cmc_data$Contraceptive_method: No-use
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94589, p-value = 2.184e-14
##
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Long-term
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.979, p-value = 8.737e-05
##
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Short-term
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97475, p-value = 1.026e-07
```

It is safe to assume that groups are independ of each other and data within each group are independent and randomization condition meets. However, based on the p values of shairo test, the distributions of wife age within each group are not nearly normal and equal variance assumption within groups does not meet. So, I can not use ANOVA test in this case. Instead I use Kruskal-Wallis test to perform the comparison. Kruskal-Wallis test can be used when we can not run a ANOVA for multiple comparisons because the groups do not follow normal distribution. [2]

3. The hypotheses for test

- H_0: All the means of wife age among the groups are the same.
- H_A: The means of wife age among the groups are not all the same.

4. Kruskal-Wallis test

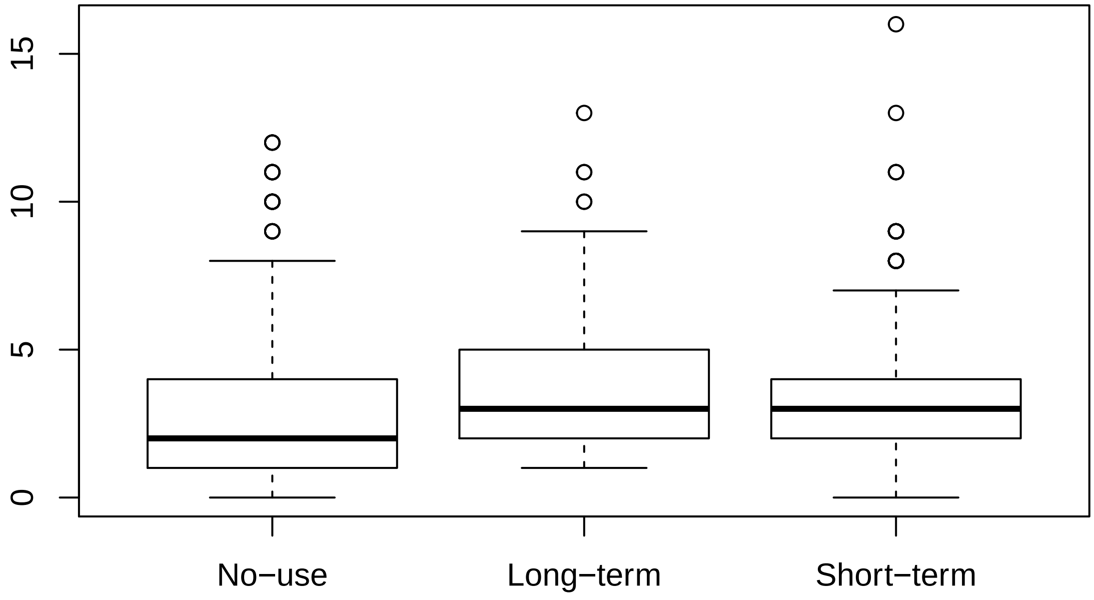
```
kruskal.test(Wife_age ~ Contraceptive_method, data=cmc_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Wife_age by Contraceptive_method
## Kruskal-Wallis chi-squared = 59.864, df = 2, p-value = 1.002e-13
```

Since p-value is less than 0.001, I reject null hypothesis and conclude all the means of wife age among different contraceptive method used groups are not the same.

Number of children ever born and contraceptive method used 1. Study question

```
boxplot(cmc_data$Number_children ~ cmc_data$Contraceptive_method)
```



```
by(cmc_data$Number_childre, cmc_data$Contraceptive_method, mean)
```

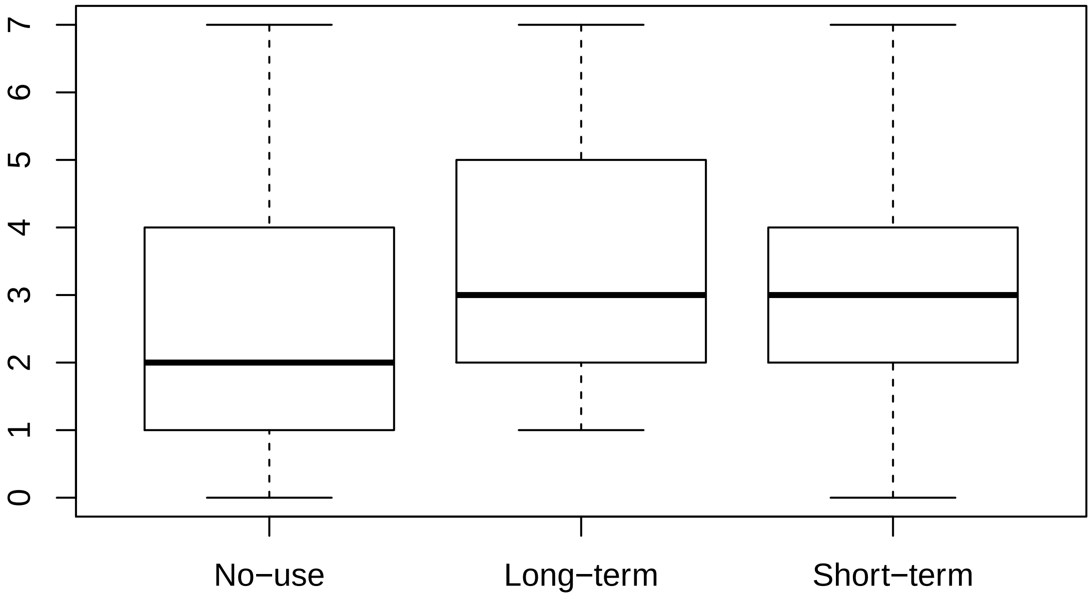
```
## cmc_data$Contraceptive_method: No-use
## [1] 2.934817
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Long-term
## [1] 3.738739
## ------------------------------------------------------------
## cmc_data$Contraceptive_method: Short-term
## [1] 3.35225
```

We see the means of number of children ever born are different among different contraceptive methods used groups. And we want evaluate whether this kind of difference is statistically significant.

2. Outliers

From the above boxplot, I see some outliers. Here I decide to ignore the data points when the number of children is equal or greater than 8.

```
cmc_data_n <- cmc_data[cmc_data$Number_children < 8,]
boxplot(cmc_data_n$Number_children ~ cmc_data_n$Contraceptive_method)
```



```
by(cmc_data_n$Number_childre, cmc_data_n$Contraceptive_method, mean)
```

```
## cmc_data_n$Contraceptive_method: No-use
## [1] 2.369792
## --------------------------------------------------------
## cmc_data_n$Contraceptive_method: Long-term
## [1] 3.452532
## --------------------------------------------------------
## cmc_data_n$Contraceptive_method: Short-term
## [1] 3.0818
```

3. Checking conditions

I will try to use ANOVA test to check the means across multiple groups. But before we do that, we need to check the following conditions.

(1) Independence Assumption

- Groups must be independent of each other
- Data within each group must be independent and nearly normal
- Randomization Condition

(2) Equal Variance Assumption

```
by(cmc_data_n$Number_children, cmc_data_n$Contraceptive_method, var)
```

```
## cmc_data_n$Contraceptive_method: No-use
## [1] 3.73606
## ------------------------------------------------------------
## cmc_data_n$Contraceptive_method: Long-term
## [1] 2.946946
## ------------------------------------------------------------
## cmc_data_n$Contraceptive_method: Short-term
## [1] 2.517885
```

```
by(cmc_data_n$Number_children, cmc_data_n$Contraceptive_method, shapiro.test)
```

```
## cmc_data_n$Contraceptive_method: No-use
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9046, p-value < 2.2e-16
##
## ------------------------------------------------------------
## cmc_data_n$Contraceptive_method: Long-term
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.93296, p-value = 9.724e-11
##
## ------------------------------------------------------------
## cmc_data_n$Contraceptive_method: Short-term
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.92995, p-value = 2.356e-14
```

It is safe to assume that groups are independ of each other and data within each group are independent and randomization condition meets. However, the distribution of number of children ever born within each group are not nearly normal and equal variance assumption within groups do not meet. So, we can not use ANOVA test in this case. Instead I use Kruskal-Wallis test to perform the comparison.

4. The hypotheses for test

- H_0: All the means of number of children ever born among the groups are the same.
- H_A: The means of number of children ever born among the groups are not all the same.

5. Kruskal-Wallis test

```
kruskal.test(Number_children ~ Contraceptive_method, data=cmc_data_n)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Number_children by Contraceptive_method
## Kruskal-Wallis chi-squared = 96.585, df = 2, p-value < 2.2e-16
```

Since p-value is less than 0.001, I reject null hypothesis and conclude that all the means of number of children ever born among different contraceptive method used groups are not the same.

Part 5 - Machine Learning

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other taks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees[3]. Random forests will give estimates of what variances are important in the classification.

```
library(caret)
```

```
## Loading required package: lattice
```

```
inTrain <- createDataPartition(y=cmc_data$Contraceptive_method ,p=0.7,list=FALSE)
training <- cmc_data[inTrain,]
testing <- cmc_data[-inTrain,]
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
model <- randomForest(Contraceptive_method ~ ., data=training)
model
```

```
##
## Call:
##  randomForest(formula = Contraceptive_method ~ ., data = training)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 49.56%
## Confusion matrix:
##            No-use Long-term Short-term class.error
## No-use        263        38        140   0.4036281
## Long-term      59        88         87   0.6239316
## Short-term    122        66        170   0.5251397
```

```
importance(model)
```

```
##                         MeanDecreaseGini
## Wife_age                    156.431247
## Wife_education               50.825124
## Husband_education            35.911501
## Number_children             128.784822
## Wife_religion                17.285052
## Wife_working                 22.926756
## Husband_occupation           45.355519
## Standard_living_index        48.627783
## Media_exposure                8.475997
```

```
predicted <- predict(model,testing)
table(predicted)
```

```
## predicted
##     No-use  Long-term Short-term
##        180         97        163
```

```
confusionMatrix(predicted, testing$Contraceptive_method)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction   No-use Long-term Short-term
##   No-use        117        21         42
##   Long-term      23        43         31
##   Short-term     48        35         80
##
## Overall Statistics
##
##                 Accuracy : 0.5455
##                   95% CI : (0.4976, 0.5927)
##      No Information Rate : 0.4273
##      P-Value [Acc > NIR] : 4.099e-07
##
##                    Kappa : 0.2972
##  Mcnemar's Test P-Value : 0.8653
##
## Statistics by Class:
##
##                      Class: No-use Class: Long-term Class: Short-term
## Sensitivity                 0.6223          0.43434            0.5229
## Specificity                 0.7500          0.84164            0.7108
## Pos Pred Value              0.6500          0.44330            0.4908
## Neg Pred Value              0.7269          0.83673            0.7365
## Prevalence                  0.4273          0.22500            0.3477
## Detection Rate              0.2659          0.09773            0.1818
## Detection Prevalence        0.4091          0.22045            0.3705
## Balanced Accuracy           0.6862          0.63799            0.6168
```

From the model, we can see that the top four important variables are wife's age, number of children ever born, wife's education and standard-of-living index. Overall, I got a model with around 55% Accuracy.

Part 6 - Conclusion

- From the above analysis, I conclude that there are strong associations between contraceptive methods used and the variables, such as wife's age and number of children ever born, which are the most important two. I even built a relatively accurate model to predict the contraceptive methods used based on the explanatory variables provided.
- In this data, the distributions of wife age and number of children ever born among different contraceptive methods used groups are not nearly normal. Although I was able to bypass this issue by conducting Kruskal-Wallis test, it might be better to sample data within all women with or without getting married.
- This study provides us interesting information about the associations between contraceptive methods used by married woman in Indonesia and her socio-economic characteristics in 1987. We could conduct the similar survey to find out what kind of factors are associated with the choice of contraceptive methods among the women in US now.

References

- [1] OpenIntro Statistics (Third Edition) Page 19
- [2] http://www.r-bloggers.com/kruskal-wallis-one-way-analysis-of-variance/
- [3] https://en.wikipedia.org/wiki/Random_forest