Discrete variable: Mean, Variance, Probability distribution, Joint and Conditional probabilities

Mandatory Exercise no. 1

Information on data

The data file is shared on moodle called data.rds. This data is based on statistics for a state in the US where all the incidents of **police stop and search** are reported during Jan 2010 - Mar 2019. The purpose of the project is to investigate whether there is statistical evidence of police stopping people on the basis of their ethnicity. You can now start with the exercise.

Importing data

You are provided with two variables from the data set (age and race). First, download the data.rds file and save it on your hard drive. You can then run the following command to import your file (you need to provide your file path where the data file is saved):

```
rm(list=ls(all=T))
# library(tidyverse)
# library(lubridate)
options(digits=2)
data <- readRDS("~/Dropbox/Teaching/Statistics/mandatory exercises/data.rds")
```

If you look at the data, it consists of age up to 99. I would like you to create a subset of the sample where we can filter the data and only consider age from 18 - 45 years. The following command in R filters the data and creates a subsample of our choice.

```
newdata <- subset(data, age >= 18 & age<=45)
```

Section A (mandatory):

- *Q1: Create a cross table in R (also known as contingency table). Preferably keep age variable in rows and ethnicity (race) in columns. Below is just a sample:*

Table 1: Cross table example

	X (ethnicity)							
	Asian	*Black*		*Hispanic*	*White*		*Other*	**Py**
Y (age)								
18	*	*	*	*	*	*	*	*
19	*	*	*	*	*	*	*	*
20	*	*	*	*	*	*	*	*
-	*	*	*	*	*	*	*	*
-	*	*	*	*	*	*	*	*
45	*	*	*	*	*	*	*	*
Px	*	*	*	*	*	*	*	**1**

Below, I define my variables and create a crosstable in R

```
age <- newdata$age
race <- newdata$race
tab <- xtabs( ~ age + race)
```

```
tab
```

```
##     race
## age  asian/pacific islander black hispanic white other unknown
##   18                    733 22414     4445 24062   259       0
##   19                    939 33250     5363 31754   297       0
##   20                    978 39806     5609 36779   279       0
##   21                   1193 43061     5520 42251   309       0
##   22                   1209 44787     5559 48121   310       0
##   23                   1446 45637     5919 55754   360       0
##   24                   1422 43751     5879 58313   366       0
##   25                   1402 43167     5972 59424   397       0
##   26                   1354 40208     5739 58087   353       0
##   27                   1401 37261     5839 56486   381       0
##   28                   1278 35722     5663 54451   372       0
##   29                   1229 32084     5190 48704   334       0
##   30                   1278 33752     5893 50107   375       0
##   31                   1138 30102     5380 44412   294       0
##   32                   1161 32164     5763 46322   327       0
##   33                   1067 28944     5295 41668   319       0
##   34                   1100 30003     5657 42081   300       0
##   35                   1085 28354     5790 40537   344       0
##   36                    941 25093     4841 35252   274       0
##   37                    940 23636     4633 33495   273       0
##   38                    956 23156     4390 32585   267       0
##   39                    955 20834     3860 29636   221       0
##   40                   1014 23358     4153 33385   281       0
##   41                    909 20487     3562 29894   238       0
##   42                    940 20944     3503 31551   221       0
##   43                    949 21295     3563 31953   227       0
##   44                    807 19263     2997 29446   185       0
##   45                    902 21956     3182 33998   226       0
```

Here, you can see there is an irrelevant categoy unknown which has zero entries, I can get rid of this in R by typing:

```
table <- tab[,-6] # It removes the 6th column (which is unknown and has zero entries)
```

- *Q2: Convert your table to discrete probability distribution (prop.table function)*

Now I can convert my table to discrete probability in R:

```
options(scipen=999)
x <- prop.table(table)
x
```

```
##     race
## age  asian/pacific islander    black hispanic    white    other
##   18               0.000333 0.010173 0.002017 0.010921 0.000118
##   19               0.000426 0.015091 0.002434 0.014412 0.000135
##   20               0.000444 0.018067 0.002546 0.016693 0.000127
##   21               0.000541 0.019544 0.002505 0.019176 0.000140
##   22               0.000549 0.020328 0.002523 0.021841 0.000141
##   23               0.000656 0.020713 0.002686 0.025305 0.000163
##   24               0.000645 0.019857 0.002668 0.026467 0.000166
##   25               0.000636 0.019592 0.002711 0.026971 0.000180
##   26               0.000615 0.018249 0.002605 0.026364 0.000160
##   27               0.000636 0.016912 0.002650 0.025637 0.000173
##   28               0.000580 0.016213 0.002570 0.024714 0.000169
##   29               0.000558 0.014562 0.002356 0.022105 0.000152
```

```
##   30               0.000580 0.015319 0.002675 0.022742 0.000170
##   31               0.000517 0.013662 0.002442 0.020157 0.000133
##   32               0.000527 0.014598 0.002616 0.021024 0.000148
##   33               0.000484 0.013137 0.002403 0.018912 0.000145
##   34               0.000499 0.013617 0.002568 0.019099 0.000136
##   35               0.000492 0.012869 0.002628 0.018399 0.000156
##   36               0.000427 0.011389 0.002197 0.016000 0.000124
##   37               0.000427 0.010728 0.002103 0.015202 0.000124
##   38               0.000434 0.010510 0.001992 0.014789 0.000121
##   39               0.000433 0.009456 0.001752 0.013451 0.000100
##   40               0.000460 0.010602 0.001885 0.015152 0.000128
##   41               0.000413 0.009298 0.001617 0.013568 0.000108
##   42               0.000427 0.009506 0.001590 0.014320 0.000100
##   43               0.000431 0.009665 0.001617 0.014503 0.000103
##   44               0.000366 0.008743 0.001360 0.013365 0.000084
##   45               0.000409 0.009965 0.001444 0.015431 0.000103
```

- *Q3: Calculate marginal probability of age (Py) and race (Px).*

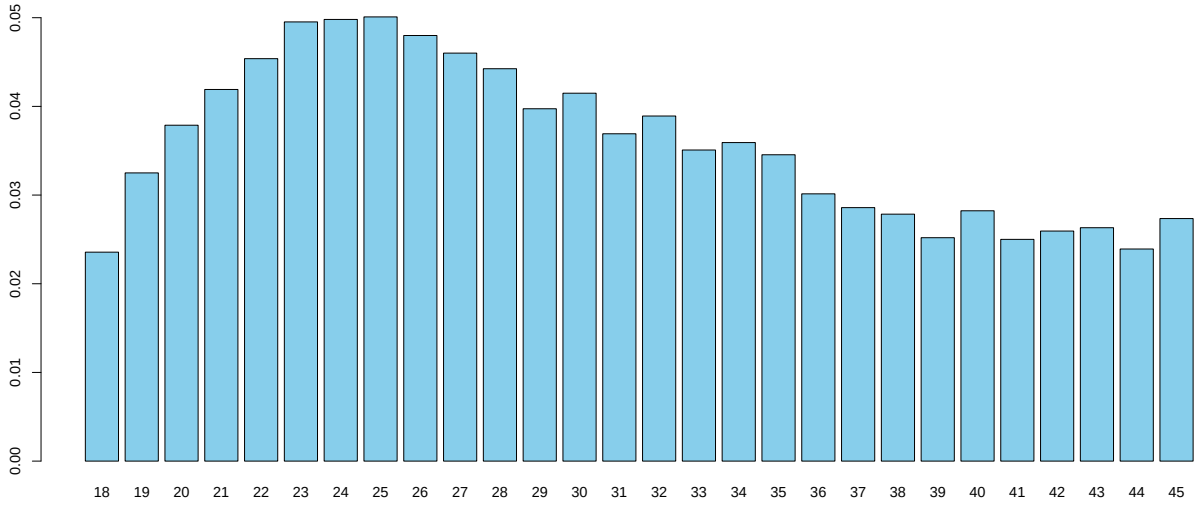
```
tab <- addmargins(x)
tab
```

```
##      race
## age   asian/pacific islander    black hispanic    white    other      Sum
##   18                0.000333 0.010173 0.002017 0.010921 0.000118 0.023562
##   19                0.000426 0.015091 0.002434 0.014412 0.000135 0.032498
##   20                0.000444 0.018067 0.002546 0.016693 0.000127 0.037876
##   21                0.000541 0.019544 0.002505 0.019176 0.000140 0.041908
##   22                0.000549 0.020328 0.002523 0.021841 0.000141 0.045381
##   23                0.000656 0.020713 0.002686 0.025305 0.000163 0.049525
##   24                0.000645 0.019857 0.002668 0.026467 0.000166 0.049804
##   25                0.000636 0.019592 0.002711 0.026971 0.000180 0.050090
##   26                0.000615 0.018249 0.002605 0.026364 0.000160 0.047993
##   27                0.000636 0.016912 0.002650 0.025637 0.000173 0.046008
##   28                0.000580 0.016213 0.002570 0.024714 0.000169 0.044246
##   29                0.000558 0.014562 0.002356 0.022105 0.000152 0.039732
##   30                0.000580 0.015319 0.002675 0.022742 0.000170 0.041486
##   31                0.000517 0.013662 0.002442 0.020157 0.000133 0.036911
##   32                0.000527 0.014598 0.002616 0.021024 0.000148 0.038914
##   33                0.000484 0.013137 0.002403 0.018912 0.000145 0.035081
##   34                0.000499 0.013617 0.002568 0.019099 0.000136 0.035920
##   35                0.000492 0.012869 0.002628 0.018399 0.000156 0.034544
##   36                0.000427 0.011389 0.002197 0.016000 0.000124 0.030137
##   37                0.000427 0.010728 0.002103 0.015202 0.000124 0.028583
##   38                0.000434 0.010510 0.001992 0.014789 0.000121 0.027847
##   39                0.000433 0.009456 0.001752 0.013451 0.000100 0.025193
##   40                0.000460 0.010602 0.001885 0.015152 0.000128 0.028227
##   41                0.000413 0.009298 0.001617 0.013568 0.000108 0.025004
##   42                0.000427 0.009506 0.001590 0.014320 0.000100 0.025943
##   43                0.000431 0.009665 0.001617 0.014503 0.000103 0.026319
##   44                0.000366 0.008743 0.001360 0.013365 0.000084 0.023918
##   45                0.000409 0.009965 0.001444 0.015431 0.000103 0.027352
##   Sum               0.013946 0.392366 0.063160 0.526720 0.003808 1.000000
```

- *Q4: Plot the probability distribution of age (i.e., plot Py) - make sure your x-axis shows age from 1-45 and y-axis shows the probabilities. Which age group has the highest probabily of being stopped and searched by police officers.*

For pltoting the probability distribution of age, I need the marginal probability of age (Py). I can extract my Py from the table in R below (note that I do not need the last value in the column because it is the sum of all marginal probabilities):

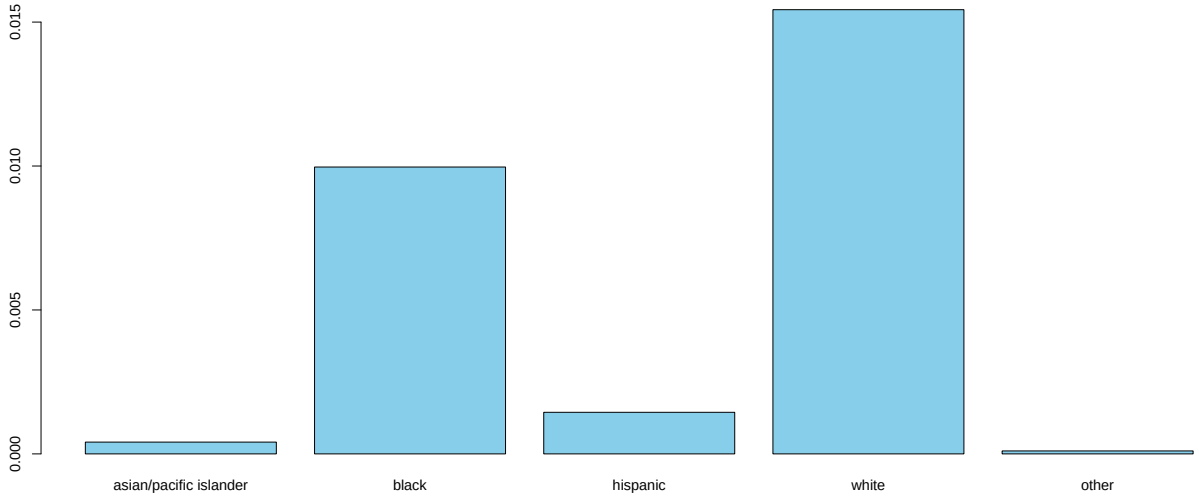
```
py= tab[1:28,6]
barplot(py, col="skyblue")
```



- *Q5: Plot the probability dist. of race (i.e, plot Px) - make sure your x-axis shows ethnic groups and y-axis shows the probabilities.*

For pltoting the probability distribution of race/ethnic group, I need the marginal probability of race (Px). I can extract my Px from the table in R below

```
px= tab[28,1:5]
barplot(px, col="skyblue")
```



- *Q6: Calculate the expected value of age, E[Y]*

To calculate the expected value of age, I use the formula $E[Y] = \sum Y(Py)$. I already have computed Py (marginal probability of age), I just need the vector for age which I can define as follows:

```
Y=c(18:45) # age is from 18-45
mean_age <- sum(Y*py)
mean_age
```

```
## [1] 30
```

- *Q7: Calculate the variance of age, Var[Y]*

To calculate the variance I can use the formula $Var[Y] = \sum (Y - \mu_y)^2(Py)$:

```
var_age <- sum((Y - mean_age)^2*(py))
var_age
```

```
## [1] 57
```

- *Q8: Calculate the conditional probability of age (give that the ethnic group is black) P(Y|black).*

So, we need to calculate the conditional probability of the whole vector (age) given that the ethnicity is black. We can compute all conditional probabilities in R by typing:

```
cp <- prop.table(x, margin=2)
cp
```

```
##     race
## age  asian/pacific islander black hispanic white other
##   18                  0.024 0.026    0.032 0.021 0.031
##   19                  0.031 0.038    0.039 0.027 0.035
##   20                  0.032 0.046    0.040 0.032 0.033
##   21                  0.039 0.050    0.040 0.036 0.037
##   22                  0.039 0.052    0.040 0.041 0.037
##   23                  0.047 0.053    0.043 0.048 0.043
##   24                  0.046 0.051    0.042 0.050 0.044
##   25                  0.046 0.050    0.043 0.051 0.047
##   26                  0.044 0.047    0.041 0.050 0.042
##   27                  0.046 0.043    0.042 0.049 0.045
##   28                  0.042 0.041    0.041 0.047 0.044
##   29                  0.040 0.037    0.037 0.042 0.040
##   30                  0.042 0.039    0.042 0.043 0.045
##   31                  0.037 0.035    0.039 0.038 0.035
##   32                  0.038 0.037    0.041 0.040 0.039
##   33                  0.035 0.033    0.038 0.036 0.038
##   34                  0.036 0.035    0.041 0.036 0.036
##   35                  0.035 0.033    0.042 0.035 0.041
##   36                  0.031 0.029    0.035 0.030 0.033
##   37                  0.031 0.027    0.033 0.029 0.033
##   38                  0.031 0.027    0.032 0.028 0.032
##   39                  0.031 0.024    0.028 0.026 0.026
##   40                  0.033 0.027    0.030 0.029 0.033
##   41                  0.030 0.024    0.026 0.026 0.028
##   42                  0.031 0.024    0.025 0.027 0.026
##   43                  0.031 0.025    0.026 0.028 0.027
##   44                  0.026 0.022    0.022 0.025 0.022
##   45                  0.029 0.025    0.023 0.029 0.027
```

In the above table, P(Y|black) is the second column which is our conditional probability of age given that the ethnic group is black. we can extract this column by writing:

```
cp_black <- cp[,2]
```

- *Q9: Calculate the conditional mean of age (given that the ethnic group is black), E(Y|black)*

To calculate the conditional mean (given that the ethnic group is black), we simply need to use the formula $E[Y|black] = \sum Y * P(Y|black)$

```
c_mean <- sum(Y*cp_black)
c_mean
```

```
## [1] 30
```

- *Q10: Calculate the conditional variance of age (given that the ethnic group is black), Var(Y|black)*

To calculate the conditional mean (given that the ethnic group is black), we simply need to use the formula $Var[Y|black] = \sum(Y - \mu_{Y|black})^2 P(Y|black)$

```
c_var <- sum((Y - c_mean)^2*(cp_black))
c_var
```

```
## [1] 58
```

Section B (optional):

- *A: Assume, the police stops a person who is of 20 years of age, what is the probability that the person is black?*

In this case, the condition is that the person is of 20 years of age. We can obtain all conditional probabilities in R by assuming that the rows are given:

```
cp1 <- prop.table(x, margin=1)
cp1
```

```
##     race
## age  asian/pacific islander  black hispanic  white  other
##   18                 0.0141 0.4318   0.0856 0.4635 0.0050
##   19                 0.0131 0.4644   0.0749 0.4435 0.0041
##   20                 0.0117 0.4770   0.0672 0.4407 0.0033
##   21                 0.0129 0.4664   0.0598 0.4576 0.0033
##   22                 0.0121 0.4479   0.0556 0.4813 0.0031
##   23                 0.0133 0.4182   0.0542 0.5110 0.0033
##   24                 0.0130 0.3987   0.0536 0.5314 0.0033
##   25                 0.0127 0.3911   0.0541 0.5384 0.0036
##   26                 0.0128 0.3802   0.0543 0.5493 0.0033
##   27                 0.0138 0.3676   0.0576 0.5572 0.0038
##   28                 0.0131 0.3664   0.0581 0.5586 0.0038
##   29                 0.0140 0.3665   0.0593 0.5564 0.0038
##   30                 0.0140 0.3693   0.0645 0.5482 0.0041
##   31                 0.0140 0.3701   0.0662 0.5461 0.0036
##   32                 0.0135 0.3751   0.0672 0.5403 0.0038
##   33                 0.0138 0.3745   0.0685 0.5391 0.0041
##   34                 0.0139 0.3791   0.0715 0.5317 0.0038
##   35                 0.0143 0.3725   0.0761 0.5326 0.0045
##   36                 0.0142 0.3779   0.0729 0.5309 0.0041
##   37                 0.0149 0.3753   0.0736 0.5319 0.0043
##   38                 0.0156 0.3774   0.0716 0.5311 0.0044
##   39                 0.0172 0.3753   0.0695 0.5339 0.0040
##   40                 0.0163 0.3756   0.0668 0.5368 0.0045
##   41                 0.0165 0.3719   0.0647 0.5426 0.0043
##   42                 0.0164 0.3664   0.0613 0.5520 0.0039
##   43                 0.0164 0.3672   0.0614 0.5510 0.0039
##   44                 0.0153 0.3655   0.0569 0.5588 0.0035
##   45                 0.0150 0.3643   0.0528 0.5642 0.0038
```

I can clearly see the answer is [0.477]

- *B: Assume, the police stops a person who is 19 years of age, what is the probability that the person is white?*

I can see the answer is [0.4435]

Section C (mandatory):

- Derive $Var(a + bX + cY)$ where a,b, and c are constants whereas X and Y are variables
- Derive $E(a + bX + cY)$
- Show that the $Cov(X, X)$ can be written as the $Var(X)$
- Show that $Cov(X, Y) = E((X - \mu_X)(Y - \mu_X))$ is also equal to $E(XY) - \mu_X\mu_Y$

Afsnit A (mandatory):

- Q1: Opstil en krydstabel i R (også kendt som contingency table). Afbild "age" på x-aksen og race på y-aksen. Nedenfor er et eksempel:

Table 2: Cross table example

	X (ethnicity)							
	Asian	*Black*		*Hispanic*	*White*		*Other*	**Py**
Y (age)								
18	*	*	*	*	*	*	*	*****
19	*	*	*	*	*	*	*	*****
20	*	*	*	*	*	*	*	*****
-	*	*	*	*	*	*	*	*****
-	*	*	*	*	*	*	*	*****
45	*****	*****	*****	*****	*****	*****	*****	*****
Px	*	*	*	*	*	*	*	**1**

- Q2: Konverter din krydstabel til diskret sandsynlighedsfordeling (prop.table funktion i R)
- Q3: Beregn den marginale sandsynlighed for age (Py) og race (Px).
- Q4: Plot sandsynlighedsfordelingen for alder (dvs. plot Py) - din x-akse skal vise alder fra 1-45 og y-aksen skal vise sandsynlighederne. Hvilken aldersgruppe har størst sandsynlighed for at blive stoppet og undersøgt af politiet?
- Q5: Plot sandsynlighedsfordelingen for race (dvs. plot Py) - din x-akse skal vise etniske grupper og y-aksen skal vise sandsynlighederne.
- Q6: Beregn den forventede værdi af alder E(Y)
- Q7: Beregn variansen af alder Var(Y)
- Q8: Beregn den betingede sandsynlighed for alder (givet betingelsen, at den etniske gruppe er black), P(Y|black).
- Q9: Beregn det betingede gennemsnit af alder (givet betingelsen, at den etniske gruppe er black), E(Y|black)
- Q10: Beregn den betingede varians af alder (givet betingelsen, at den etniske gruppe er black), Var(Y|black)

Afsnit B (optional):

- A: Antag, at politiet stopper en person på 20 år. Hvad er sandsynligheden for, at personens hudfarve er sort?
- B: Antag, at politiet stopper en person på 19 år. Hvad er sandsynligheden for, at personens hudfarve er hvid?

Section C (mandatory):

- Udled $Var(a + bX + cY)$ hvor a,b, and c er konstanter, mens X og Y er variable
- Udled $E(a + bX + cY)$

- Vis at $Cov(X, X)$ kan skrives som $Var(X)$
- Vis at $Cov(X, Y) = E((X - \mu_X)(Y - \mu_X))$ er også lig med $E(XY) - \mu_X \mu_Y$