# 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):

```r
rm(list=ls(all=T))
# library(tidyverse)
# library(lubridate)
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.

```r
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** |

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

- Q3: Calculate marginal probability of age (Py) and race (Px).
- 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.
- 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.
- Q6: Calculate the expected value of age, E[Y]
- Q7: Calculate the variance of age, Var[Y]
- Q8: Calculate the conditional probability of age (give that the ethnic group is black) P(Y|black).
- Q9: Calculate the conditional mean of age (given that the ethnic group is black), E(Y|black)
- Q10: Calculate the conditional variance of age (given that the ethnic group is black), Var(Y|black)

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

## 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 $E(XY) - \mu_X \mu_Y$