Проект то Статистика и Емпирични Методи

# На Виктор Христов, №62151

1. **Цел**

Да се анализират данните за военните сили на всяка държава през 2017 година ([хипервръзка към данни](https://www.kaggle.com/blitzr/gfp2017)) . Данните са разделени на много категории, което позволява държавите да се сравняват по различни критерии като например земни сили, въздушни сили, морски сили, външни дългове и т.н.

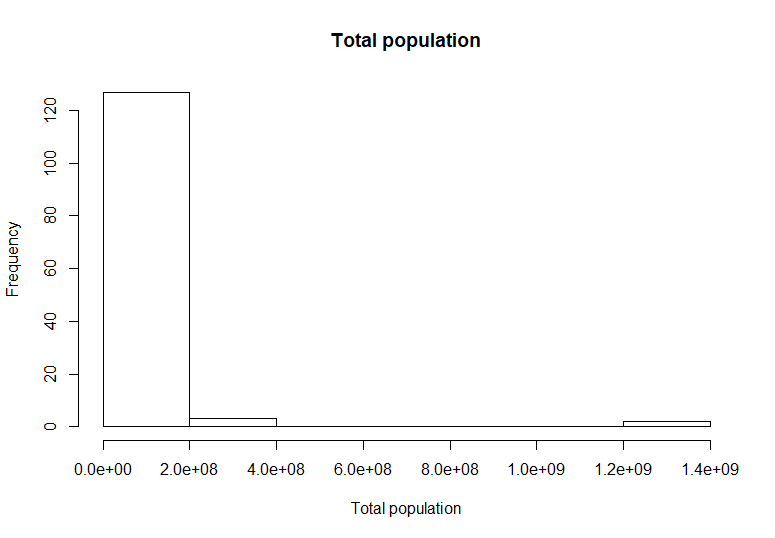
С цел да се избегне затрупване на читателя с информация, в рамките на този проект ще се разглеждат и анализират единствено данните за армиите на всяка държава спрямо броя хора. Това ще включва държавите, общото население на съответната държава , пълния състав на армията, както и активния участък от тази армия.

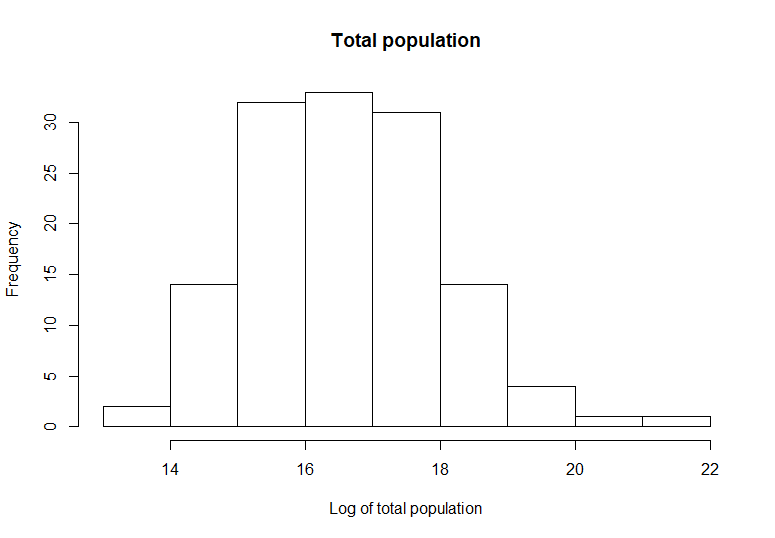
1. **Променливи**
2. Държава (Country)

Вектор от категорийни стойности, който описва всяка от 133те държави, включени в тази извадка от данни. Стойностите са уникални и няма как да се сравнят помежду си.

1. Общо население (Total Population)

Вектор от числени стойности, който описва населението на всяка държава, включена в тази извадка от данни.

1. Хистограма

Поради големият брой държави с население под 200 милиона души, заедно с огромното население на Китай и Индия, хистограмата получава следния логаритмичен вид. За да нормализира визуализацията на този вектор, ще приложим команда log на неговите стойности и ще визуализираме резултата. В резултат се получава следната хистограма:

Данните вече са нормално разпределени. По същият начин трябва да се коригират и останалите числени променливи, защото техните хистограми имат същия логаритмичен вид.

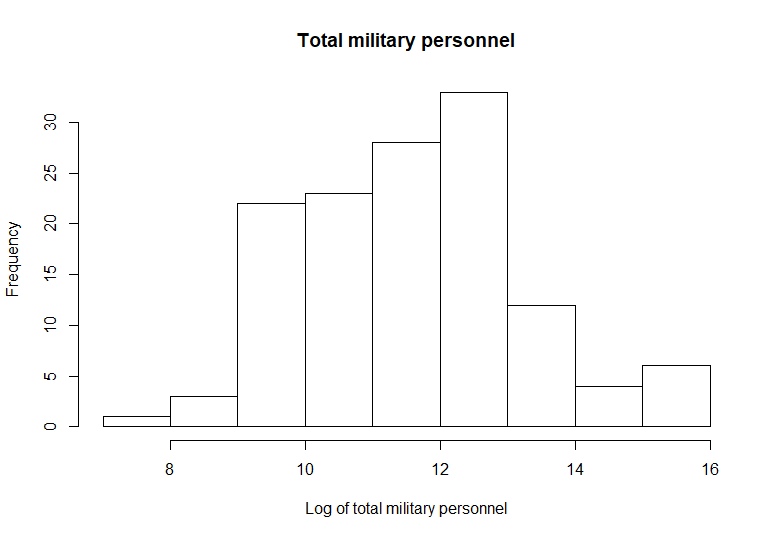
В нормален вид:

1. Локация = 15,510,711;
2. Очакване = 53,600,869;
3. Дисперсия = 165,351,348

В коригиран вид:

1. Локация = 16.55704;
2. Очакване = 16.60369;
3. Дисперсия = 1.40884
4. Обща човешка военна сила (Total military personnel)

Вектор от числени стойности, който описва общата човешка военна сила на всяка държава, включена в тази извадка от данни.

1. Хистограма

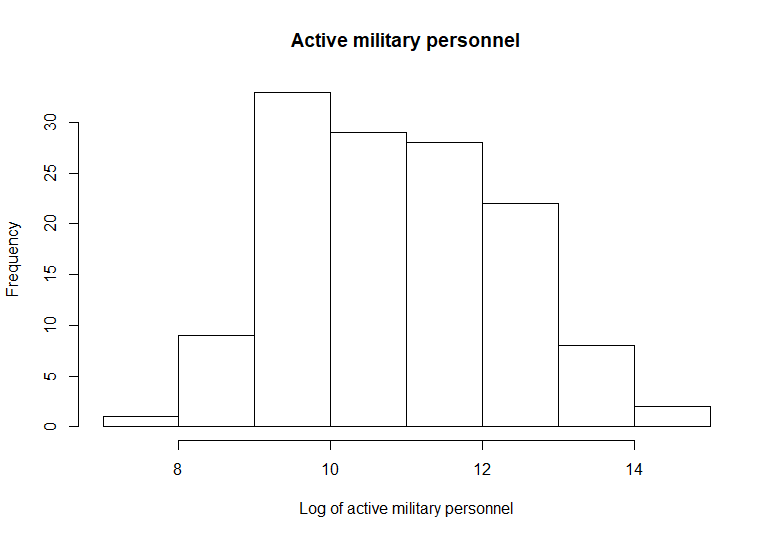
В нормален вид:

1. Локация = 111,250;
2. Очакване = 449,237;
3. Дисперсия = 1,060,239

В коригиран вид:

1. Локация = 11.61954;
2. Очакване = 11.61427;
3. Дисперсия = 1.645619
4. Активна човешка военна сила (Active military personnel)

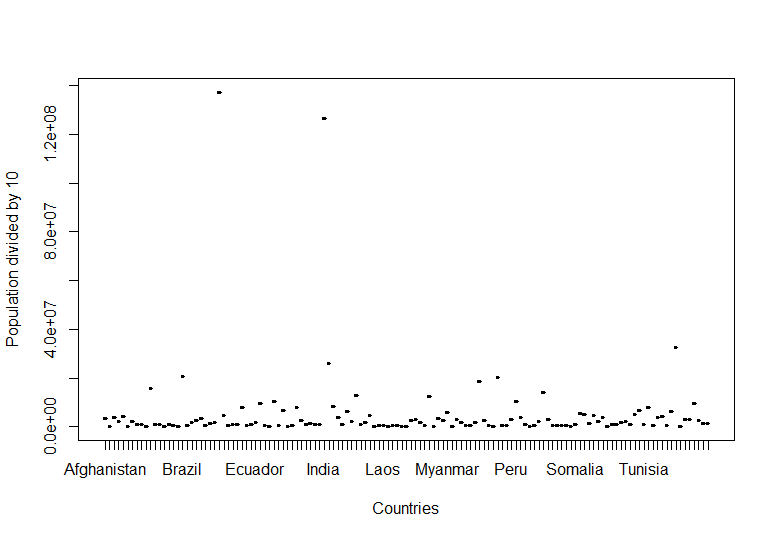
Армията на всяка държава се разделя на активен участък и резервен участък. Този вектор от числени стойности описва активния участък от армиите в тази извадка от данни.

1. Хистограма

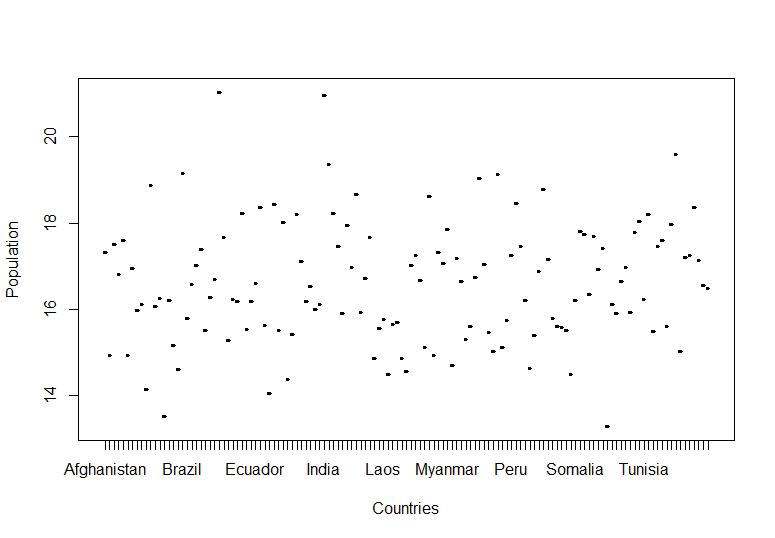
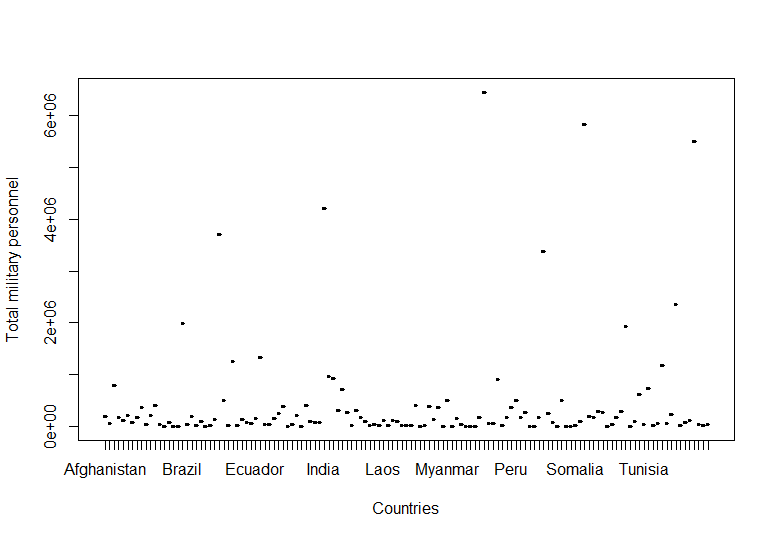
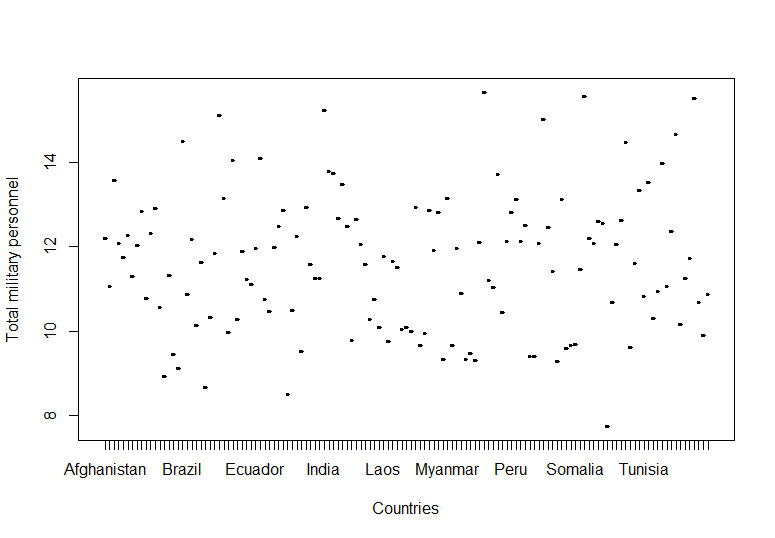
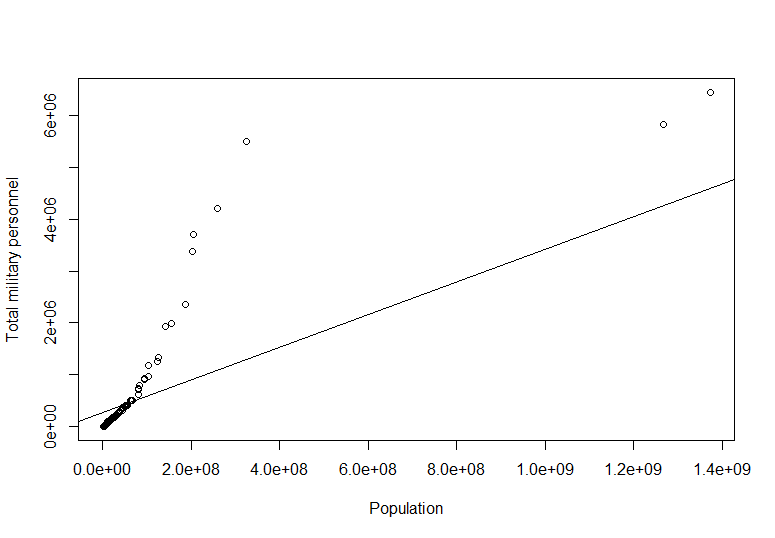
В нормален вид:

1. Локация = 45,000;
2. Очакване = 150,872;
3. Дисперсия = 289,545

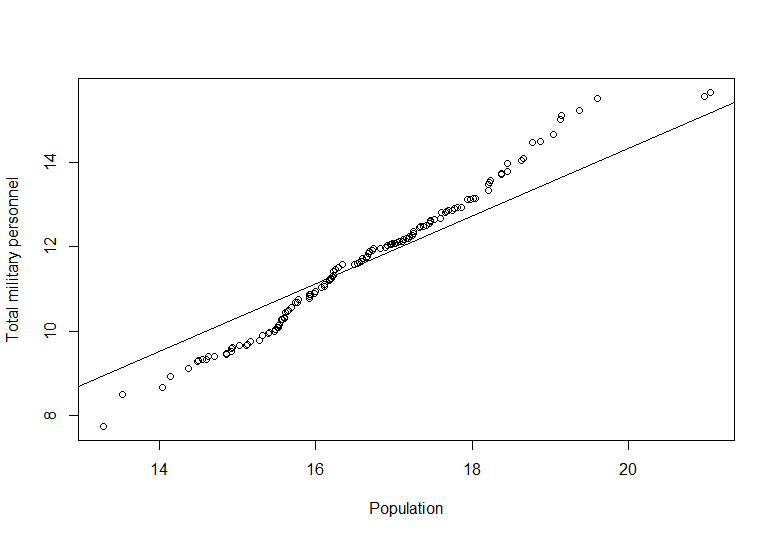
В коригиран вид:

1. Локация = 10.71442;
2. Очакване = 10.898;
3. Дисперсия = 1.426126
4. **Взаимодействие между променливи**
5. Между държава и общо население
6. Нормален вид

Забележка: Нужно е да се раздели населението на 10, за да се избегне преливане на числото и представяне на некоректна информация.

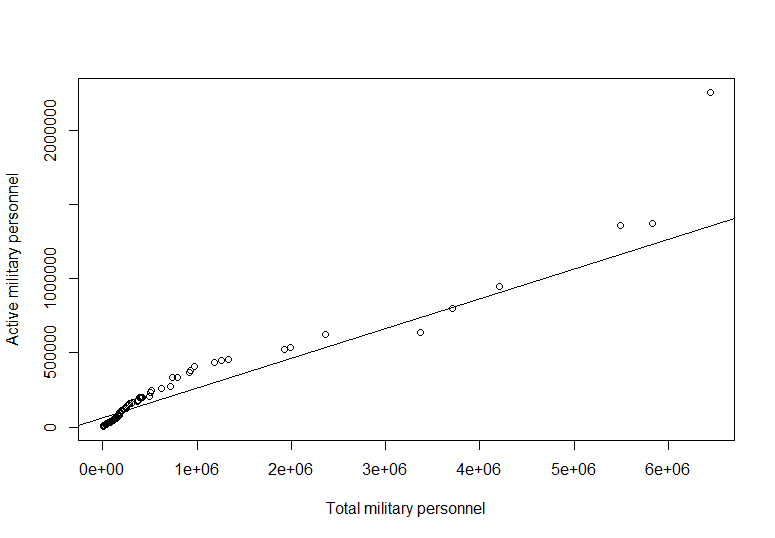
1. Коригиран вид
2. Между държава и обща човешка военна сила
3. Нормален вид
4. Коригиран вид
5. Между общо население и обща човешка военна сила
6. Нормален вид

Корелация по метода на Spearman = 0.6630594

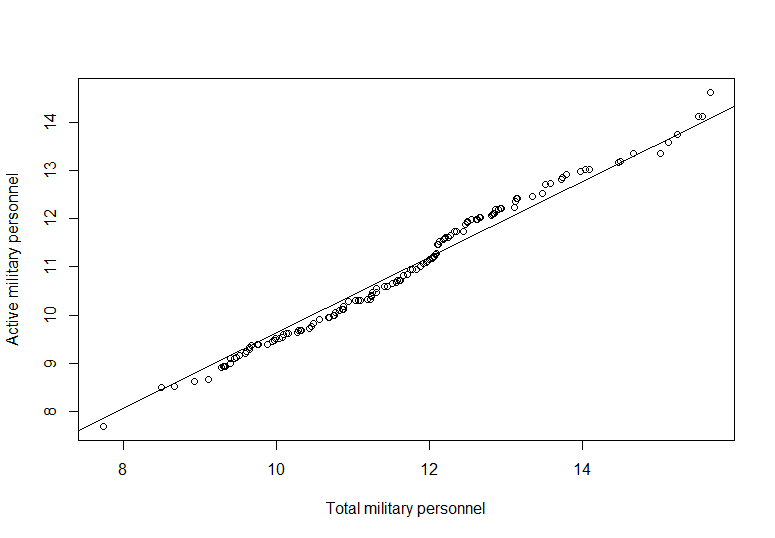
1. Коригиран вид

Корелация по метода на Pearson = 0.6904726

1. Между обща човешка военна сила и активна човешка военна сила
2. Нормален вид



Корелация по метода на Spearman = 0.8960795

1. Коригиран вид

Корелация по метода на Pearson = 0.9032227

1. **Заключение**
2. Данните в първоначалния си вид не са удобни за изследване, поради което е нужно да се коригират. Трансформацията на данните чрез функция log води до следните промени:
3. Хистограмите на числените стойности се приближават към нормалното разпределение;
4. Разликата между локацията и очакването се смалява значително, което също е признак за нормализиране на данните;
5. Дисперсията намалява драстично като резултат от изчезването на outlier-ите
6. Boxplot и QQ-plot графиките също придобиват по-нормално разпределение като резултат от изчезването на outlier-ите
7. Липсата на сериозни outlier-и позволява корелацията на две числени променливи по-точно да се определи чрез метода на Pearson вместо да е нужно да се използва метода на Spearman, който е устойчив към outlier-и, но по-неточен
8. Анализиране на данните по двойки:
9. Boxplot визуализира графично числените данни спрямо всяка държава и така става възможно нагледно да се представи как дадена променлива като общо население и човешка военна сила варира между държавите. Вижда се, че:

* От гледна точка на населението, почти всички държави имат много малко като брой население, сравнено с outlier-ите Индия и Китай
* От гледна точка на общата човешка военна сила има по-голям на брой outlier-и сравнено с населението, което обяснява не толкова силната връзка между населението и общата човешка военна сила.

1. QQ-plot визуализира връзката между две числени променливи. Като резултат от корелационния анализ става ясно, че:

* Има средна по сила връзка между населението и общата човешка военна сила, т.е не може да се каже, че във всеки случай колкото повече расте населението на някоя държава това ще доведе до толкова по-голямо нарастване на общата човешка военна сила.
* Има силна връзка между общата човешка военна сила и активната човешка военна сила, т.е с малко изключения растежа на общата човешка военна сила на някоя държава ще доведе до растеж в същата степен на активната човешка военна сила.