SI 424: Statistical Inference
Project Report

Krushnakant Bhattad	Devansh Jain
190100036	190100044

Autumn 2021

Project Title:

Experiments on Parameter Estimation and Hypothesis Testing

Abstract:

We first devise two problems on parameter estimation. We perform experiments for different sample size and observe the varied value of estimates.
We also devise two problems on hypothesis testing. We perform experiments for different sample size and different true values and observe the variation in Type I error and Type II error.

Associated GitHub Repository:

The GitHub repository can accessed at: `https://github.com/devansh-dvj/SI424-Project`

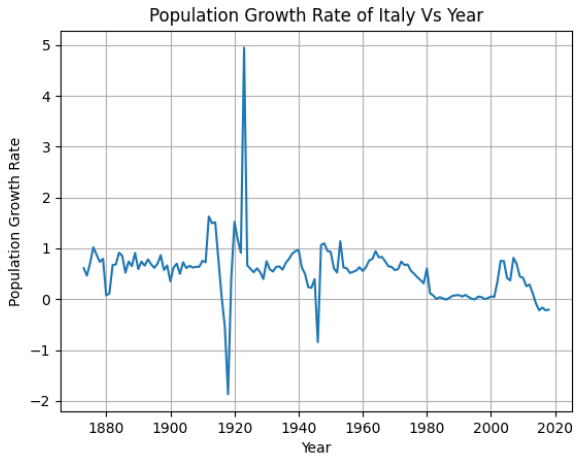
Common Notations:

$\mathcal{N}(\mu, \sigma^2)$ denotes the Normal distribution with mean μ and variance σ^2.
Binomial(p, n) denotes the Binomial distribution with probability of success p and number of trials n.
Uniform$[\theta_1, \theta_2]$ denotes the Uniform distribution with lower bound θ_1 and upper bound θ_2.
mean(x) denotes mean of elements in vector x.
max(x) denotes maximum of elements in vector x.
min(x) denotes minimum of elements in vector x.

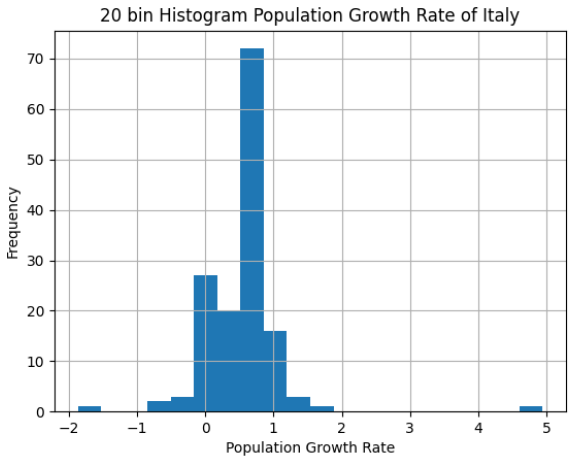
---

1 Parameter estimation Problem 1

1.1 Analysis of Population Growth Rate



(a) Plot of Population Growth Rate of Italy Vs Year



(b) Histogram of Population Growth Rate of Italy

Figure 1: Analysis of Population Growth Rate of Italy

From the above obtained plot and histogram, we can approximate population growth rate of Italy as a constant μ with zero-mean additive Gaussian noise, i.e. population growth rate of Italy $\sim \mathcal{N}(\mu, \sigma^2)$.

1.2 Problem Description

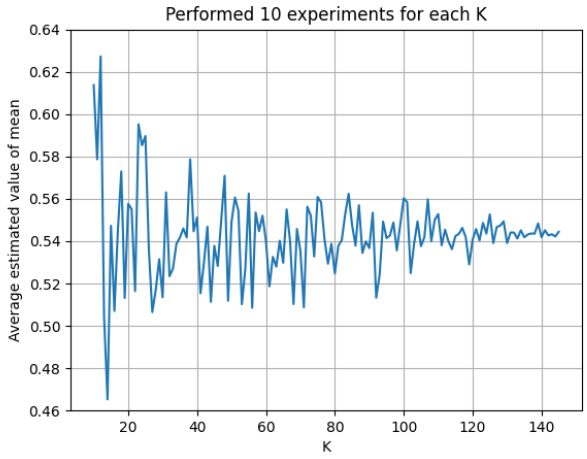
We assume that the population growth rate of Italy follows $\mathcal{N}(\mu, \sigma^2)$ distribution.
We are given a sample of population growth rate $g = (g_1, \cdots, g_K)$ where K is the sample size.
Estimated values are $\hat{\mu}_{MLE} = \text{mean}(g)$ and $\hat{\sigma}_{MLE} = \sqrt{\text{mean}((g - \hat{\mu}_{MLE})^2)}$.
We perform this experiment for N iterations for a fixed K.
We analyze the observed estimated of μ and σ for different K and N.

1.3 Generation of Sample

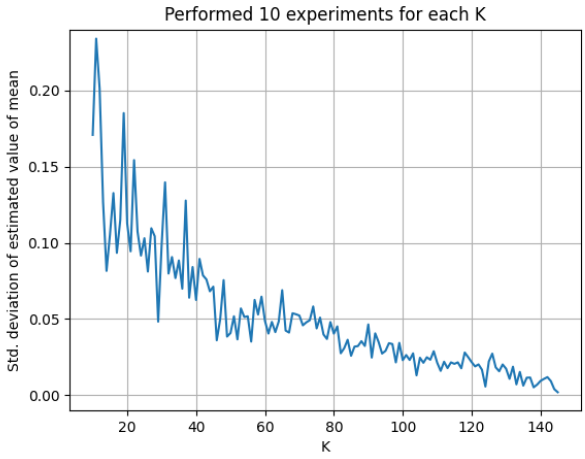
From `population_Country.csv`, we determine year-wise population growth rate for Italy from 1872 to 2018 (list of size 147).
The population growth rates are computed as percentage change in population (using `Total2` column) for every year.
Our K-sized sample is a randomly chosen K-sized subset ($K \leq 147$) of the above obtained list.
We use function `numpy.random.choice` with suitable arguments to obtain this sample.

1.4 Results

1.4.1 Estimate of μ

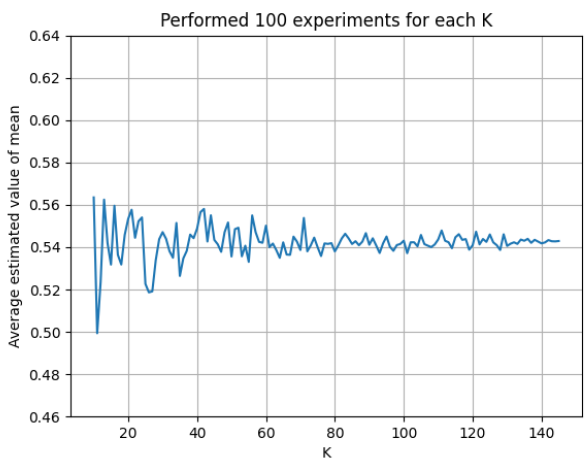


(a) Average value of $\hat{\mu}_{MLE}$ observed

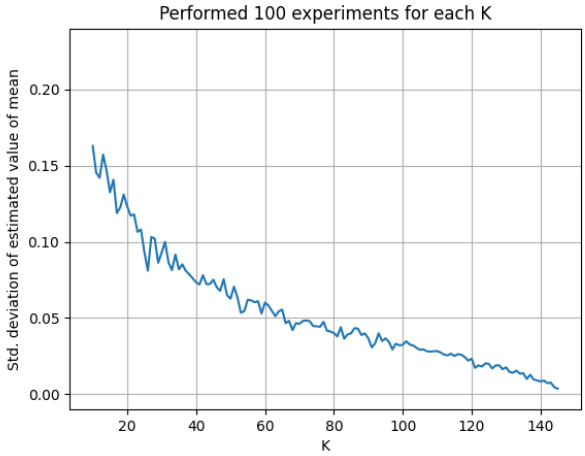


(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

Figure 2: Number of iterations, $N = 10$

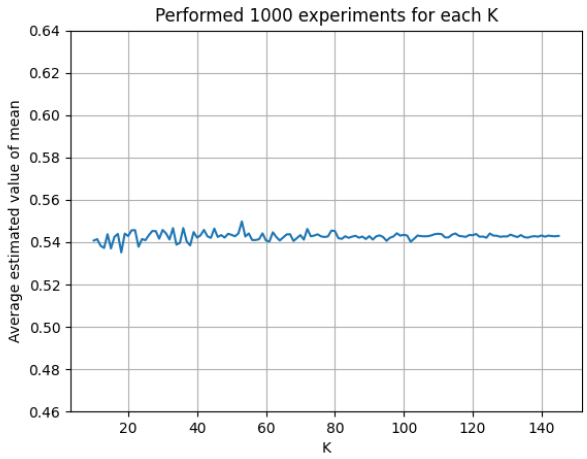


(a) Average value of $\hat{\mu}_{MLE}$ observed

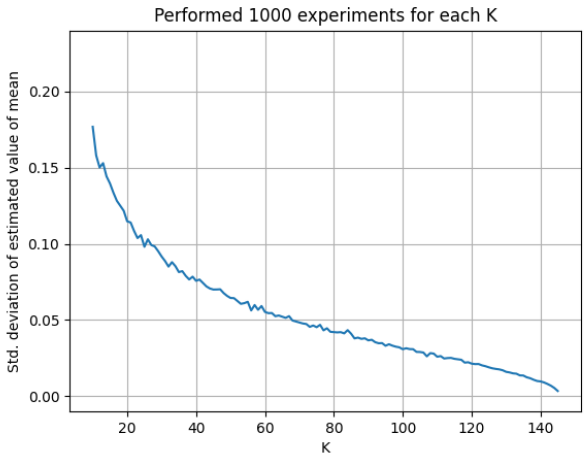


(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

Figure 3: Number of iterations, $N = 100$



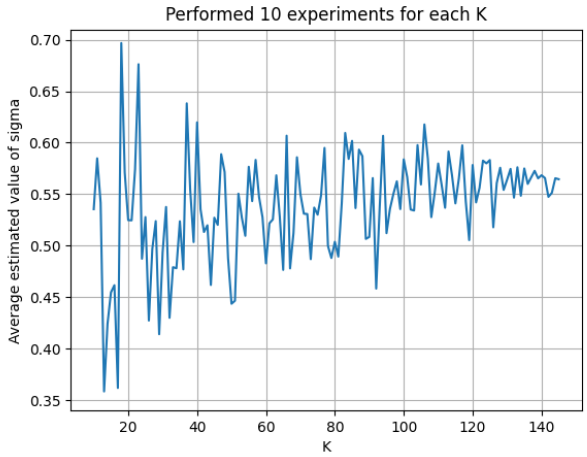
(a) Average value of $\hat{\mu}_{MLE}$ observed



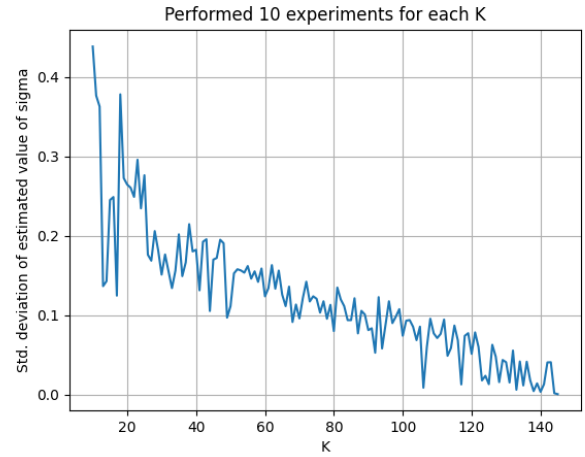
(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

Figure 4: Number of iterations, $N = 1000$

1.4.2 Estimate of σ

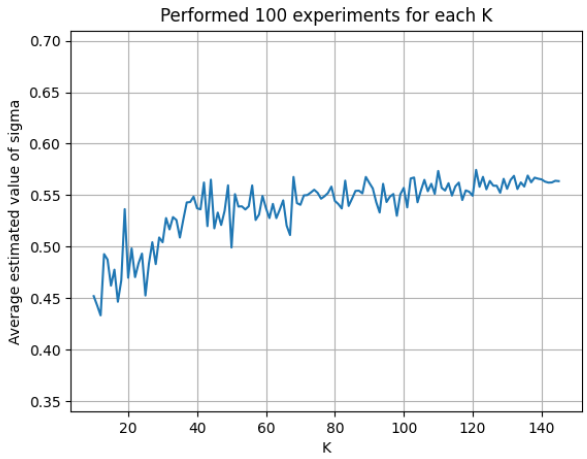


(a) Average value of $\hat{\sigma}_{MLE}$ observed

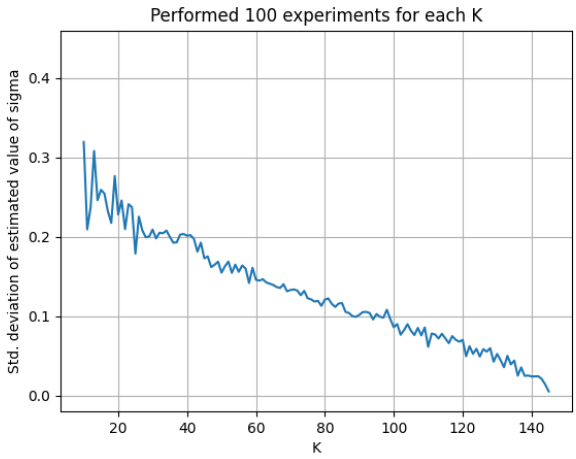


(b) Std. deviation of $\hat{\sigma}_{MLE}$ observed

Figure 5: Number of iterations, $N = 10$

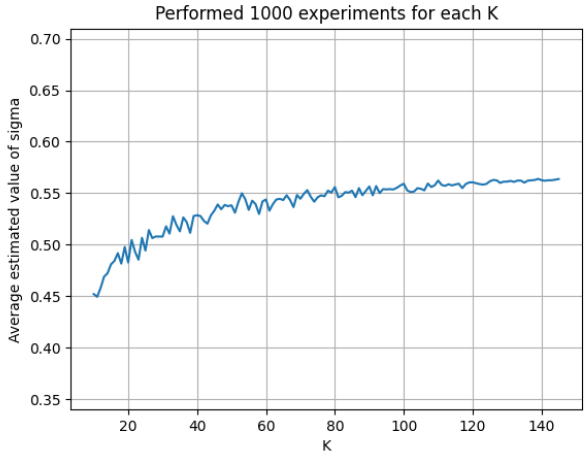


(a) Average value of $\hat{\sigma}_{MLE}$ observed

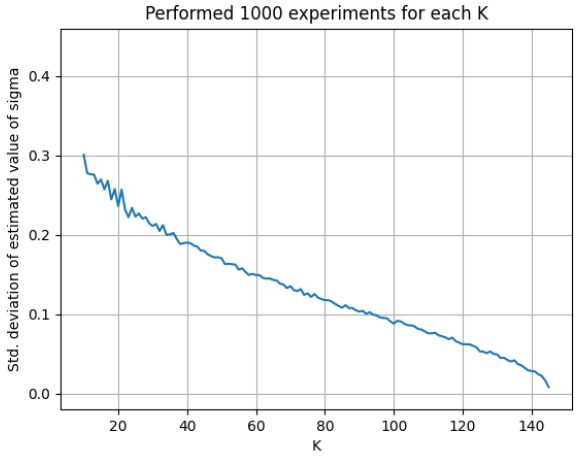


(b) Std. deviation of $\hat{\sigma}_{MLE}$ observed

Figure 6: Number of iterations, $N = 100$



(a) Average value of $\hat{\sigma}_{MLE}$ observed



(b) Std. deviation of $\hat{\sigma}_{MLE}$ observed

Figure 7: Number of iterations, $N = 1000$

1.5 Inference

- We can clearly observe that as N increases, we are able to capture relation of variability of the estimates with K very well.
- For large N, the expected value of estimate of μ for large K is consistent (0.543).
- The expected value of estimate of σ for large N is dependent on K because $\hat{\sigma^2}_{MLE}$ is proportional to a chi squared RV which has degrees of freedom dependent upon K.
- For small N, the average estimates for both μ and σ vary more for small K.
  This can be explained by the large variance for small K.

2 Parameter estimation Problem 2

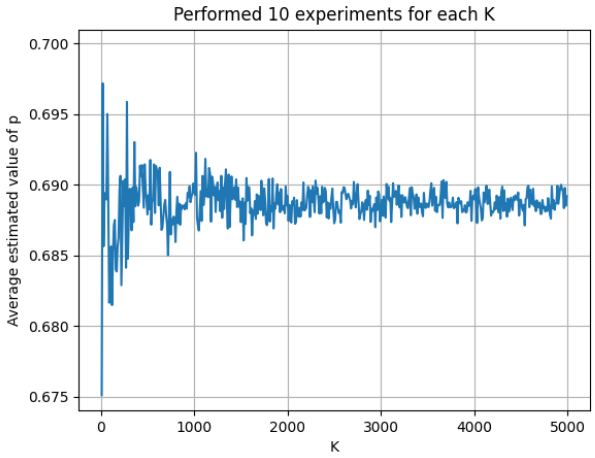
2.1 Description

The age of people who died in Greece in 2005 follows Binomial$(p, 110)$.
We have age $a = (a_1, \cdots, a_n)$ of people who died in Greece in 2005 given to us.
Estimated value of p is $\hat{p}_{MLE} = \text{mean}(a)/110$.

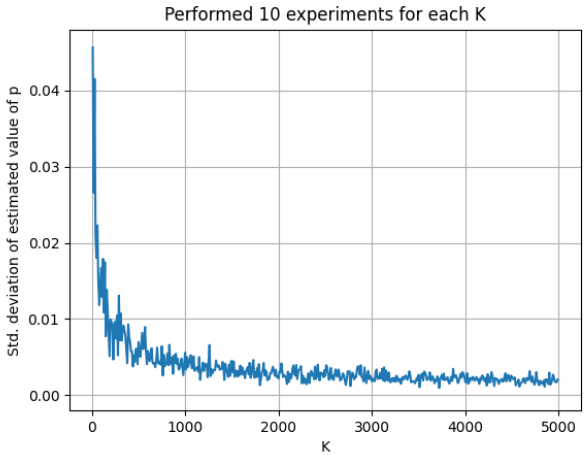
2.2 Experiment

`mortality_Country.csv` contains total deaths per age interval, we cannot generate a directly.
We extract percentage of deaths in Greece in 2005 for each age interval.
We define a procedure which returns age of a person based on this percentage distribution.
We do this by using the CDF generated using the obtained percentage distribution and randomly sample real value in $[0, 1]$ and use this to get the age.
Using the above defined procedure, we generate a K sized list of ages, and compute the estimate of p ($\hat{p}_{MLE}$). We repeat this for N iterations.
We intend to observe the variation in our estimate of p for different K and N.

2.3 Results

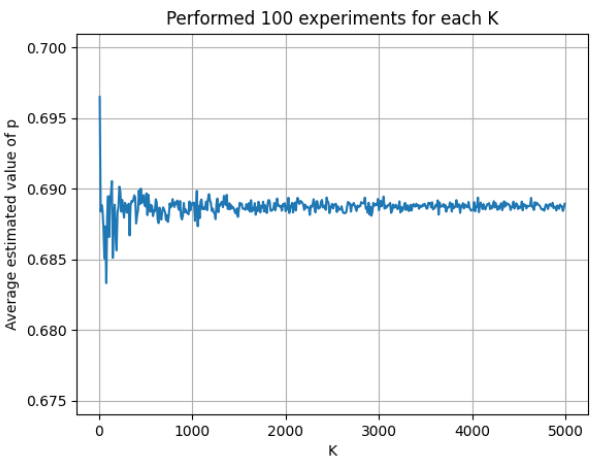


(a) Average value of $\hat{p}_{MLE}$ observed

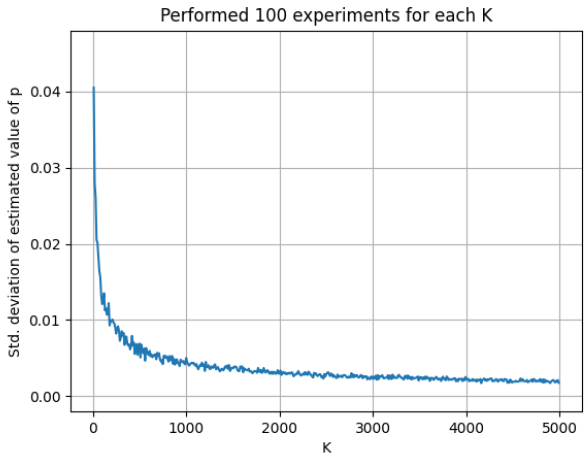


(b) Std. deviation of $\hat{p}_{MLE}$ observed

Figure 1: Number of iterations, $N = 10$

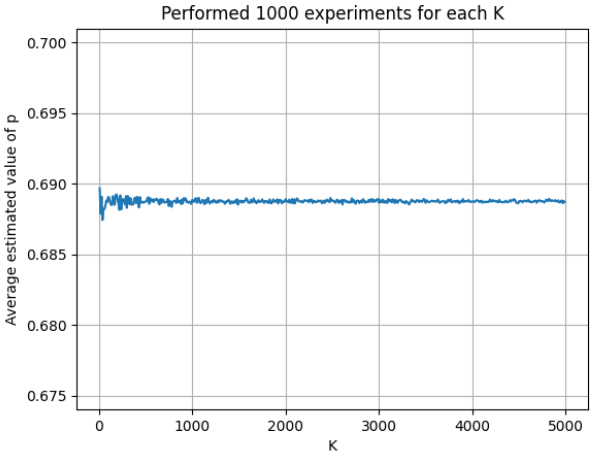


(a) Average value of $\hat{p}_{MLE}$ observed

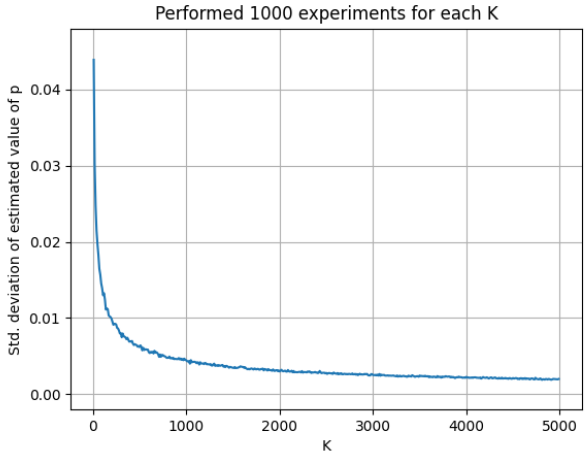


(b) Std. deviation of $\hat{p}_{MLE}$ observed

Figure 2: Number of iterations, $N = 100$



(a) Average value of $\hat{p}_{MLE}$ observed



(b) Std. deviation of $\hat{p}_{MLE}$ observed

Figure 3: Number of iterations, $N = 1000$

2.4 Inference

- We can clearly observe that as N increases, we are able to capture relation of variability of the estimated value with K very well.
- For large N, the expected value of estimate for all K is consistent (almost 0.689).
  This sits well with law of large numbers.
- For small N, the average estimated varies more for small K.
  This can be explained by the large variance for small K.

3 Hypothesis Testing Problem 1

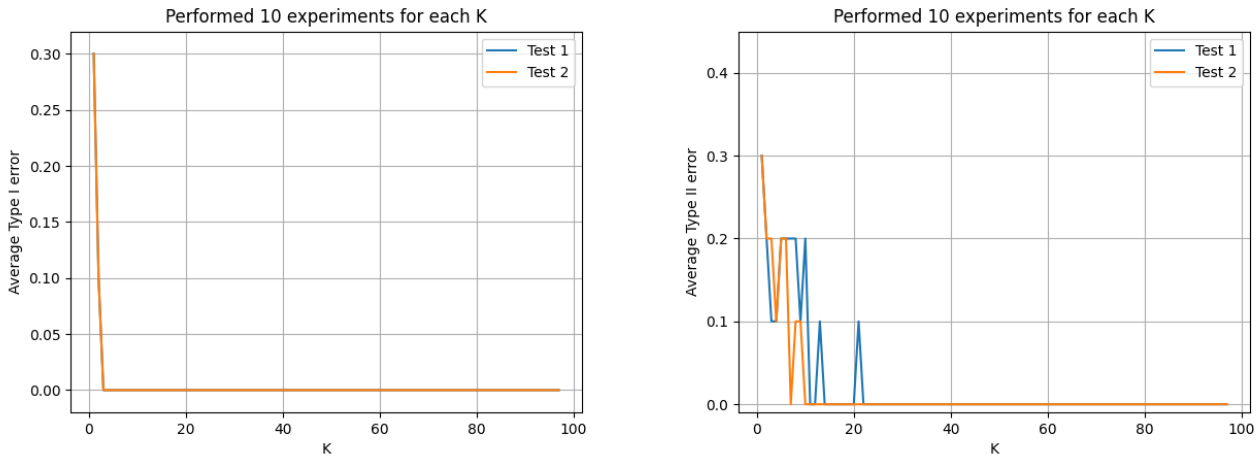
3.1 Description

We have two countries - Italy and Australia.
We have gender ratios $r = (r_1, \cdots, r_n)$ of different years given to us for a specific country.
We observe that $r_i \sim \text{Uniform}[\theta - \beta, \theta + \beta]$.
Given a sample, we hypothesise H0 (Null hypothesis) and H1 (Alternative hypothesis).
H0: Country is Italy
H1: Country is Australia
We perform following tests to accept/reject H0.
Test1: Reject H0 if $\text{mean}(r) < 1.0$
Test2: Reject H0 if $\hat{\theta}_{MLE} = \dfrac{\max(r) + \min(r)}{2} < 1.0$.

3.2 Experiment

From `population_Country.csv`, we extract one unordered list of gender ratios for both countries.
The gender ratios are computed as ratio of total female population (using `Female2` column) and total male population (using `Male2` column) for every year.
The size of the unordered list is 147 for Italy and 98 for Australia.
We randomly choose a K sized subset ($K \leq 98$) of one of the lists. We compare the output of both the tests with the true value. We repeat this for $2N$ iterations, where both countries have true value for N iterations (to avoid dominance of either side of hypothesis).
We intend to observe the values of Type I error and Type II error for both the tests for different K and N.

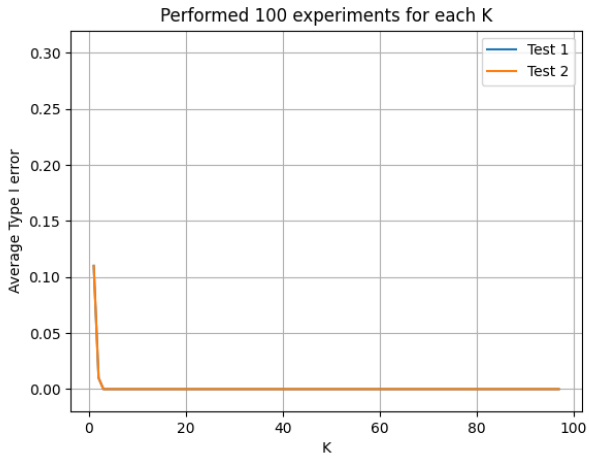
3.3 Results



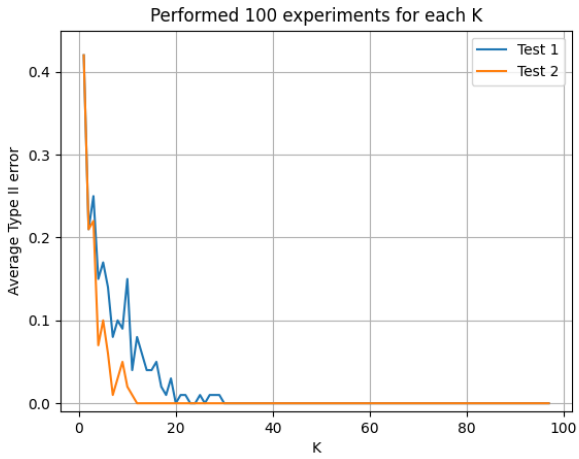
(a) Average value of Type I error observed

(b) Average value of Type II error observed

Figure 1: Number of iterations for each side of hypothesis as truth, $N = 10$

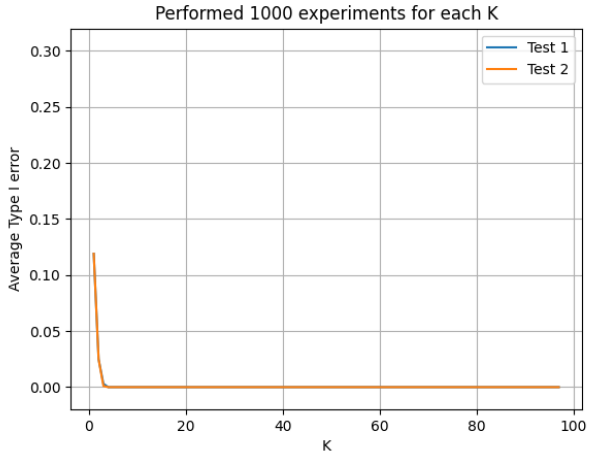


(a) Average value of Type I error observed

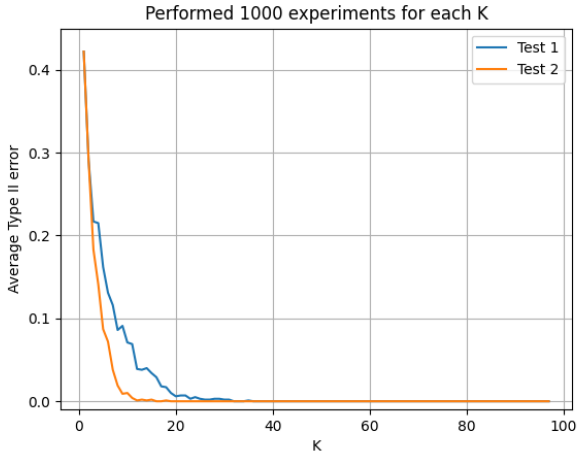


(b) Average value of Type II error observed

Figure 2: Number of iterations for each side of hypothesis as truth, $N = 100$

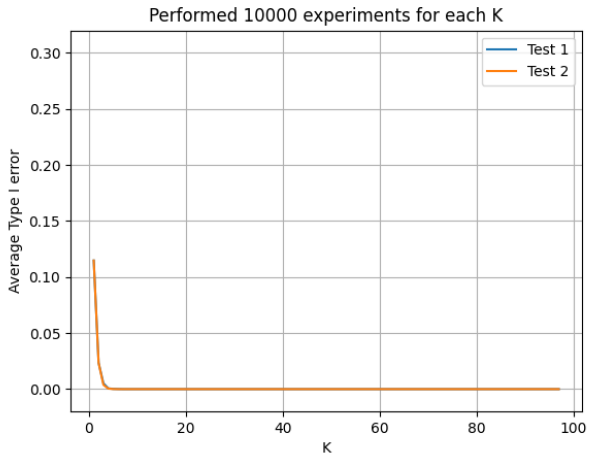


(a) Average value of Type I error observed

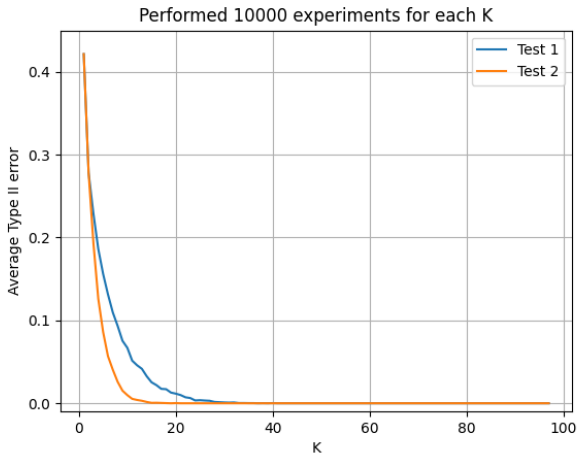


(b) Average value of Type II error observed

Figure 3: Number of iterations for each side of hypothesis as truth, $N = 1000$



(a) Average value of Type I error observed



(b) Average value of Type II error observed

Figure 4: Number of iterations for each side of hypothesis as truth, $N = 10000$

3.4 Inference

- As N (i.e. the number of experiments for a given sample size) increases, we are able to capture relation of errors with K (i.e the sample size) very well.
- The average value of Type I and Type II error for both tests decreases with the sample size for all values of N.
- Type I error is same for both tests. Type II error is less for Test 2.
  We can conclude that Test 2 is more powerful than Test 1.

4 Hypothesis Testing Problem 2

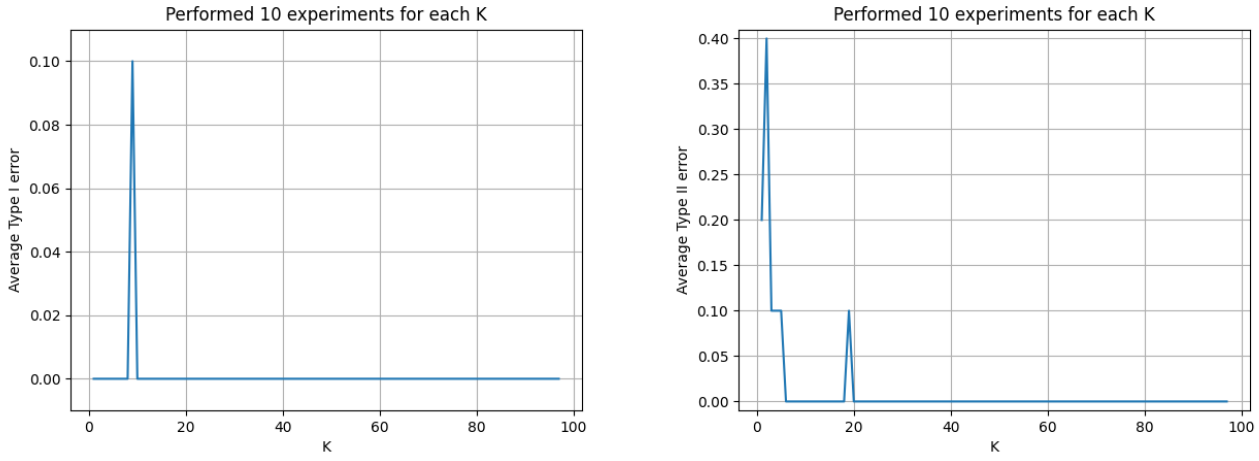
4.1 Description

We have two countries - Italy and Australia.
We have population growth rate $g = (g_1, \cdots, g_n)$ of different years given to us for a specific country.
We observe that, if the "specific country" is Italy, $g_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$; and if the "specific country" is Australia, $g_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
Given a sample, we hypothesise H0 (Null hypothesis) and H1 (Alternative hypothesis). Our tests are based on the sample mean (i.e. since g is our given sample, this would be mean(g) which is a sufficient statistic). Our samples belong to either Italy(where true mean is g_1) or Australia(where true mean is g_2).
H0: Country is Italy (i.e. $\mu = \mu_1$)
H1: Country is Not Italy (in which case the only possibility is that the country is Australia) (i.e. $\mu \neq \mu_1$, i.e. $\mu = \mu_2$)
We perform following test to accept/reject H0.
Test: Reject H0 if $\hat{\mu}_{MLE} = \text{mean}(g) > 1.2$

4.2 Experiment

From `population_Country.csv`, we determine population growth rate for both countries.
The population growth rates are computed as percentage change in population (using `Total2` column) for every year.
The size of the list is 147 for Italy and 98 for Australia.
We randomly choose a K sized subset ($K \leq 98$) from one of the lists. We compare the output of the test with the true value. We repeat this for $2N$ iterations, where both countries have true value for N iterations (to avoid dominance of either side of hypothesis).
We intend to observe the values of Type I error and Type II error for the test for different K and N.

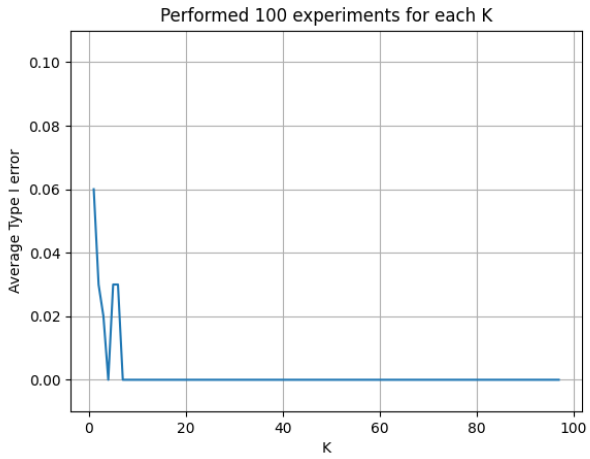
4.3 Results



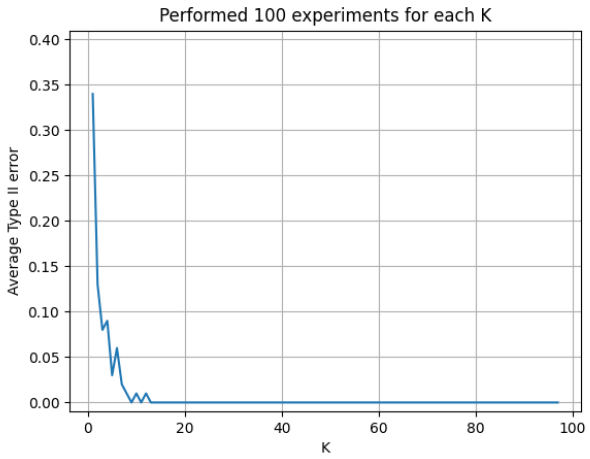
(a) Average value of Type I error observed

(b) Average value of Type II error observed

Figure 1: Number of iterations for each side of hypothesis as truth, $N = 10$

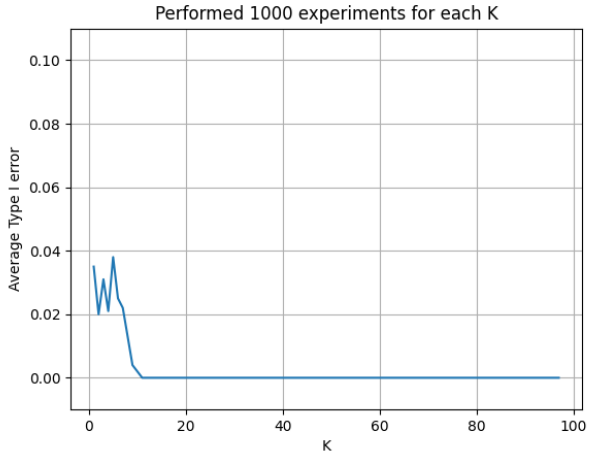


(a) Average value of Type I error observed

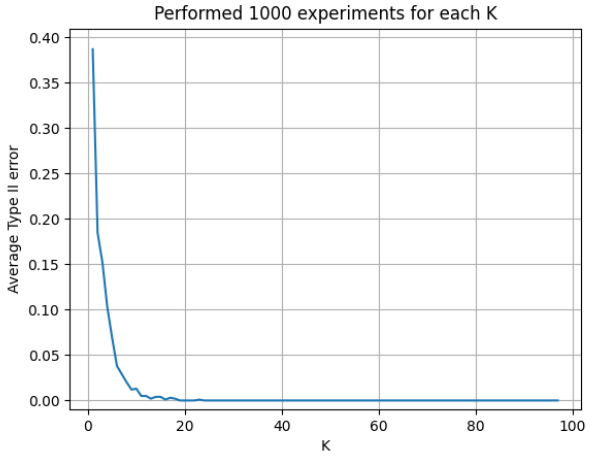


(b) Average value of Type II error observed

Figure 2: Number of iterations for each side of hypothesis as truth, $N = 100$

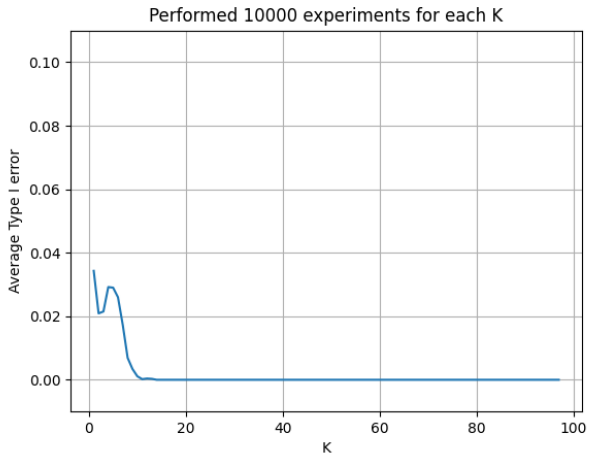


(a) Average value of Type I error observed

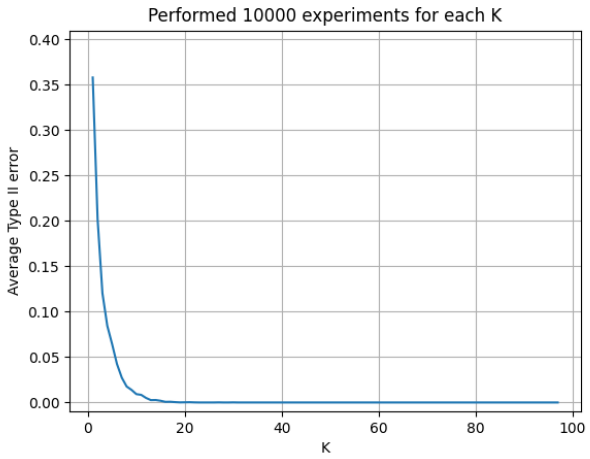


(b) Average value of Type II error observed

Figure 3: Number of iterations for each side of hypothesis as truth, $N = 1000$



(a) Average value of Type I error observed



(b) Average value of Type II error observed

Figure 4: Number of iterations for each side of hypothesis as truth, $N = 10000$

4.4 Inference

- As N (i.e. the number of experiments for a given sample size) increases, we are able to capture relation of errors with K (i.e the sample size) very well.
- The average value of Type I and Type II error decreases with the sample size for all values of N.
- The Type II error is more prevalent (i.e. has a higher probability of occurrence) for small values of K. At larger K values, both errors are practically zero.