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`

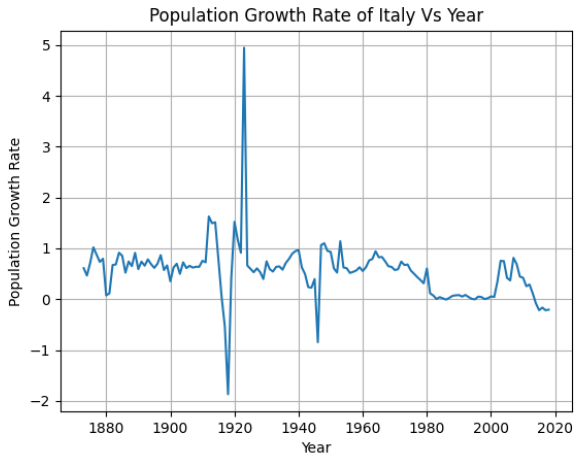
Common Notations:

$\mathcal{N}(\mu, \sigma^2)$ 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.

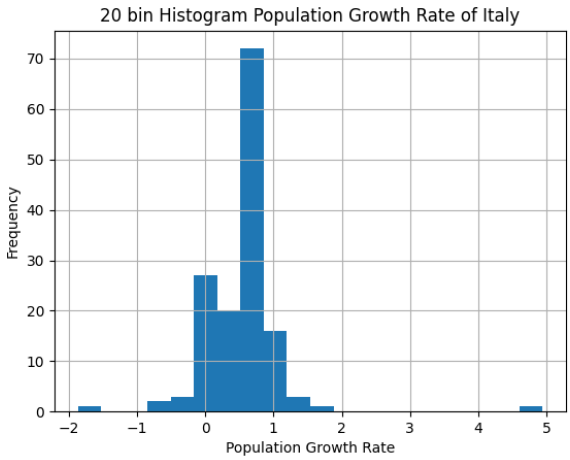
---

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

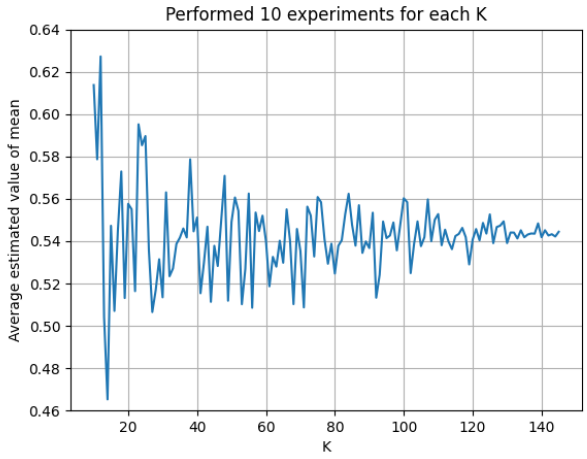
We assume that the population growth rate of Italy follows $\mathcal{N}(\mu, \sigma^2)$ distribution.
We are given a sample of population growth rate $g = (g_1, \ldots, g_K)$ where K is the sample size.
Estimated values are $\hat{\mu}_{MLE} = \text{mean}(g)$ and $\hat{\sigma}_{MLE} = \sqrt{\text{mean}((g - \hat{\mu}_{MLE})^2)}$.
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

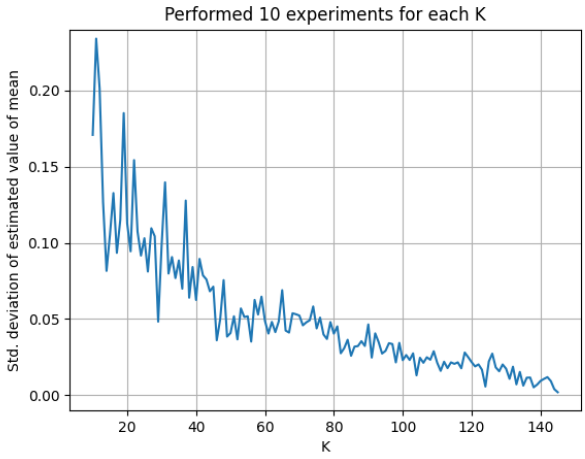
From `population_Country.csv`, we determine year-wise population growth rate for Italy from 1872 to 2018 (list of size 147).
The population growth rates are computed as percentage change in population (using `Total2` column) for every year.
Our K-sized sample is a randomly chosen K-sized subset ($K \leq 147$) 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

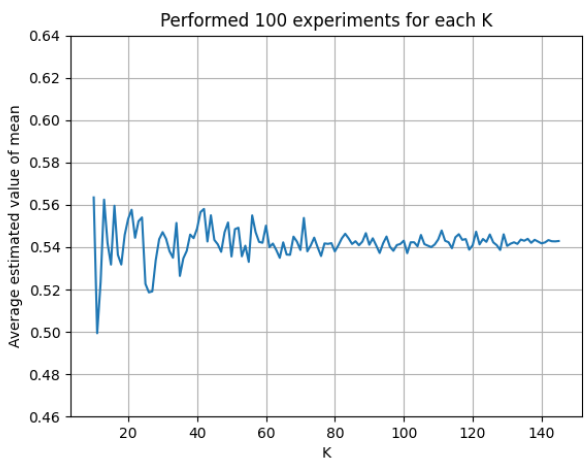


(a) Average value of $\hat{\mu}_{MLE}$ observed

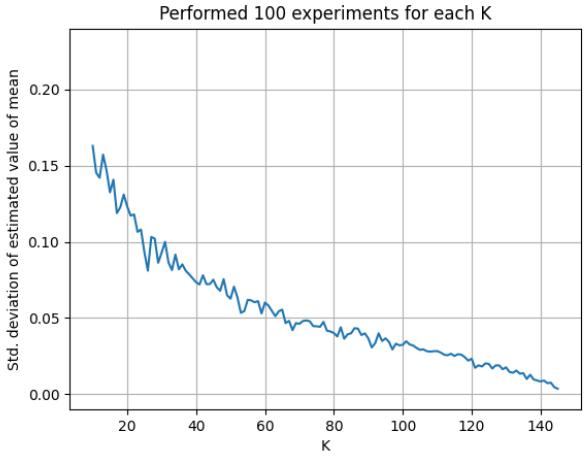


(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

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

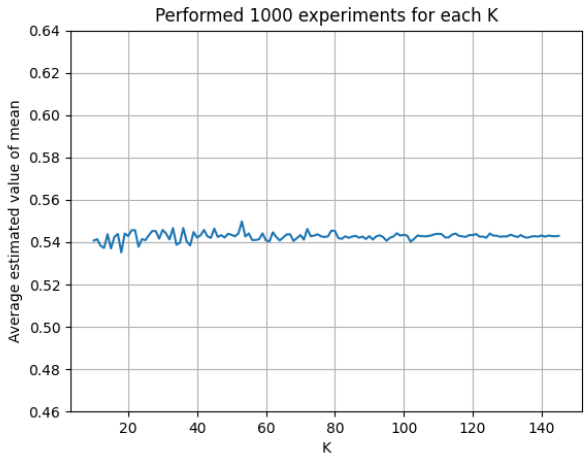


(a) Average value of $\hat{\mu}_{MLE}$ observed

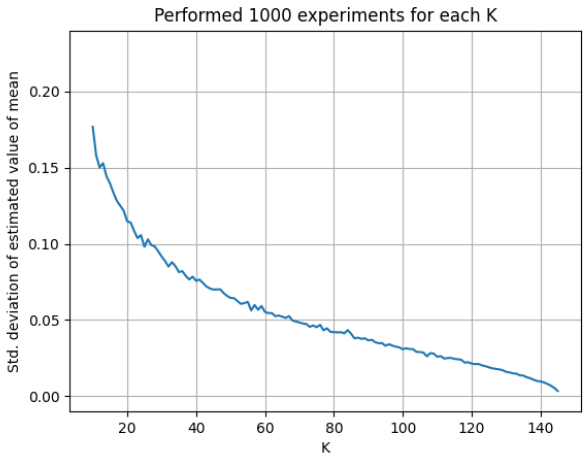


(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

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



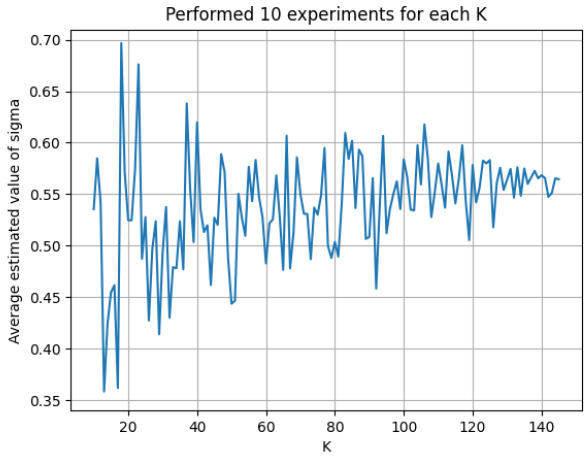
(a) Average value of $\hat{\mu}_{MLE}$ observed



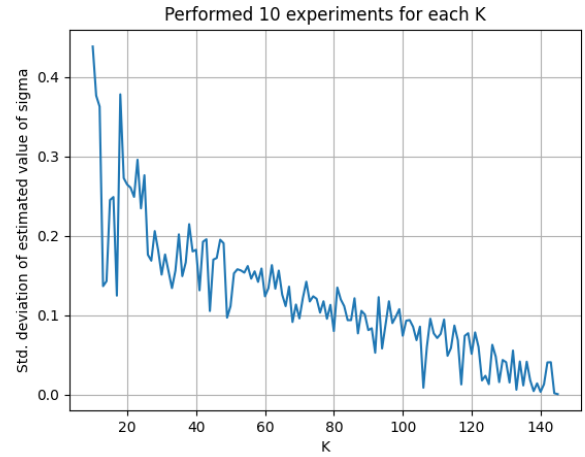
(b) Std. deviation of $\hat{\mu}_{MLE}$ observed

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

1.4.2 Estimate of σ

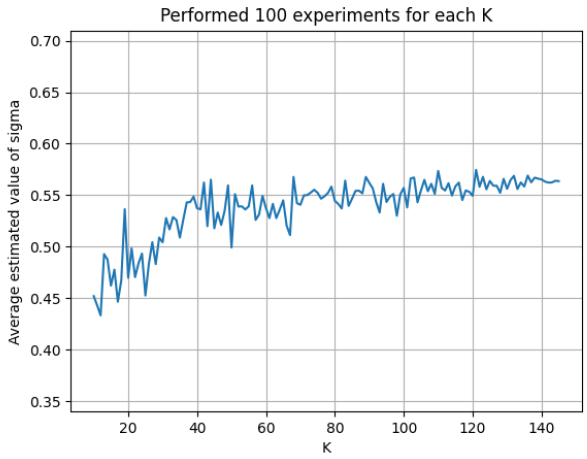


(a) Average value of $\hat{\sigma}_{MLE}$ observed

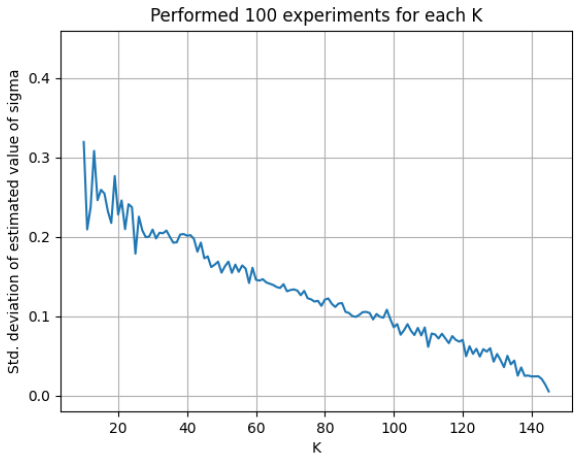


(b) Std. deviation of $\hat{\sigma}_{MLE}$ observed

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

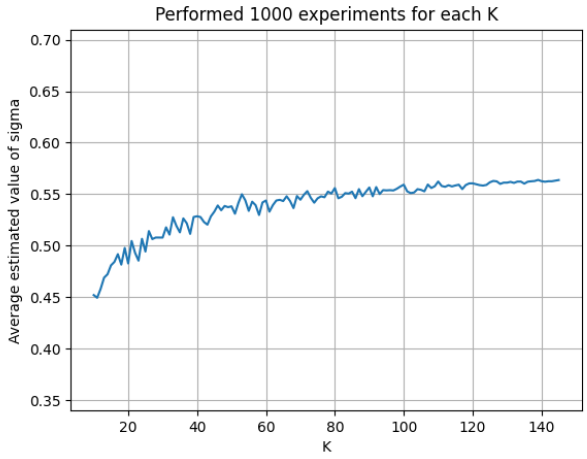


(a) Average value of $\hat{\sigma}_{MLE}$ observed

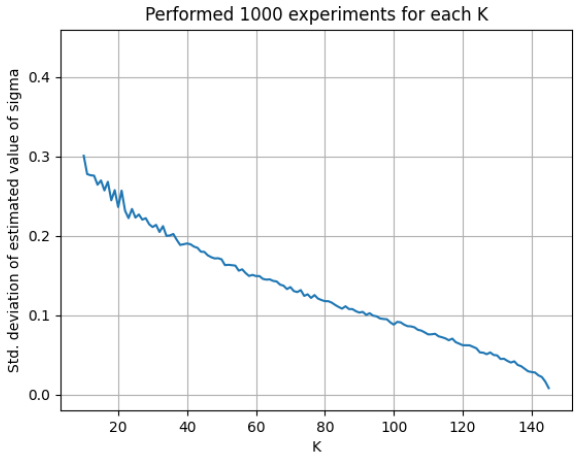


(b) Std. deviation of $\hat{\sigma}_{MLE}$ observed

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



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



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

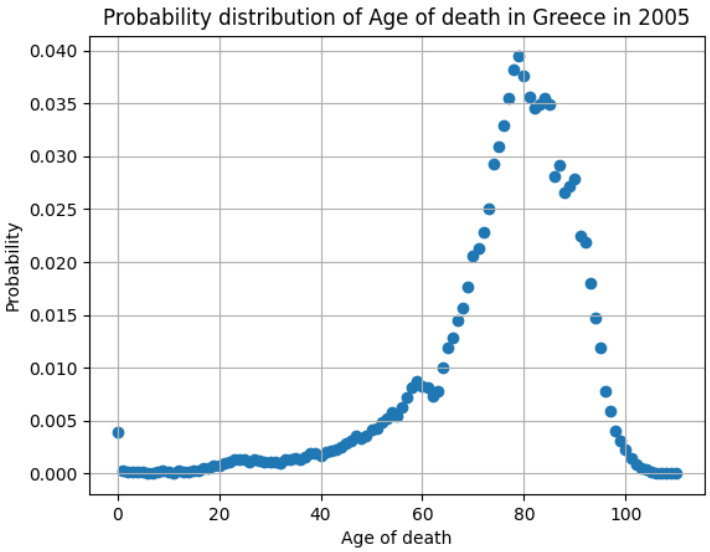
Figure 7: Number of iterations, $N = 1000$

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 Analysis of Age of death



(a) Probability distribution of Age of death in Greece in 2005

Figure 1: Analysis of Age of death in Greece in 2005

We can approximate the obtained probability mass function as Binomial$(p, 110)$.
Here, the random variable is the age of death in Greece in 2005 which is a discrete random variable taking value in $\{0, \ldots, 110\}$.

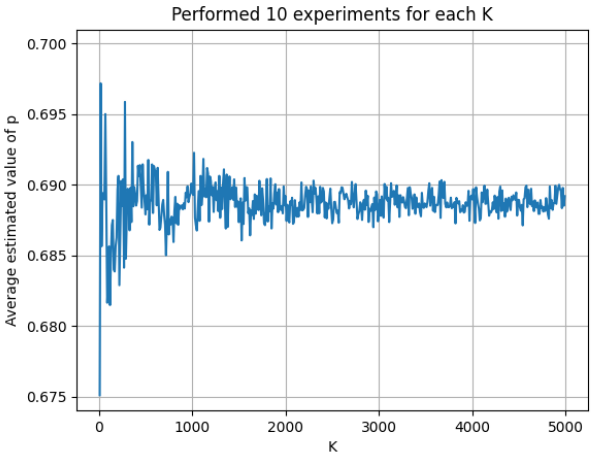
2.2 Problem Description

We assume that the age of people who died in Greece in 2005 follows Binomial$(p, 110)$.
We are given a sample of age of death $a = (a_1, \ldots, a_K)$ where K is the sample size.
Estimated value of p is $\hat{p}_{MLE} = \text{mean}(a)/110$. We perform this experiment for N iterations for a fixed K.
We analyze the observed estimated of p for different K and N.

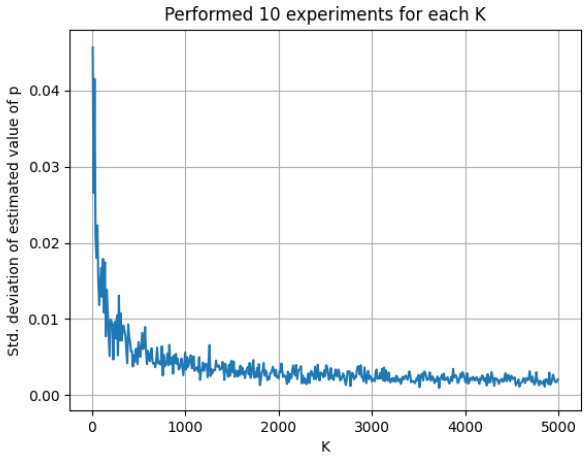
2.3 Generation of Sample

`mortality_Country.csv` contains total deaths per age interval, we cannot generate a directly.
We extract probability distribution of age of death in Greece in 2005 for each age interval.
Our K-sized sample is a randomly generated K-sized list of ages with above obtained probability distribution.
We use function `numpy.random.choice` with suitable arguments to obtain this sample.

2.4 Results

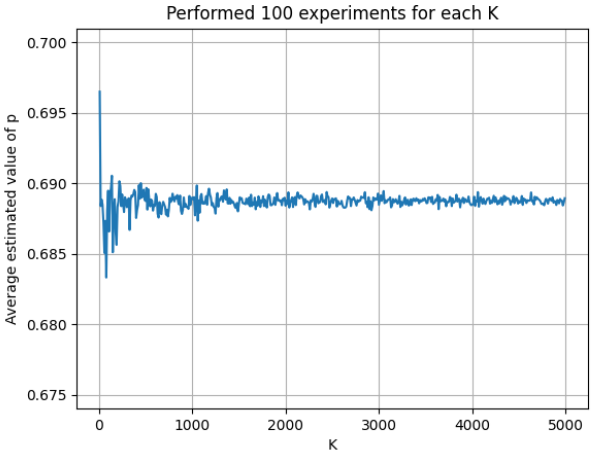


(a) Average value of $\hat{p}_{MLE}$ observed

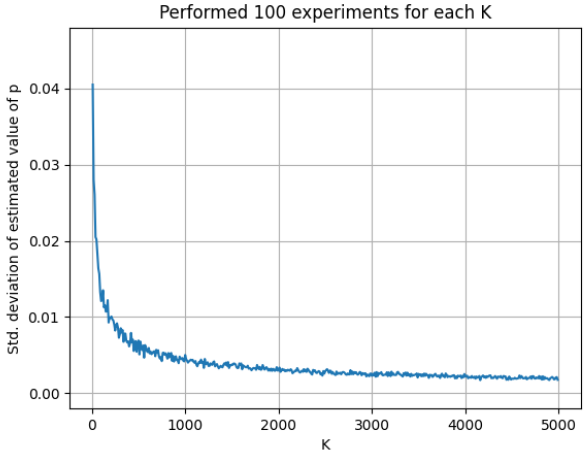


(b) Std. deviation of $\hat{p}_{MLE}$ observed

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

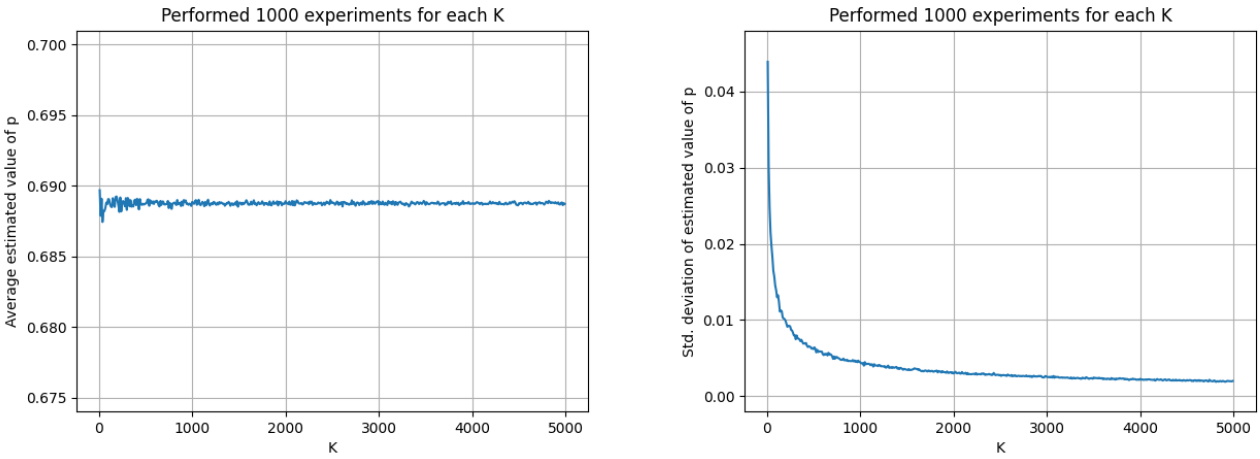


(a) Average value of $\hat{p}_{MLE}$ observed



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

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



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

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

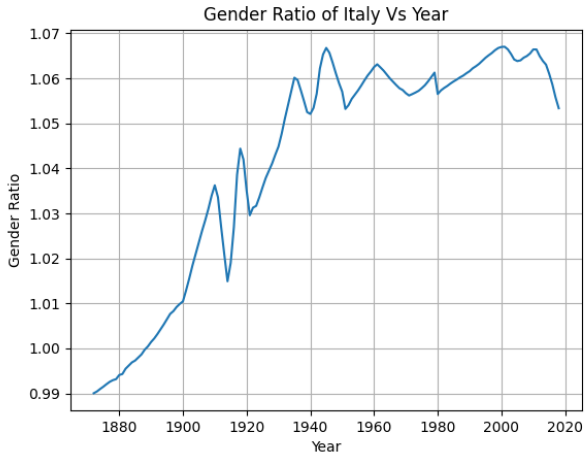
Figure 4: Number of iterations, $N = 1000$

2.5 Inference

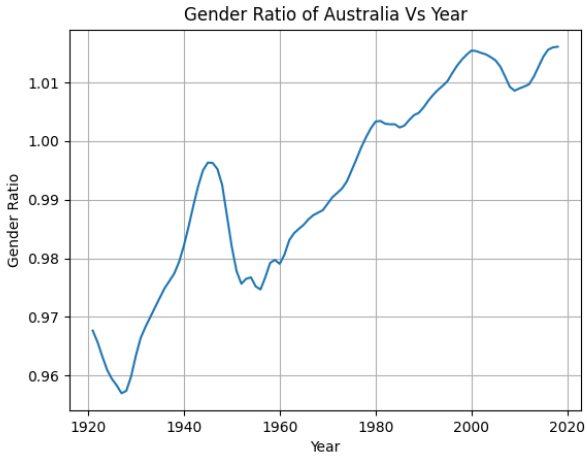
- We can clearly observe that as N increases, we are able to capture relation of variability of the estimated value with K very well.
- For large N, the expected value of estimate for all K is consistent (almost 0.689).
  This sits well with law of large numbers.
- For small N, the average estimated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

3 Hypothesis Testing Problem 1

3.1 Analysis of Gender Ratio



(a) Plot of Gender Ratio of Italy Vs Year



(b) Plot of Gender Ratio of Australia Vs Year

Figure 1: Analysis of Gender Ratio for Italy and Australia

From the above obtained plot, we can approximate gender ratios to be linear w.r.t. year.
Thus, we can conclude that gender ratio of Italy $\sim$ Uniform$[\theta_1 - \beta_1, \theta_1 + \beta_1]$.
Similarly, gender ratio of Australia $\sim$ Uniform$[\theta_2 - \beta_2, \theta_2 + \beta_2]$.

3.2 Problem Description

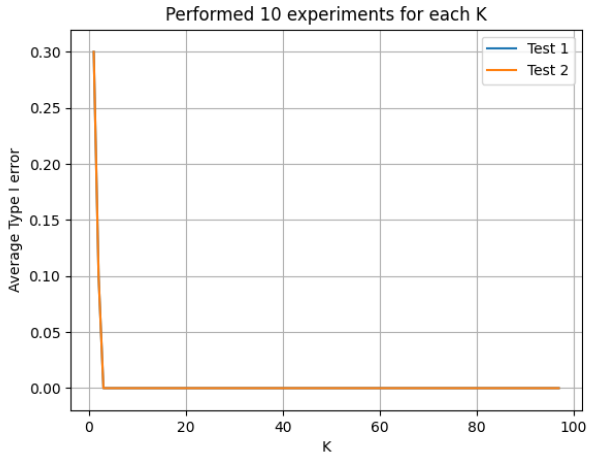
We assume that the gender ratio of Italy follows Uniform$[\theta_1 - \beta_1, \theta_1 + \beta_1]$ distribution and that of Australia follows Uniform$[\theta_2 - \beta_2, \theta_2 + \beta_2]$ distribution.
We are given a sample of gender ratio $\mathbf{r} = (r_1, \ldots, r_K)$ where K is the sample size which belongs to one of the above two mentioned distributions.
We devise a simple hypothesis:
H0: Country is Italy (i.e. $\theta = \theta_1$)
H1: Country is Australia (i.e. $\theta = \theta_2$)
We perform following tests to accept/reject H0.
Test1: Reject H0 if $\text{mean}(\mathbf{r}) < 1.0$
Test2: Reject H0 if $\hat{\theta}_{MLE} = \dfrac{\max(\mathbf{r}) + \min(\mathbf{r})}{2} < 1.0$
We perform this experiment for $2N$ iterations for a fixed K, choosing each distribution as the true distribution for N iterations.
We analyze the observed values of Type I and Type II errors for the tests for different K and N.

3.3 Generation of Sample

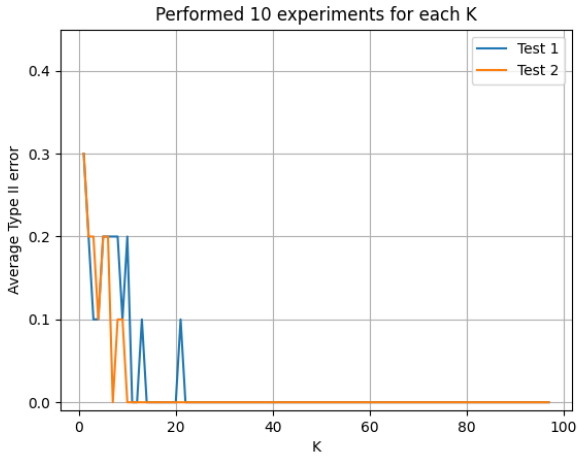
From `population_Country.csv`, we determine year-wise gender ratio for Italy from 1872 to 2018 (list of size 147) and for Australia from 1921 to 2018 (list of size 98).
The gender ratios are computed as ratio of total female population (using `Female2` column) and total male population (using `Male2` column) for every year.
Our K-sized sample is a randomly chosen K-sized subset ($K \leq 98$) of one of the above obtained list.

For first N iterations, we choose from list of Italy. For latter N iterations, we choose from list of Australia. We use function `numpy.random.choice` with suitable arguments to obtain this sample.

3.4 Results

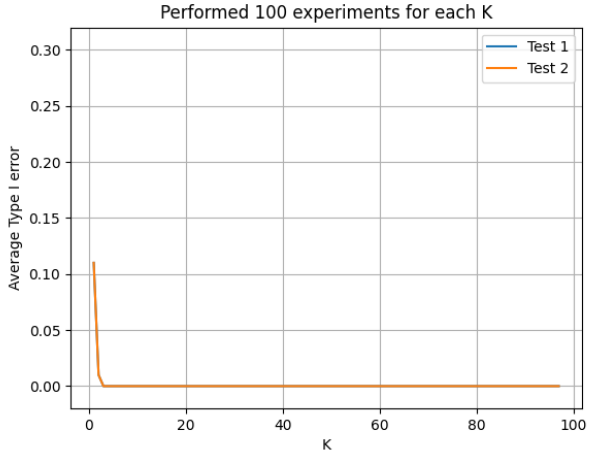


(a) Average value of Type I error observed

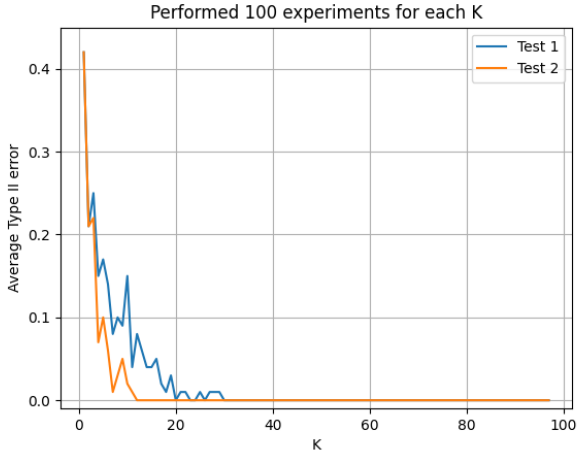


(b) Average value of Type II error observed

Figure 2: Number of iterations for each side of hypothesis as truth, $N = 10$

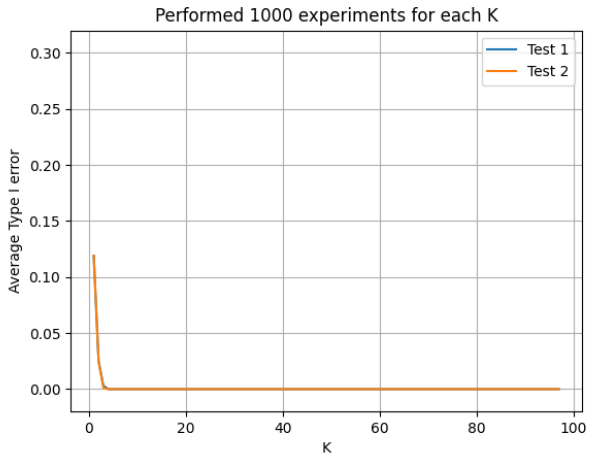


(a) Average value of Type I error observed

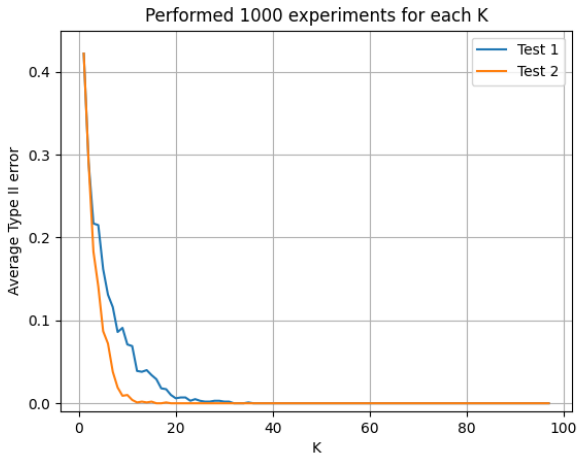


(b) Average value of Type II error observed

Figure 3: Number of iterations for each side of hypothesis as truth, $N = 100$

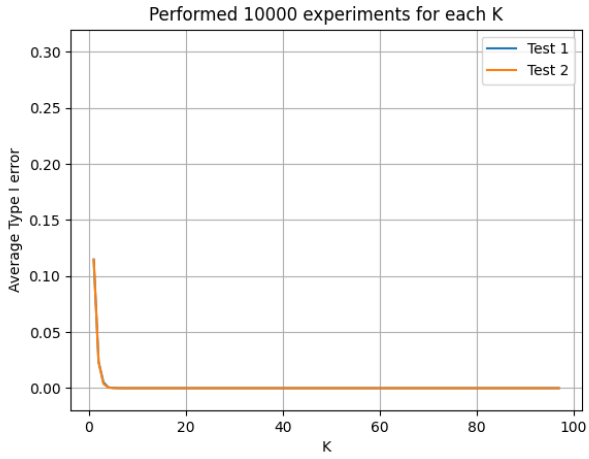


(a) Average value of Type I error observed

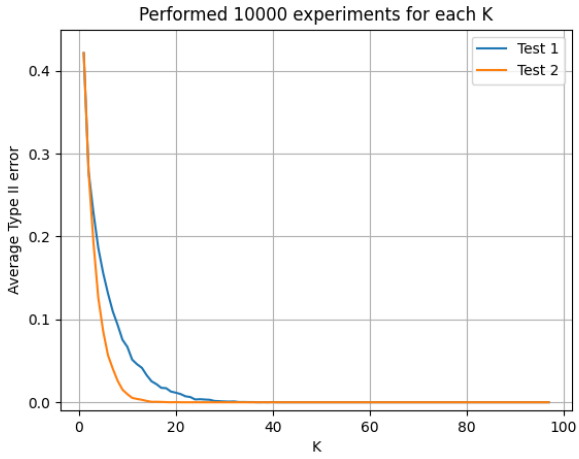


(b) Average value of Type II error observed

Figure 4: Number of iterations for each side of hypothesis as truth, $N = 1000$



(a) Average value of Type I error observed



(b) Average value of Type II error observed

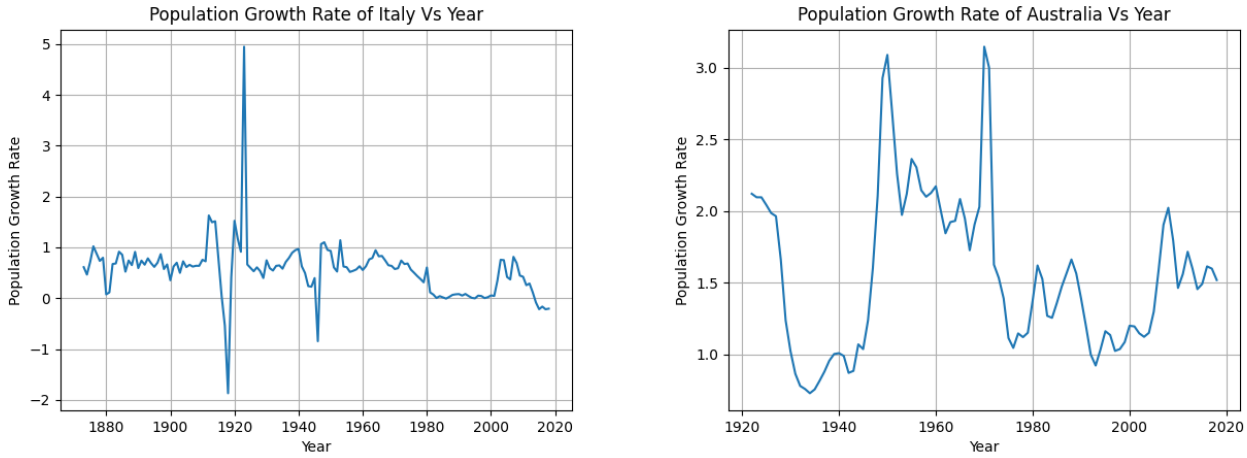
Figure 5: Number of iterations for each side of hypothesis as truth, $N = 10000$

3.5 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 Analysis of Population Growth Rate



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

Figure 1: Analysis of Population Growth Rate for Italy and Australia

From the above obtained plot, we can approximate population growth rate of Italy as a constant μ_1 with zero-mean additive Gaussian noise, i.e. population growth rate of Italy $\sim \mathcal{N}(\mu_1, \sigma_1^2)$.
Similarly, population growth rate of Australia $\sim \mathcal{N}(\mu_2, \sigma_2^2)$.

4.2 Problem Description

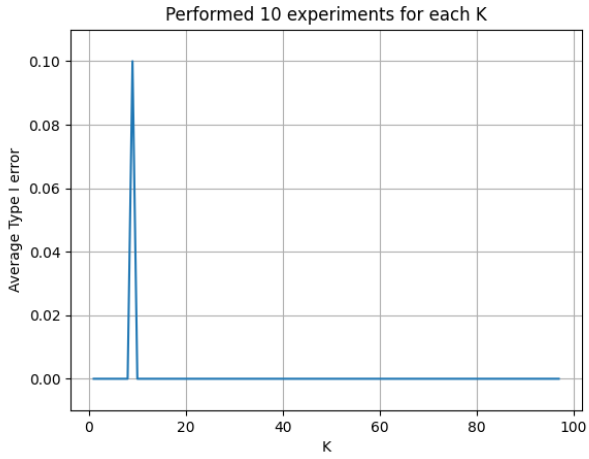
We assume that the population growth rate of Italy follows $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution and that of Australia follows $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution.
We are given a sample of population growth rate $\mathbf{g} = (g_1, \ldots, g_K)$ where K is the sample size which belongs to one of the above two mentioned distributions.
We devise a simple hypothesis:
H0: Country is Italy (i.e. $\mu = \mu_1$)
H1: Country is Australia (i.e. $\mu = \mu_2$)
We perform following test to accept/reject H0.
Test: Reject H0 if $\hat{\mu}_{MLE} = \text{mean}(\mathbf{g}) > 1.2$
We perform this experiment for $2N$ iterations for a fixed K, choosing each distribution as the true distribution for N iterations.
We analyze the observed values of Type I and Type II errors for the test for different K and N.

4.3 Generation of Sample

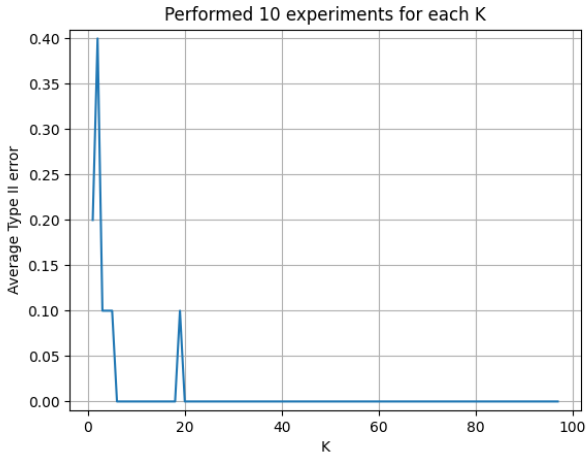
From `population_Country.csv`, we determine year-wise population growth rate for Italy from 1872 to 2018 (list of size 147) and for Australia from 1921 to 2018 (list of size 98).
The population growth rates are computed as percentage change in population (using `Total2` column) for every year.
Our K-sized sample is a randomly chosen K-sized subset ($K \leq 98$) of one of the above obtained list.
For first N iterations, we choose from list of Italy. For latter N iterations, we choose from list of Australia.

We use function `numpy.random.choice` with suitable arguments to obtain this sample.

4.4 Results

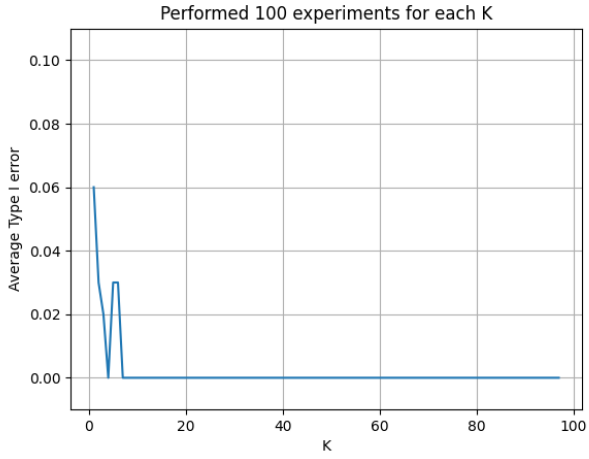


(a) Average value of Type I error observed

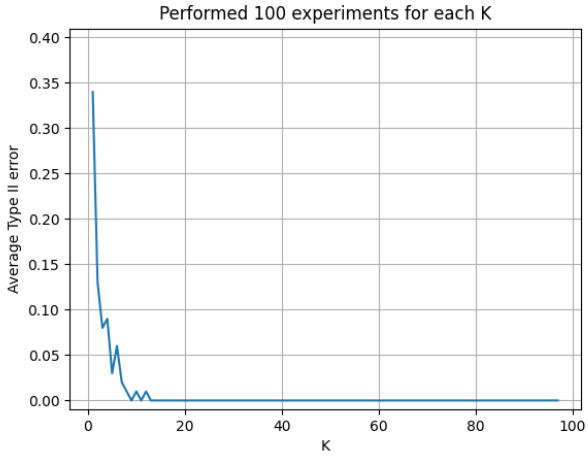


(b) Average value of Type II error observed

Figure 2: Number of iterations for each side of hypothesis as truth, $N = 10$

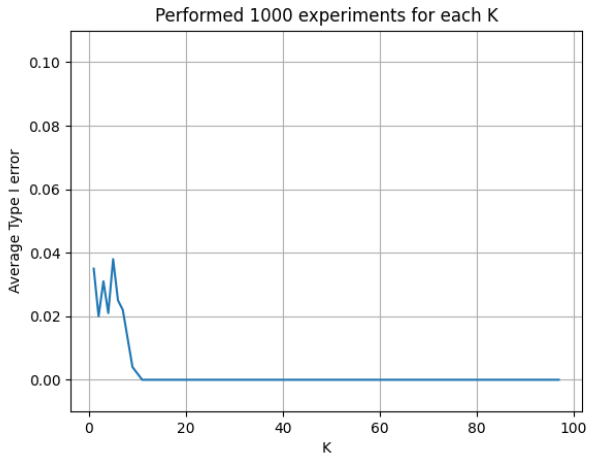


(a) Average value of Type I error observed

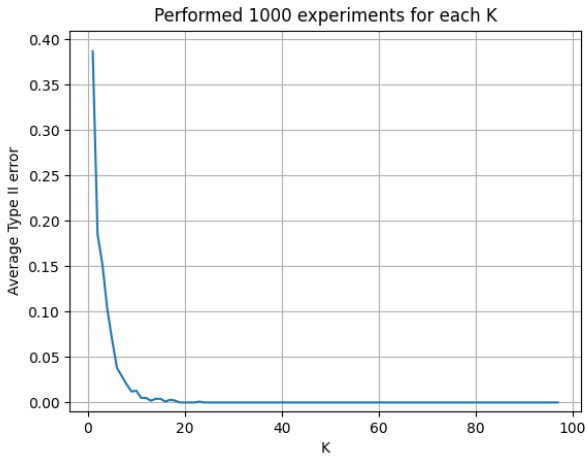


(b) Average value of Type II error observed

Figure 3: Number of iterations for each side of hypothesis as truth, $N = 100$

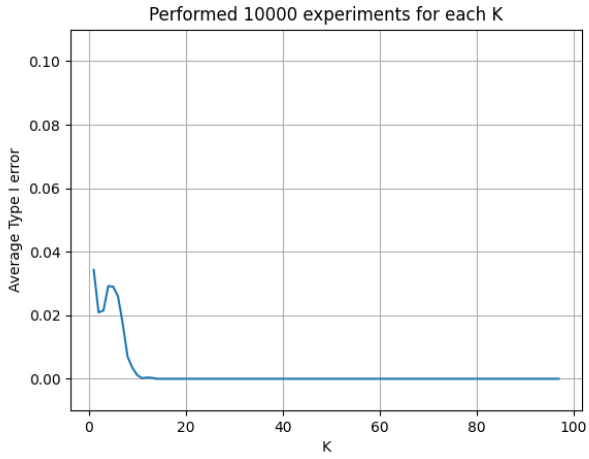


(a) Average value of Type I error observed

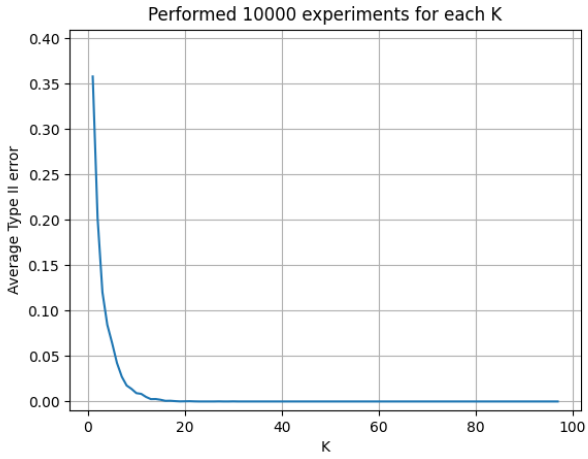


(b) Average value of Type II error observed

Figure 4: Number of iterations for each side of hypothesis as truth, $N = 1000$



(a) Average value of Type I error observed



(b) Average value of Type II error observed

Figure 5: Number of iterations for each side of hypothesis as truth, $N = 10000$

4.5 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.