SI424: Statistical Inference
Project Report

Krushnakant Bhattad	Devansh Jain
190100036	190100044

Autumn 2021

Project Title:

Data set used:

Associated GitHub Repository:

The GitHub repository can accessed at:

ABSTRACT:

COMMON NOTATIONS:

$\mathcal{N}(\mu, \sigma^2)$ denotes the Normal distribution with mean μ and variance σ^2.

1 Parameter estimation Problem 1

1.1 Description

The population growth rate of Italy follows $\mathcal{N}(\mu, \sigma^2)$ distribution.
We have population growth rate $g = (g_1, ..., g_n)$ of different years given to us for Italy.
Estimated values are $\hat{\mu}_{MLE} = \text{mean}(g)$ and $\hat{\sigma^2}_{MLE} = \text{mean}((g - \hat{\mu}_{MLE})^2)$.

1.2 Experiment

From `population_Country.csv`, we extract one unordered lists of 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 unordered list is 147 for Italy.
We randomly choose a K sized subset ($K \leq 147$) of the list. We compute the estimate of μ and σ^2 ($\hat{\mu}_{MLE}$ and $\hat{\sigma^2}_{MLE}$). We repeat this for N iterations.
We intend to observe the variation in our estimate of μ and σ^2 for different K and N.

1.3 Results

1.4 Inference

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 $a = (a_1, ..., a_n)$ of people who died in Greece in 2005 given to us.
Estimated value of p is $\hat{p}_{MLE} = \text{mean}(a)$.

2.2 Experiment

From `mortality_Country.csv`, we extract percentage of deaths in Greece in 2005 for each age interval.
We generate 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 to return.
Using the above defined procedure, we generate a K sized list of ages. We 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

2.4 Inference

3 Hypothesis Testing Problem 1

3.1 Description

We have two countries - Italy and Australia.
We have gender ratios $r = (r_1, ..., r_n)$ of different years given to us for a specific country.
We observe that $r_i \sim \text{Uniform}[\theta - \beta, \theta + \beta]$.
H0: Country is Italy
H1: Country is Australia
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` columm) 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

3.4 Inference

4 Hypothesis Testing Problem 2

4.1 Description

We have two countries - Italy and Australia.
We have population growth rate $g = (g_1, ..., g_n)$ of different years given to us for a specific country.
We observe that $g_i \sim \mathcal{N}(\mu, \sigma^2)$
H0: Country is Italy
H1: Country is Australia
Test: Reject H0 if $\hat{\mu}_{MLE} = \text{mean}(g) < 1.2$

4.2 Experiment

From `population_Country.csv`, we extract one unordered lists of 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 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 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 the test for different K and N.

4.3 Results

4.4 Inference