AE102 : Data Analysis and Interpretation

Data Analysis of Solar Eclipse Database

Soham S. Phanse[1]

Guide:
Prof. Prabhu Ramchandran[1] Prof. Amuthan Ramabathiran[1]

[1]Department of Aerospace Engineering, IIT Bombay

Objectives

Objective

To analyse the Solar Eclipse Database dating from 1900 AD to 2100 AD.

* Identify the Random Variables, find population parameters and computer sampling distributions.
* Find confidence intervals for the Sample Parameters
* Formulate a Hypothesis and test it
* Visualize the data effectively
* Perform Regression Analysis on the variables

Outline

Outline

About The Topic

We have chosen the topic analysing data related to solar eclipses in the last 200 years. The primary reference we will be using for this is *NASA's 5 Millennium Solar Eclipse Catalog - NASA/TP2009-214174*. There were and will be about 11898 eclipses in the 5 millennia from 2000 BC to 3000 AD.Solar eclipses are a topic of interest among astronomers, astrologers and the general public as well. Through this project we will try and analyse various parameters related to it with the help of the database.

What is a Solar Eclipse?

A solar eclipse occurs when a portion of the Earth is engulfed in a shadow cast by the Moon which fully or partially blocks sunlight. This occurs when the Sun, Moon and Earth are aligned in a straight line. Such alignment coincides with a new moon (*syzygy*[1]) indicating the Moon is closest to the *ecliptic plane*[2]. In a total eclipse, the disk of the Sun is fully obscured by the Moon. In partial and annular eclipses, only part of the Sun is obscured.

[1] *syzygy* is a word used by astronomers to denote an alignment of 3 bodies in a straight line

[2] The *ecliptic plane* is the plane of Earth's elliptical orbit around the Sun

Data

Sr	Yr	Month	D	Days_b/w_ecl	TD_Gr_Ecl	DELTA T	L_N	S_N	Type	QLE	Gamma	Ecl_Mag	Lat	Long	Alt	Width	Cent_Dur (min)	Cent_Dur (sec)	Tot_Dur (sec)
9283	1901	May	18	177.00	05:33:48	-1	-1220	136	T	n-	-0.3626	1.068	2S	98E	69	238	6	29	389
9284	1901	Nov	11	148.00	07:28:21	0	-1214	141	A	p-	0.4758	0.9216	11N	69E	62	336	11	1	661
9285	1902	Apr	8	29.00	14:05:06	0	-1209	108	Pe	-t	1.5024	0.0643	72N	142W	0	0	0	0	0
9286	1902	May	7	177.00	22:34:16	0	-1208	146	P	t-	-1.0837	0.8593	70S	125W	0	0	0	0	0
9287	1902	Oct	31	149.00	08:00:18	1	-1202	151	P	t-	1.1556	0.696	71N	101E	0	0	0	0	0
9288	1903	Mar	29	176.00	01:35:23	2	-1197	118	A	-p	0.8413	0.9767	56N	130E	32	153	1	53	113
9289	1903	Sep	21	178.00	04:39:52	2	-1191	123	T	-p	-0.8967	1.0316	58S	77E	26	241	2	12	132
9290	1904	Mar	17	176.00	05:40:44	3	-1185	128	A	nn	0.1299	0.9367	6N	95E	82	237	8	7	487
9291	1904	Sep	9	178.00	20:44:21	3	-1179	133	T	-n	-0.1625	1.0709	4S	135W	81	234	6	20	380
9292	1905	Mar	6	177.00	05:12:26	4	-1173	138	A	p-	-0.5768	0.9269	40S	117E	55	334	7	58	478
9293	1905	Aug	30	177.00	13:07:26	5	-1167	143	T	p-	0.5708	1.0477	42N	4W	55	192	3	46	226

Figure: The 5 Millennium Solar Eclipse Catalog - NASA/TP2009-214174 - First 10 rows

Outline

RV1: Total Central Duration

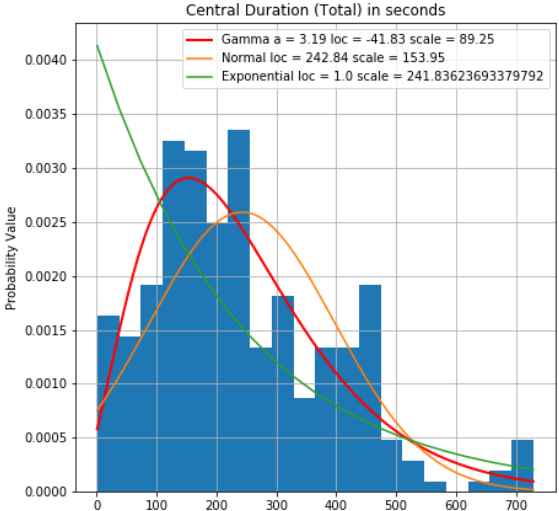
- For central eclipses (total, annular, or hybrid), the central line duration of the total or annular phase (in minutes and seconds) is given at the geographic position intersected by the axis of the lunar shadow cone at the instant of greatest eclipse.
- In the case of a total or hybrid eclipse, this duration is very nearly, the maximum duration of the total phase along the entire umbral path.
- For an annular eclipse, the duration at the greatest eclipse may be near either the minimum or maximum duration of the annular phase along the path.

True Population Parameters

The True Population Parameters (Total Central Duration of Eclipses in seconds) are:

- Population Mean : 242.83 seconds
- Population Variance : 23784.91 seconds
- Population Standard Deviation : 154.22 seconds

Fitting Distributions



Sampling Distribution of Sample Mean

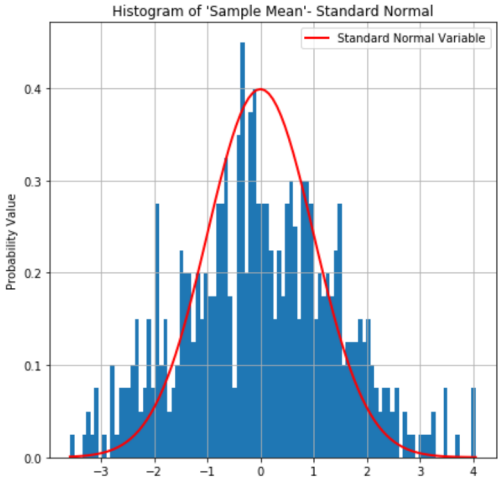
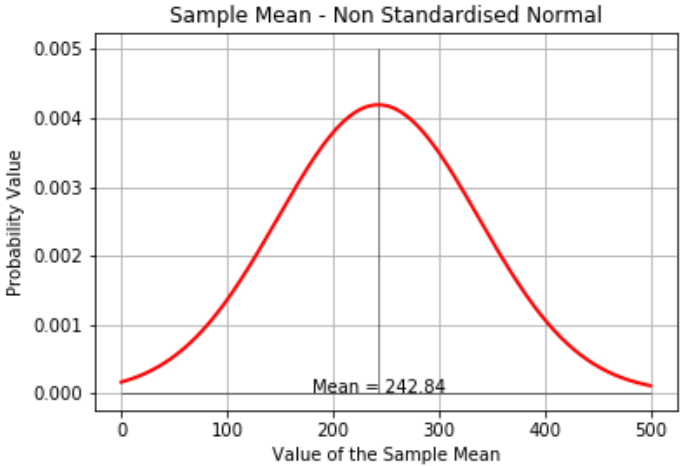
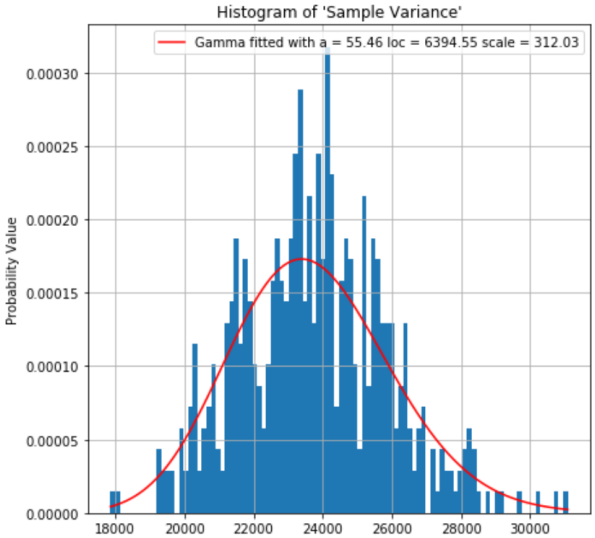


Figure: Standard Normal Approximation to the Sample Mean

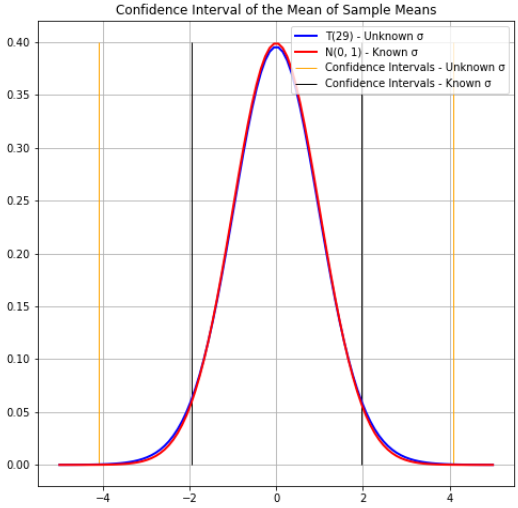
Sampling Distribution of Sample Mean



Sampling Distribution of Sample Variance



Confidence Interval of the Mean of Sample Means



Confidence Interval of the Variance of Sample Means

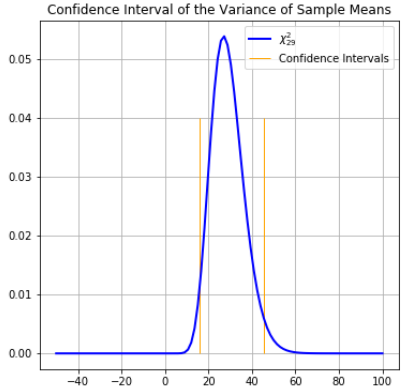


Figure: Confidence Intervals for the Variance of Sample Mean

Confidence Intervals for the Variance

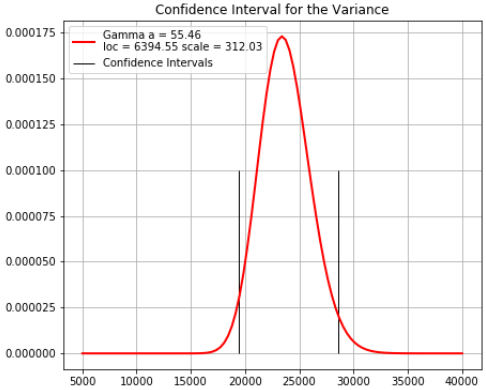


Figure: Confidence Intervals for the Variance of the Original Distributions

RV2: Eclipse Latitude

- Only eclipses with non-zero central duration are considered in the following analysis.
- The geographic latitude and longitude corresponds to the position of greatest eclipse.
- Negative values correspond to the Southern Hemisphere and Positive Values to the Northern Hemisphere.
- '0' corresponds to the Equator.

True Population Parameters

The True Population Parameters (Eclipse Latitude) are:

- Population Mean : -0.92 = 0.92 S
- Population Variance : 1540.29
- Population Standard Deviation : 39.24

Fitting Distributions

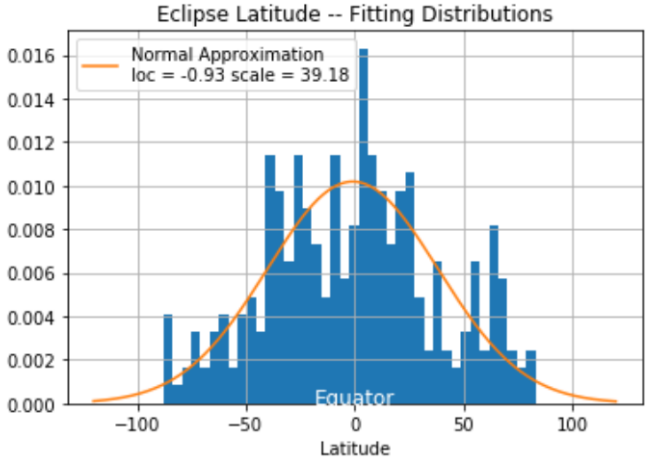


Figure: Distribution Fitting

Sampling Distribution of Sample Mean

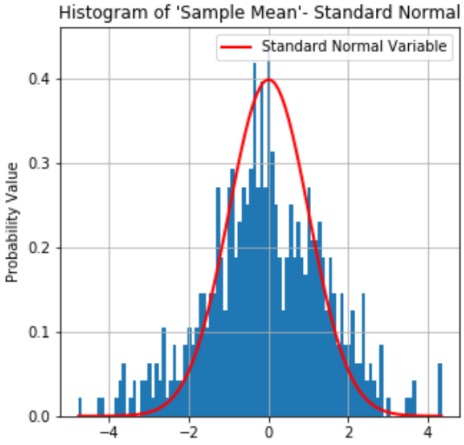


Figure: Caption

Sampling Distribution of Sample Variance

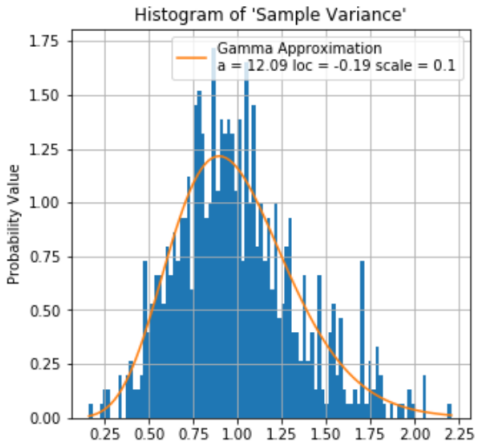
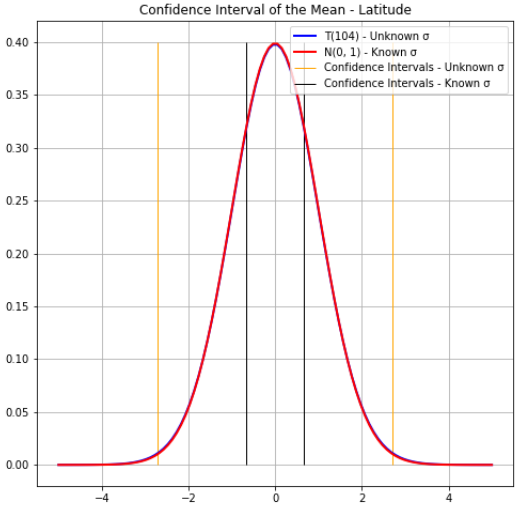
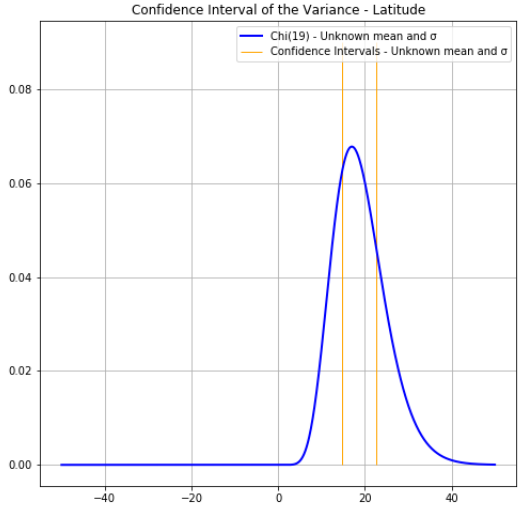


Figure: Caption

Confidence Interval of the True Mean



Confidence Interval of the True Variance



Outline

Formulating the Hypothesis

We have formulated the following hypothesis for the *RV2: Eclipse Magnitude*

- **Null Hypothesis**
  We formulate the hypothesis about the mean of this population, it is as follows:

$$H_0 : \mu(\textit{Eclipse Latitude}) = \mu_0 \;\; \textit{such that} \;\; \mu_0 = 0$$

  This hypothesis says that the mean latitude of the eclipses is zero $\longrightarrow$ that is the eclipse latitudes are expected to be near to the equator.

Formulating the Hypothesis

We have formulated the following hypothesis for the *RV2: Eclipse Magnitude*

- **Null Hypothesis**
  We formulate the hypothesis about the mean of this population, it is as follows:
  $$H_0 : \mu(\textit{Eclipse Latitude}) = \mu_0 \ \textit{ such that } \ \mu_0 = 0$$
  This hypothesis says that the mean latitude of the eclipses is zero $\longrightarrow$ that is the eclipse latitudes are expected to be near to the equator.
- **Alternate Hypothesis**
  The alternate hypothesis is as follows :
  $$H_0 : \mu(\textit{Eclipse Latitude}) \neq \mu_0 \ \textit{ such that } \ \mu_0 = 0$$

Hypothesis Tests - Variance Known

Thus, the significance level α test **is reject** H_0 **if**

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\frac{\alpha}{2}}$$

and **accept** H_0 **if**

$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\frac{\alpha}{2}}$$

Results

```
The null hypothesis is accepted at Significance Level = 0.496 Sample Size = 250

* Monitoring Data -- (for the developer) Test Statistic = -1.328 Zalpha = -0.681

The P-value of the test is :  0.18422190921522041
Since the p-value is smaller than the significance level the Hypothesis is **rejected**
```

Figure: Hypothesis Testing Results

Hypothesis Tests - Variance Unknown

Thus, the significance level α test is **reject** H_0 **if**

$$\textit{accept } H_0 \textit{ if } \left|\frac{\sqrt{n}(X-\mu_0)}{S}\right| \leq t_{\alpha/2,n-1}$$

and **accept** H_0 **if**

$$\textit{reject } H_0 \textit{ if } \left|\frac{\sqrt{n}(X-\mu_0)}{S}\right| > t_{\alpha/2,n-1}$$

Outline

Best Regression Pair

As we had a lot of arrays of data and selecting any two of them for conducting the *Regression Analysis* was difficult we constructed a function which checked the *Determination of Co-relation (R^2)* for all possible data pairs. After the analysis we took the best where the R^2 value was the greatest.

Best Regression Pair

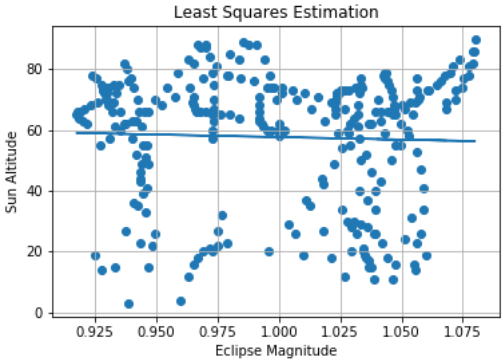


Figure: Eclipse Magnitude vs Sun Altitude - Least Squares Estimation

Distribution of Regression Parameters

$$A \approx N(\alpha, \frac{\sigma^2 \sum_i x_i^2}{nS_{xx}})$$

$$B \approx N(\beta, \frac{\sigma^2}{S_{xx}})$$

Distribution of Regression Parameters - α

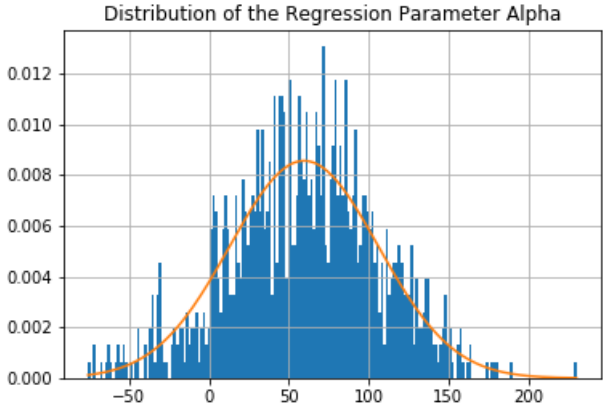


Figure: Distribution of Regression Parameter α

Distribution of Regression Parameters - β

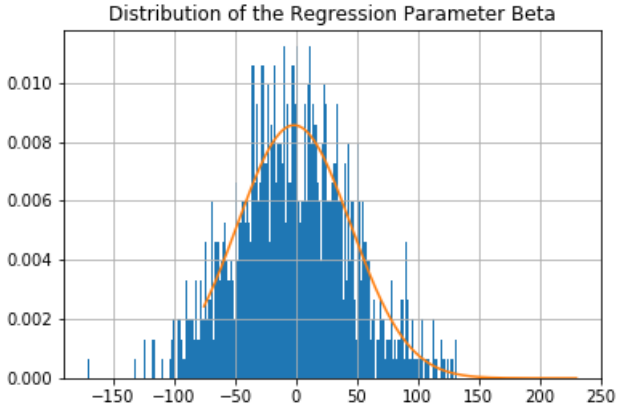


Figure: Distribution of Regression Parameter β

Beautiful Patterns . . .

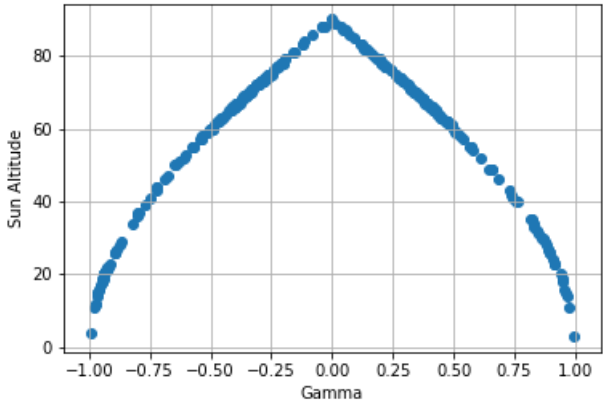


Figure: Gamma vs Sun Altitude

Beautiful Patterns . . .

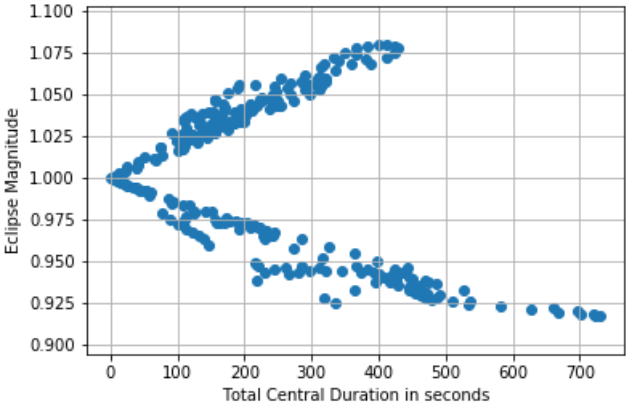


Figure: Total Central Duration vs Eclipse Magnitude