Statistical Modeling - 24DS636 (2024-25)

Abhijith M S

2025-01-25

Table of contents

Course Introduction

The course contents of “Statistical Modeling” offered by Abhijith M S, PhD to Masters students pursuing M.Tech in Data Science, during the even semester of the academic year 2024-25.

Syllabus

(As given in the curriculum)

- Probability, Random Variables & Probability Distributions.
- Sampling, analysis of sample data-Empirical Distributions, Sampling from a Population Estimation, confidence intervals, point estimation–Maximum Likelihood, Probability mass functions, Modeling distributions, Hypothesis testing- Z, t, Chi-Square.
- ANOVA & Designs of Experiments - Single, Two factor ANOVA, Factorials ANOVA models.
- Linear least squares, Correlation & Regression Models-linear regression methods, Ridge regression, LASSO, univariate and Multivariate Linear Regression, probabilistic interpretation, Regularization, Logistic regression, locally weighted regression.
- Exploratory data analysis, Time series analysis, Analytical methods – ARIMA and SARIMA.

Evaluations: A Tentative Timeline

- Best two marks out of three quizzes (Total = 20 marks)
- Quiz-1 (10 marks): (January First week)
- Quiz-2 (10 marks):(March First week)
- Quiz-3 (10 marks):(April First week)
- Assignments (Total = 30 marks)
- Assignment-1 (10 marks):(Submission: End of January)
- Assignment-2 (10 marks):(Submission: End of March)

Course Introduction

- Project Review - 1 (10 marks):(February second week)
- Mid Sem (Total = 20 marks)
- Mid-Semester Exam (20 marks):(Feb first week, as per Academic calender)
- End Sem (Total = 30 marks)
- End-Semester Project Presentation (20 marks):(April second week, as per Academic calender) \end{itemize}$

Contact: ms_abhijith@cb.amrita.edu

1 Parameter Estimation

1.1 Point Estimation: Maximum Likelihood Estimators

1.2 Interval Estimates

- Consider a sample $X_1, X_2, \ldots, X_n$ drawn from a known distribution with an unknown mean μ.
- It is established that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ serves as the maximum likelihood estimator for μ.
- However, the sample mean $\bar{X}$ is not expected to be exactly equal to μ, but rather close to it.
- Therefore, instead of providing a single point estimate, it is often more useful to specify an interval within which we are confident that μ lies.
- To determine such an interval estimator, we utilize the probability distribution of the point estimator.

1.2.1 Confidence Intervals for the Mean of a normal population with known Variance

- Consider a sample $X_1, X_2, \ldots, X_n$ drawn from a normal distribution with an unknown mean μ and a known variance σ^2.
- The point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n.
- Therefore, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution.

> ℹ What to do
>
> Consider that I want to find an interval around $\bar{X}$ such that the actual population mean μ falls within the interval, say 95 % of the times.

Tip

- For finding such an interval, I can use the Z-table. From the Z-table I can find:

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.9750 - 0.0250 = 0.95$$

- Rewriting the above equation:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$
$$P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu-\bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$
$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu-\bar{X} < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$
$$P\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- We have P(Z < -1.96) = 0.025, similarly P(Z > 1.96) = 0.025. Usually 1.96 is represented generally as $z_{0.025}$. Thus, P(Z < -z$_{0.025}$) = 0.025 and P(Z > z$_{0.025}$) = 0.025.

- Hence, 100(1-0.05) percent confidence interval for the mean of a normal population with known variance is:

$$P\left(\bar{X}-z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{0.025}\frac{\sigma}{\sqrt{n}}\right) = 0.95$$
$$P\left(\bar{X}-z_{0.05/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{0.05/2}\frac{\sigma}{\sqrt{n}}\right) = (1-0.05)$$

- For a confidence level of $100(1-\alpha)$ percent, the corresponding critical value from the standard normal distribution is $z_{\alpha/2}$.
- The $100(1-\alpha)$ percent confidence interval for μ is given by:

$$\mu \in \left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{1.1}$$

- The interval as given in Equation 1.1 is called a two-sided confidence interval.

- Also the term $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called the margin of error.

ℹ Derivation of two-sided confidence interval

- To find 100(1-α) percent confidence interval of mean (μ), we have;

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}/2\right) = 1-\alpha$$

- Doing the same manipulations we did earlier for obtaining the 95percent confidence interval we can obtain:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- The above equation give us the required confidence interval, as given in Equation 1.1.

What if !?

What if we are interested in one sided confidence intervals !!?

One-sided Upper Confidence Iterval

- To determine such an interval, for a standard normal random variable Z, we have;

$$P\left(Z < 1.645\right) = 0.95$$

- Thus,

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.645\right) = 0.95$$

$$P\left(\mu - \bar{X} > -1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\mu > \bar{X} - 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- Thus a 95 percent one-sided upper confidence interval for μ is

$$\mu \in \left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$$

or in other words; 100(1-0.05) percent one-sided upper confidence interval for μ is

$$\mu \in \left(\bar{X} - z_{0.05}\frac{\sigma}{\sqrt{n}}, \infty\right)$$

⚠ Oneside interval!

Can you think of another one sided confidence interval?

ℹ One-sided lower confidence interval

- We have
$$P\left(Z > -1.645\right) = 0.95$$
- Proceed just like in the previous case and you will find a 100(1-0.05) percent one-sided lower confidence interval for μ as;

$$\mu \in \left(-\infty, \bar{X} + z_{0.05}\frac{\sigma}{\sqrt{n}}\right)$$

- In general, 100(1-α) percent one-sided upper confidence interval for μ is given in Equation 1.2.

$$\mu \in \left(\bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right) \tag{1.2}$$

- Also, 100(1-α)percent one-sided lower confidence interval for μ is given in Equation 1.3.
$$\mu \in \left(-\infty, \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) \tag{1.3}$$

- The python code below creates a sample and find 95% confidence interval for the mean if the population standard deviation is assumed to be 10. Other values are specified in the code.

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
mu = 50  # true mean
sigma = 10  # known standard deviation
n = 30  # sample size
```

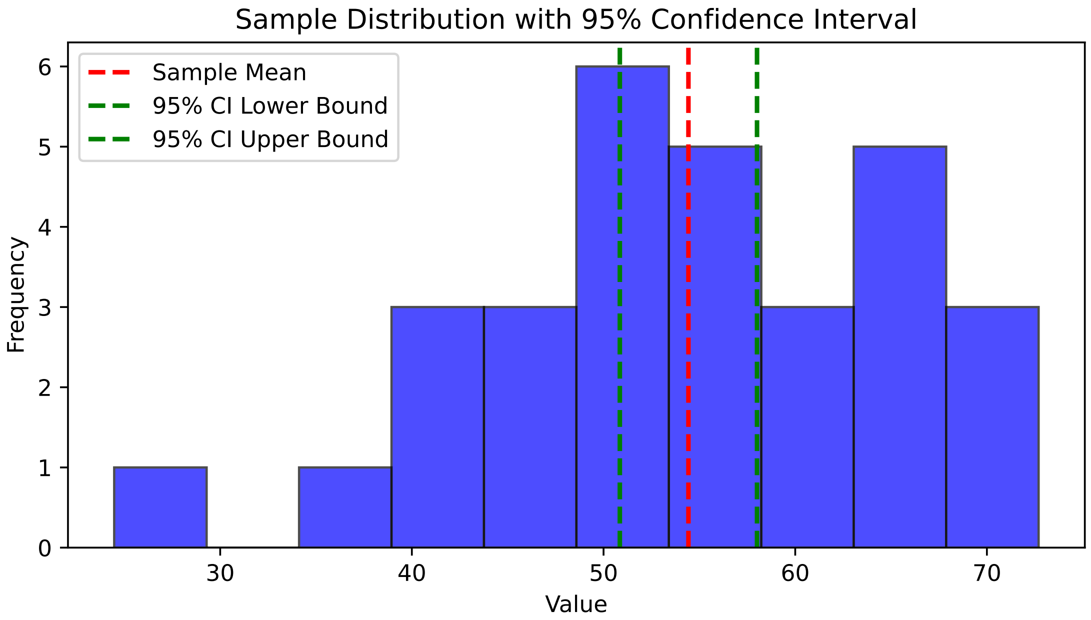
```
alpha = 0.05  # significance level

# Generate a sample
np.random.seed(0)
sample = np.random.normal(mu, sigma, n)
sample_mean = np.mean(sample)

# Calculate the confidence interval
z = 1.96  # z-value for 95% confidence
margin_of_error = z * (sigma / np.sqrt(n))
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

# Plot the sample and confidence interval
plt.figure(figsize=(8, 4))
plt.hist(sample, bins=10, alpha=0.7, color='blue', edgecolor='black')
plt.axvline(sample_mean, color='red', linestyle='dashed', linewidth=2, label='Sample Mean')
plt.axvline(confidence_interval[0], color='green', linestyle='dashed', linewidth=2, label='95% CI Lower Bound')
plt.axvline(confidence_interval[1], color='green', linestyle='dashed', linewidth=2, label='95% CI Upper Bound')
plt.title('Sample Distribution with 95% Confidence Interval')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.show()

print(f”Sample Mean: {sample_mean}”)
print(f”95% Confidence Interval: {confidence_interval}”)
```



```
Sample Mean: 54.42856447263174
95% Confidence Interval: (50.85011043026466, 58.007018514998826)
```

Problem

Suppose that when a signal having value μ is transmitted from location A the value received at location B is normally distributed with mean μ and variance 4. That is, if μ is sent, then the value received is μ + N where N, representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent 9 times. If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5; (a). construct a 95 percent two-sided confidence interval for μ. (b). construct a 95 percent one-sided upper and lower confidence intervals for μ.

1.2.2 Confidence Intervals for the Mean of a normal population with unknown Variance

- If you recollect the discussion we had about the sample mean from a normal population with unknown variance we saw that variable t$_{n-1}$ given by:

$$t_{n-1} = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

has a t-distribution with n-1 degrees of freedom.

- Because of the symmetry of the t-distribution we can write for any $\alpha \in (0,\ 1/2)$;

$$P\left(-t_{\alpha/2,n-1} < \sqrt{n}\frac{\bar{X}-\mu}{S} < t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$P\left(-\bar{X} - t_{\alpha/2,n-1}\frac{\sqrt{n}}{S} < -\mu < -\bar{X} + t_{\alpha/2,n-1}\frac{\sqrt{n}}{S}\right) = 1-\alpha$$

$$P\left(\bar{X} + t_{\alpha/2,n-1}\frac{\sqrt{n}}{S} > \mu > \bar{X} - t_{\alpha/2,n-1}\frac{\sqrt{n}}{S}\right) = 1-\alpha$$

$$P\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

- If the sample mean is $\bar{X}$ and sample standard deviation S, then we can say that with 100(1-α) percent confidence that

$$\mu \in \left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

- In this case 100(1-α) percent one-sided upper confidence interval can be obtained from the fact that:

$$P\left(\sqrt{n}\frac{(\bar{X}-\mu)}{S} < t_{\alpha,n-1}\right) = 1-\alpha$$

$$P\left(\mu > \bar{X} - \frac{S}{\sqrt{n}}t_{\alpha,n-1}\right) = 1-\alpha$$

- Thus $100(1-\alpha)$ percent one-sided upper confidence interval for the mean in this case is given by;

$$\mu \in \left(\bar{X} - \frac{S}{\sqrt{n}}t_{\alpha,n-1}, \infty\right)$$

- Thus $100(1-\alpha)$ percent one-sided lower confidence interval for the mean in this case is given by;

$$\mu \in \left(-\infty, \bar{X} + \frac{S}{\sqrt{n}}t_{\alpha,n-1}\right)$$

Problem

Let us again consider the previous problem but let us now suppose that when the value μ is transmitted at location A then the value received at location B is normal with mean μ and variance σ^2 but with σ^2 being unknown. If 9 successive values are, 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, and 10.5, compute a 95 percent confidence interval for μ.

1.2.3 Confidence Intervals for the Variance of a Normal Distribution

- If we are sampling from a normal distribution with unknown mean and unknown variance then;

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

follows a chi-squared distribution.

- We have

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

- Hence, 100(1-α) percent two-sided confidence interval for the variance in this case;

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

- The 100(1-α) percent one-sided upper and lower confidence intervals in this case will be respectively;

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty\right)$$

and

$$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$$

Problem

A standardized procedure is expected to produce washers with very small deviation in their thicknesses. Suppose that 10 such washers were chosen and measured. If the thicknesses of these washers were, in inches; .123, .133, .124, .125, .126, .128, .120, .124, .130, and .126. What is a 90 percent confidence interval for the standard deviation of the thickness of a washer produced by this procedure?

All problems and most part of text are taken from Ross (2009) .

2 Hypothesis Testing

2.0.1 Introduction

- A statistical hypothesis is typically a statement regarding a set of parameters of a population distribution.
- It is termed a hypothesis because its truth value is unknown.
- The main challenge is to devise a method to determine whether the values of a random sample from this population align with the hypothesis.
- For example, consider a normally distributed population with an unknown mean value and a known variance of 1. The statement “ is less than 0.5” is a statistical hypothesis that we can test by observing a random sample from this population. If the random sample is consistent with the hypothesis, we say the hypothesis is “accepted”; otherwise, it is “rejected.”

References

Ross, Sheldon. 2009. “Probability and Statistics for Engineers and Scientists.” Elsevier, New Delhi 16: 32–33.