

Statistical Modeling - 24DS636 (2024-25)

Abhijith M S

2025-01-27

Table of contents

Course Introduction	1
Syllabus	1
Evaluations: A Tentative Timeline	1
1 Parameter Estimation	3
1.1 Point Estimation: Maximum Likelihood Estimators	3
1.2 Interval Estimates	3
1.2.1 Confidence Intervals for the Mean of a normal population with known Variance	3
1.2.2 Confidence Intervals for the Mean of a normal population with unknown Variance	9
1.2.3 Confidence Intervals for the Variance of a Normal Distribution . .	10
2 Hypothesis Testing	13
2.0.1 Introduction and Important Terminology	15
2.0.2 Testing a Null Hypothesis	16
2.0.3 Significance Level and Classical Approach	17
2.0.4 Testing Hypotheses with Point Estimators	17
2.0.5 Hypothesis Tests Concerning the mean of a normal population . .	18
2.0.6 Choosing the Significance Level	20
References	21

Course Introduction

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

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

- 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.

Derivation of two-sided confidence interval

- To find $100(1-\alpha)$ 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 Interval

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

$$P(Z < 1.645) = 0.95$$

- 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

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

1 Parameter Estimation

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?

📖 One-sided lower confidence interval

- We have

$$P(Z > -1.645) = 0.95$$

- 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.

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

- Also, 100(1- α)percent one-sided lower confidence interval for μ is given in Equation 1.3.

$$\mu \in \left(-\infty, \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right) \quad (1.3)$$

- 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)

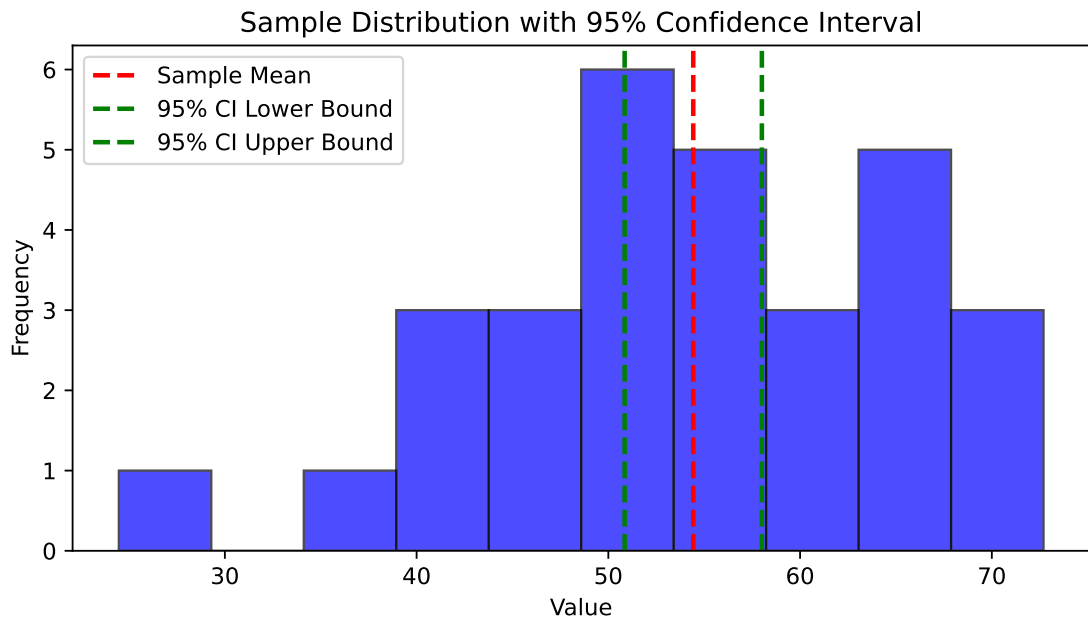
# Calculate the confidence interval
z = 1.96 # z-value for 95% confidence
margin_of_error = z * (sigma / np.sqrt(n))
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

# 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}")

```

1 Parameter Estimation



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 $\mu + N$ 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;

- construct a 95 percent two-sided confidence interval for μ .
- construct 95 percent one-sided upper and lower confidence intervals for μ .

🔥 Problem

Suppose a quality control manager at a factory wants to ensure that the average weight of a product is at least 500 grams. They take a random sample of 30 products and find the sample mean weight to be 495 grams with a standard deviation of 10 grams. Help the manager to estimate the minimum average weight of the products with 95% confidence.

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 $100(1-\alpha)$ 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 $100(1-\alpha)$ 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;

1 Parameter Estimation

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

- 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;

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

follows a chi-squared distribution.

- We have

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

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

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

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

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

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

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

and

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

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

Problem

A quality control manager at a factory wants to ensure that the average weight of a product is at least 500 grams. They take a random sample of 30 products and find the sample mean weight to be 495 grams with a standard deviation of 10 grams. The manager wants to test if the average weight of the products is significantly less than 500 grams at a 5% significance level.

- Null Hypothesis (H_0): The average weight of the products is at least 500 grams ($\mu \geq 500$).
- Alternative Hypothesis (H_1): The average weight of the products is less than 500 grams ($\mu < 500$).

Examples Highlighting the Need for Hypothesis Testing

1. Medical Research:

- Scenario: A pharmaceutical company develops a new drug intended to lower blood pressure.
- Hypothesis Testing: The null hypothesis (H_0) might state that the new drug has no effect on blood pressure, while the alternative hypothesis (H_1) states that the drug does lower blood pressure. Hypothesis testing helps determine if the observed effects in clinical trials are statistically significant or if they could have occurred by random chance.

2. Quality Control:

- Scenario: A factory produces light bulbs, and the quality control team wants to ensure that the average lifespan of the bulbs is 1000 hours.
- Hypothesis Testing: The null hypothesis (H_0) could be that the mean lifespan of the bulbs is 1000 hours. The alternative hypothesis (H_1) might be that the mean lifespan is not 1000 hours. Hypothesis testing helps the team decide whether to accept the production process or take corrective actions.

3. Marketing:

2 Hypothesis Testing

- Scenario: A company launches a new advertising campaign and wants to know if it has increased sales.
- Hypothesis Testing: The null hypothesis (H_0) might state that the advertising campaign has no effect on sales, while the alternative hypothesis (H_1) states that the campaign has increased sales. Hypothesis testing helps the company determine if the increase in sales is statistically significant.

4. Education:

- Scenario: An educator wants to test if a new teaching method is more effective than the traditional method.
- Hypothesis Testing: The null hypothesis (H_0) could be that there is no difference in effectiveness between the new and traditional methods. The alternative hypothesis (H_1) might be that the new method is more effective. Hypothesis testing helps in making data-driven decisions about adopting new teaching strategies.

5. Environmental Science:

- Scenario: Researchers want to determine if a new policy has reduced pollution levels in a city.
- Hypothesis Testing: The null hypothesis (H_0) might state that the policy has no effect on pollution levels, while the alternative hypothesis (H_1) states that the policy has reduced pollution levels. Hypothesis testing helps in evaluating the effectiveness of environmental policies.

These examples illustrate how hypothesis testing is a crucial tool in various fields for making informed decisions based on data.

Important Terminology

- Null Hypothesis (H_0): The hypothesis that there is no effect or no difference. It is the default assumption that any observed effect is due to random chance. It is the hypothesis that researchers aim to test against.
- Alternative Hypothesis (H_1 or H_a): The hypothesis that there is an effect or a difference. It is what researchers want to prove.
- Test Statistic: A standardized value that is calculated from sample data during a hypothesis test. It is used to decide whether to reject the null hypothesis.
- P-value: The probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true. A smaller p-value

indicates stronger evidence against the null hypothesis.

- Significance Level (α): A threshold set by the researcher which the p-value must be below in order to reject the null hypothesis. Common significance levels are 0.05, 0.01, and 0.10.
- Critical Value: The value that the test statistic must exceed in order to reject the null hypothesis. It is determined based on the significance level and the distribution of the test statistic.
- Power of a Test: The probability that the test correctly rejects a false null hypothesis (i.e., it does not make a type II error). Higher power indicates a greater ability to detect an effect when there is one.
- Type I Error: The error made when the null hypothesis is true, but is incorrectly rejected. The probability of making a type I error is denoted by α .
- Type II Error: The error made when the null hypothesis is false, but is incorrectly accepted. The probability of making a type II error is denoted by β .
- Confidence Interval: A range of values derived from the sample data that is likely to contain the population parameter. It provides an estimate of the parameter with a certain level of confidence (e.g., 95%).
- One-tailed Test: A hypothesis test in which the region of rejection is on only one side of the sampling distribution. It tests for the possibility of the relationship in one direction.
- Two-tailed Test: A hypothesis test in which the region of rejection is on both sides of the sampling distribution. It tests for the possibility of the relationship in both directions.

2.0.1 Introduction and Important Terminology

- A statistical hypothesis is typically a statement regarding a set of parameters of a population distribution.
- It is termed a hypothesis because its true 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.
- Consider a population with distribution F_θ , where θ is unknown.
- We aim to test a specific hypothesis about θ .

2 Hypothesis Testing

- This hypothesis is denoted by H_0 and is referred to as the null hypothesis.
- For instance, if F_θ is a normal distribution function with mean θ and variance equal to 1, two possible null hypotheses about θ are:

$$H_0 : \theta = 1$$

$$H_0 : \theta > 1$$

$$H_0 : \theta \leq 1$$

- It is important to note that the null hypothesis in the first case fully specifies the population distribution.
- Whereas the null hypothesis in the second and third cases do not.

i Simple and Composite Hypotheses

- A hypothesis that fully specifies the population distribution when true is known as a simple hypothesis. (Eg; $H_0 : \theta = 1$)
- A hypothesis that does not fully specifies the population distribution is referred to as a composite hypothesis. (Eg; $H_0 : \theta > 1$, $H_{\neq\{0\}} : \theta \leq 1$)

2.0.2 Testing a Null Hypothesis

- To test a specific null hypothesis H_0 , we take a sample of size n from the population, say X_1, X_2, \dots, X_n .
- Based on these n values, we decide whether to accept or reject H_0 .
- We define a region C in the n -dimensional space. This region is called the critical region.
- If the sample X_1, X_2, \dots, X_n falls within the critical region C , we reject H_0 . Otherwise, we accept H_0 .
- In simple terms, the critical region C helps us determine the outcome of the statistical test.

accepts H_0 if $(X_1, X_2, \dots, X_n) \notin C$

and

rejects H_0 if $(X_1, X_2, \dots, X_n) \in C$

i Types of Errors in Hypothesis Testing

- When developing a procedure for testing a given null hypothesis H_0 , it is crucial to recognize that two different types of errors can occur.
- A type I error occurs if the test incorrectly rejects H_0 when it is actually true.
- A type II error occurs if the test incorrectly accepts H_0 when it is actually false.

! Note

The goal of a statistical test for H_0 is not to definitively determine its truth but to assess if the data is consistent with H_0 .

2.0.3 Significance Level and Classical Approach

- H_0 should be rejected only if the observed data is highly unlikely under H_0 .
- The classical method involves specifying a value α , known as the level of significance.
- The test is designed so that the probability of rejecting H_0 when it is true does not exceed α .
- Common choices for α are 0.1, 0.05, and 0.005.
- This approach ensures that the probability of a type I error (incorrectly rejecting H_0) is controlled and does not exceed the chosen α .

2.0.4 Testing Hypotheses with Point Estimators

- Suppose we are interested in testing a hypothesis concerning θ , an unknown parameter of the population.
- Specifically, for a given set of parameter values w , we aim to test:

2 Hypothesis Testing

$$H_0 : \theta \in w$$

- A common approach to developing a test of H_0 at a significance level α is to start by determining a point estimator of θ , say $d(X)$.
- The hypothesis H_0 is then rejected if $d(X)$ is “far away” from the region w .
- To determine how “far away” it needs to be to justify rejection of H_0 , we need to determine the probability distribution of $d(X)$ when H_0 is true.
- This distribution helps us determine the appropriate critical region to ensure the test has the required significance level α .

! Example

- For instance, consider testing the hypothesis that the mean of a normal distribution with parameters $(\theta, 1)$ is equal to 1.
- The test rejects the null hypothesis if the point estimate of θ (i.e., the sample mean) deviates more than $\frac{1.96}{\sqrt{n}}$ from 1.
- As we will discuss in the next section, the value $\frac{1.96}{\sqrt{n}}$ is selected to achieve a significance level of $\alpha = 0.05$.

2.0.5 Hypothesis Tests Concerning the mean of a normal population

2.0.5.1 With known Variance

- Let X_1, X_2, \dots, X_n be a sample of size n from a normal distribution with an unknown mean μ and a known variance σ^2 .
- We are interested in testing the null hypothesis:

$$H_0 : \mu = \mu_0$$

- Against the alternative hypothesis:

$$H_1 : \mu \neq \mu_0$$

- Where μ_0 is a specified constant.
- Since $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is a natural point estimator of μ , it is reasonable to accept H_0 if \bar{X} is not too far from μ_0 .

- Thus, the critical region of the test would be of the form:

$$C = \{X_1, \dots, X_n : |\bar{X} - \mu_0| > c\}$$

for some suitably chosen value c .

- To ensure that the test has a significance level α , we must determine the critical value c in the above equation such that the type I error is equal to α . This means c must satisfy:

$$P_{\mu_0}\{|\bar{X} - \mu_0| > c\} = \alpha$$

where P_{μ_0} denotes that the probability is computed under the assumption that population mean, $\mu = \mu_0$.

- When $\mu = \mu_0$, \bar{X} follows a normal distribution with mean μ_0 and variance $\frac{\sigma^2}{n}$. Therefore, the standardized variable Z defined by:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will have a standard normal distribution.

- The probability of a type I error is given by:

$$P(|\bar{X} - \mu_0| > c) = \alpha$$

- Equivalently, this can be written as:

$$2P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) = \alpha$$

- Where Z is a standard normal random variable. We know that:

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

- Therefore, we have:

$$\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2}$$

- Solving for c , we get:

2 Hypothesis Testing

$$c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

- Thus, the test at significance level α is to reject H_0 if:

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

- And accept H_0 otherwise. Equivalently, we can reject H_0 if:

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

- And accept H_0 if:

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

2.0.6 Choosing the Significance Level

- The appropriate significance level α depends on the specific context and consequences of the hypothesis test.
- If rejecting the null hypothesis H_0 would lead to significant costs or consequences, a more conservative significance level (e.g., 0.05 or 0.01) should be chosen.
- If there is a strong initial belief that H_0 is true, strict evidence is required to reject H_0 , implying a lower significance level.
- The test can be described as follows: For an observed value of the test statistic $\sqrt{n} \frac{|\bar{X} - \mu_0|}{\sigma}$, denoted as v , reject H_0 if the probability of the test statistic being as large as v under H_0 is less than or equal to α .
- This probability is known as the p-value of the test. H_0 is accepted if α is less than the p-value and rejected if α is greater than or equal to the p-value.
- In practice, the significance level is sometimes not set in advance. Instead, the p-value is calculated from the data, and decisions are made based on the p-value.
- If the p-value is much larger than any reasonable significance level, H_0 is accepted. Conversely, if the p-value is very small, H_0 is rejected.

References

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

