Introduction to Statistical Inference

Overview of Statistical Inference

Statistical inference is the process of extrapolating from sample data to make estimates about a larger population. This is commonly seen in election polling where a sample is used to estimate the support for a candidate with a specified margin of error.

Margin of Error

- Margin of error acknowledges that a sample estimate (e.0.562 or 56.2% in our example) is not the exact population value.
- It attempts to quantify the accuracy of the estimate.

The Role of Normal Distributions

- Classical statistics often assumes normally distributed populations.
- Normal distributions are not always precise in real-world data, like corporate revenues.
- Despite limitations, these assumptions are effective due to the Central Limit Theorem.

Central Limit Theorem and Its Importance

- The Central Limit Theorem (CLT) is pivotal for working with non-normal populations.
- It states that the distribution of sample means approximates a normal distribution.

Approximate Distribution of Standardized $\bar{X}$

The key formula presented in Section 9.8 is:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

where:

- Z has an approximately N(0, 1) distribution.
- σ^2 is the population variance.
- n is the sample size.
- This applies even for skewed data samples.

Application of Central Limit Theorem

With the Central Limit Theorem, we can see that a histogram of many sample means (e.g., mean revenues $\overline{R}$) will approximate a bell-shaped curve, even if the original data is skewed.

- Statistical Inference allows for extrapolation from samples to populations.
- The Central Limit Theorem justifies the use of normal distribution assumptions in practical scenarios.

Confidence Intervals for Means

Introduction to Confidence Intervals

- Recall that the sample mean as a random variable.
- Will develop the concept of margin of error as seen in election polls.
- Key questions: What is margin of error and how can we calculate it?

Basic Formulation

- Consider a random sample $W_1, ..., W_n$ from a population with mean μ and variance σ^2 (both unknown).
- The central 95% of the N(0,1) distribution is our focus.
- Using standard normal distribution to find cutoff points at -1.96 and 1.96.

Calculating the Confidence Interval

The confidence interval is given by:

$$P\left(-1.96 < Z < 1.96\right) = 0.95$$

$$P\left(-1.96 < \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95 \tag{1}$$

$$P\left(\overline{W} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95 \tag{2}$$

Adjusting for Unknown σ

- Substitute sample standard deviation s for σ in the formula.
- The adjusted confidence interval is:

$$0.95 \approx P\left(\overline{W} - 1.96\frac{s}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{s}{\sqrt{n}}\right) \tag{3}$$

95% Confidence Interval and Margin of Error

- Thus, the (approximate) 95% confidence interval for μ is:

$$(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}}) \tag{4}$$

- The term $1.96\frac{s}{\sqrt{n}}$ represents the margin of error.

- A 95% confidence interval provides a range within which we are fairly confident the true mean lies.
- The margin of error reflects the accuracy of our estimate.

Confidence Intervals from Approximately Normal Estimators

- Recall the estimation of parameters in a parametric family.
- Will focus on estimators ($\hat{\theta}$) that are approximately normally distributed.

Approximate 95% Confidence Interval for θ

- For an approximately normally distributed estimator $\hat{\theta}$, an approximate 95% confidence interval for θ is given by:

$$\hat{\theta} \pm 1.96 \times \text{s.e.}(\hat{\theta})$$

where $\text{s.e.}(\hat{\theta})$ is the standard error of $\hat{\theta}$.

Standard Errors and Confidence Intervals

- Forming confidence intervals in terms of standard errors is a common practice.
- This approach is applicable to many estimators that exhibit approximate normality.

Application to Maximum Likelihood Estimators

- Maximum Likelihood Estimators (MLEs) often are approximately normal (under mild conditions).
- This characteristic facilitates easy derivation of confidence intervals for MLEs.

- Confidence intervals from approximately normal estimators provide a method to estimate the range within which a parameter lies.
- The concept is widely applicable in statistical estimation and inferential statistics.

Example: Pima Diabetes Study

A famous data set involves Pima Indian women, with Y being 1 or 0, depending on whether the patient does ultimately develop diabetes, and the predictors being the number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin level, body mass index, diabetes pedigree function and age.

- will compare the (mean) Body Mass Index (BMI) values between diabetic and nondiabetic women.
- Population mean and variance denoted as μ_1, σ_1^2 for diabetics, and μ_0, σ_0^2 for nondiabetics.

Forming a Confidence Interval

Estimating the Difference

- Interest in estimating the difference $\theta = \mu_1 - \mu_0$.
- Sample estimate found to be 4.84 (BMI value difference), with a standard error of 0.56.
- Question: Is (mean) BMI substantially higher among diabetics in the population?

- Formulating a confidence interval: Our $\hat{\theta}$ was $\bar{U} - \bar{V} = 35.14 - 30.30 = 4.84$, with a standard error of 0.56, then

$$4.84 \pm 1.96 \times 0.56 = (3.74, 5.94)$$

- This interval provides an estimate range instead of a single point estimate.
- Margin of error calculated as $1.96 \times 0.56 = 1.10$.

Interpreting the Results

- The confidence interval suggests diabetics have a higher average BMI.
- The interval width is substantial but supports the conclusion of a significant difference.
- This approach acknowledges the use and possible limitations of sample-based estimates.

- The Pima Diabetes Study provides a practical example of how confidence intervals can be used to understand population parameters.
- The study indicates a substantial difference in BMI between diabetic and nondiabetic women.

Meaning of Confidence Intervals

- **Objective:** Estimating the mean weight (μ) of all adults in Auburn.
- **Method:** Sample 1000 people randomly and record their weights W_i.
- The true population mean μ is unknown; thus, we estimate it using the sample mean $\overline{W}$.
- Forming a confidence interval provides a measure of the accuracy of $\overline{W}$.

Interpreting Confidence Intervals

- Example 95% confidence interval: (142.6,158.8).
- Interpretation: We are a 95% confident that the mean weight μ is within this interval.

Understanding Through Repetition

- Imagine conducting the same survey repeatedly and recording each interval.
- Each sample would give a different interval with a different center and radius (i.e. margin of error).
- About 95% of these intervals will contain the true mean weight μ (alas, we don't know which ones).

Group Experiment Scenario

- Example: 100 people (you and 99 friends) each conduct the survey independently.
- Each person will get a different sample and thus a different confidence interval.
- Approximately 95 of these 100 intervals will contain the true population mean weight.

Practical Application

- In reality, only one sample of 1000 people is usually taken.
- The repeated sampling and notebook idea helps in understanding the meaning of a 95% confidence level.
- It emphasizes that our single interval has a 95% chance of containing the true μ.

- Confidence intervals provide an estimated range for population parameters.
- Understanding their meaning is crucial for correctly interpreting statistical results.

Confidence Intervals for Proportions

- We now explore how to find confidence intervals for proportions.
- **Example scenario:** estimating the proportion of people voting for a candidate in an election.
- Confidence intervals for proportions provide a method to estimate the range within which a population proportion lies.
- The approach simplifies computations, especially in cases where outcomes are binary.

Derivation: Sample Proportion

- Estimate the population proportion p using a sample proportion $\widehat{p}$ (p-hat).
- Assign value Y_i: 1 if a person votes for candidate A, 0 otherwise.
- Sample proportion $\widehat{p}$ is the mean of Y_i: $\widehat{p} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

Confidence Interval for Proportion

- Since we are working with means, we can use the formula for confidence interval of means.
- An approximate 95% confidence interval for p is $\widehat{p} \pm 1.96 \times s/\sqrt{n}$.
- Here, s^2 is the sample variance among the Y_i.

Simplifying the Confidence Interval

- Notice that each Y_i is either 1 or 0 (i.e., Bernoulli trial).
- Recall $Var(Y_i) = p(1-p)$, so the sample variance is $\widehat{p}(1-\widehat{p})$.
- The simplified confidence interval for p is then:

$$\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right)$$

Confidence Intervals for Machine Classification of Forest Covers

- *Remote sensing* is machine classification of type from variables observed aerially, typically by satellite. The application we'll consider here involves forest cover type for a given location; there are seven different types.
- Direct observation of the cover type is either too expensive or may suffer from land access permission issues. Hence, we guess cover type from variables that can be more easily obtained.
- Objective: Estimate population mean differences in hillside shade at noon (HS12) for different cover types. Let μ_1 and μ_2 be the population mean HS12 among sites having cover types 1 and 2, respectively. We aim to estimate $\mu_1 - \mu_2$ from our data.

Estimation and Confidence Interval for Mean Difference

- Over 50,000 observations, using the first 1,000 for analysis.
- Aim: Estimate the difference in population mean HS12 between cover types 1 and 2.
- Sample means for HS12 in cover types 1 and 2: 223.8 and 226.3, with s values of 15.3 and 14.3, and the sample sizes were 226 and 585.
- We will find an approximate 95% confidence interval for $\mu_1 - \mu_2$.
- Confidence interval for mean difference: $(-4.8, -0.3)$.
- Interpretation: The difference is not very large, suggesting HS12 might not be a strong predictor for cover type.

Population Proportions for Forest Covers Data

- Analyzing the difference in population proportions for cover types 1 and 2.
- Sample proportion difference:
  $\widehat{p}_1 - \widehat{p}_2 = 0.226 - 0.585 = -0.359$.
- Standard error calculation:
  $\sqrt{0.001 \cdot 0.226 \cdot 0.774 + 0.001 \cdot 0.585 \cdot 0.415} = 0.02043769$

Confidence Interval for Proportion Difference

- Confidence interval for proportion difference:
  $-0.359 \pm 1.96 \cdot 0.020 = (-0.399, -0.319)$.
- **Conclusion:** This interval suggests a substantial difference in proportions, likely indicating more sites of type 2.

Conclusion

- The study illustrates the use of confidence intervals in environmental data analysis.
- Highlights the importance of confidence intervals in interpreting the significance of results, both for means and proportions.

Student-t Distribution

- John Tukey (a pioneering statistician) emphasized on "an approximate answer to the right question" being better than "an exact answer to the wrong question".
- Will introduce the Student-t distribution and its application in statistical analysis.

Definition of the Student-t Distribution

- The Student-t distribution is defined for the quantity $T = \frac{\overline{W} - \mu}{\tilde{s}/\sqrt{n-1}}$.
- $\tilde{s}^2$ is the version of sample variance where we divide by $n - 1$ instead of n.

Assumptions and Characteristics

- Assumes the sampled population has a normal distribution.
- The general definition of the Student-t family is distribution of ratios $U/\sqrt{V/k}$, where
  - U has a N(0,1) distribution
  - V has a chi-squared distribution with k degrees of freedom
  - U and V are independent

Student-t Distribution in Practice

- The distribution is tabulated or use a statistical software like R.
- **Example:** When computing confidence interval for a sample size of 10, you would use the Student-t distribution.

Why Not Always Use the Student-t Distribution?

- The parent population must have an exact normal distribution, which is rarely the case in reality.
- For large n, the difference between the t-distribution and (standard) normal distribution is negligible.

- The Student-t distribution provides a more precise confidence interval for small sample sizes.
- However, its assumptions limit its applicability in many real-world scenarios.

Significance (or Hypothesis) Tests

Introduction to Significance Tests

- Significance tests form the core of statistical methods used across various scientific disciplines.
- Their presence is ubiquitous in scientific journals in fields like medicine, psychology, and economics.

Criticism of Significance Tests

- In 2016, the American Statistical Association issued a policy statement highlighting the overuse and misinterpretation of significance tests.
- The statement emphasizes that concerns about significance tests are not new but have persisted for decades.

Further Developments

- In 2019, a Nature article echoed the ASA's statement, further emphasizing the issue.
- These developments raise questions about the role and interpretation of significance tests in scientific research.

Understanding Significance Tests

- To understand the concept, consider a simple example: deciding whether a coin is fair (i.e., has a heads probability of 0.5).
- This example sets the stage for exploring the mechanics and implications of significance testing.

- While significance tests are fundamental to statistical analysis, their application requires careful consideration and interpretation.
- The recent critiques highlight the need for a deeper understanding of their implications in scientific research.

The Proverbial Fair Coin: Assessing Fairness in Coin Toss

- **Context:** Using a coin flip at the Super Bowl to determine the first kickoff.
- **Objective:** Assess the fairness of the coin.
- **Defining fairness:** p represents the probability of the coin landing heads, $p = 0.5$.

Methodology for Assessing Fairness

- Toss the coin 100 times.
- Form a confidence interval for p to determine fairness.
- Margin of error and interval location provide insights into the coin's fairness.

- **Example:** Interval (0.49, 0.54) suggests reasonable fairness.
- **Key point:** Even an interval like (0.502, 0.506) indicates fairness in practice, as it is near 0.5.

Traditional Statistical Approach

- **Traditional method:** Test the null hypothesis $H_0 : p = 0.5$ against the alternate hypothesis $H_A : p \neq 0.5$.
- This procedure is known as significance testing.
- Significance testing is central to statistical inference.

- While widely used, significance testing has recognized problems.
- Next, we'll explore both the mechanics and the criticisms of significance testing.

- Understanding the use and implications of significance tests is crucial for statistical inference.
- Careful interpretation is needed to assess "fairness" (of the coin) in real-world scenarios.

The Basics of Significance Testing

- **Approach:** Treat H_0 (null hypothesis) as true unless data strongly suggest otherwise.
- **Plan:** Toss a coin n times, consider it fair unless the number of heads is extremely high or low.

Statistical Methodology

- Let p be the true probability of heads, and $\widehat{p}$ the proportion in our sample.
- The standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/n}$, leading to the standard error of $\sqrt{\widehat{p}(1-\widehat{p})/n}$.

Computing the Z-score

- Assuming H_0 is true (i.e., $p = 0.5$), compute Z using the formula:

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{\frac{1}{n} \times 0.5(1 - 0.5)}}$$

- Z follows an approximate N(0,1) distribution under H_0.

Interpreting the Z-score

- Use Z to test the fairness of the coin.
- If $Z < -1.96$ or $Z > 1.96$, we reject H_0 at the 5% significance level (α).
- **Example:** For 100 coin tosses, reject H_0 if we get fewer than 40 or more than 60 heads.

Significance vs. Importance

- "Significant" does not imply "important".
- It indicates that the observed value of Z is a rare event under H_0.
- Significance testing forms a core part of statistical inference.
- Requires careful interpretation to determine the fairness or bias in a scenario like a coin toss.

General Testing Based on Normally Distributed Estimators

- We discussed the construction of confidence intervals before, and now we will discuss significance testing for normally distributed estimators.
- Then extend the methodology to general statistical estimators.

Formulating the Test Statistic

- Consider an estimator $\widehat{\theta}$ for a population value θ.
- To test $H_0 : \theta = c$, use the test statistic:

$$Z = \frac{\widehat{\theta} - c}{\text{s.e.}(\widehat{\theta})}$$

where s.e.$(\widehat{\theta})$ denotes the standard error of $\widehat{\theta}$.

Significance Testing Procedure

- The process involves rejecting H_0 at the significance level $\alpha = 0.05$ if $|Z| \geq 1.96$.
- This methodology is a generalization of the approach used in specific cases, like coin tosses.
- This approach provides a systematic method to conduct significance tests for a variety of estimators.
- Essential for statistical inference and hypothesis testing in various research contexts.

Understanding the Notion of "p-Values"

- Revisiting the coin example: 62 heads resulted in $Z = 2.4$.
- Strong rejection of H_0 at the 5% level.
- We next introduce the concept of **p-values** as a measure of significance.

Defining p-Values

- **p-value:** The smallest level at which H_0 would be rejected.
- Computed from the area under the N(0,1) distribution curve.
- The smaller the p-value, the more significant the result.

Computing p-Values

- **Example:** For $Z = 2.40$, p-value is 0.016.
- Indicates rejection of H_0 at the 1.6% level.
- Significance is inversely related to the p-value.

Interpreting p-Values

- Extremely small p-values indicate very high significance.
- In practice, p-values are often denoted by asterisks in statistical reports.
- **Example:** In R, one asterisk for $p < 0.05$, two for $p < 0.01$, etc.

Conclusion

- p-values provide a quantitative measure of the strength of evidence against H_0.
- Essential for understanding the significance of statistical tests

Understanding Randomness in Hypothesis Testing

- Note that H_0 is not a random event.
- The true value of p does not change; it's a property of the coin.
- Incorrect to speak of the probability that H_0 is true.

Misconceptions in Probability Statements

- Thus, it's incorrect to write $0.05 = P(|Z| > 1.96|H_0)$ as H_0 is not a random event.
- Proper notation is $0.05 = P_{H_0}(|Z| > 1.96)$.
- This means the probability that $|Z|$ exceeds 1.96 under the assumption that H_0 is true.

What's Wrong with Significance Testing and Alternatives

- **Significance testing:** Mathematically correct, but often noninformative or misleading.
- Origin of the 5% significance level by Sir Ronald Fisher in the 1920s.
- Continued use despite opposition and recognition of its limitations.
- Many statisticians are aware of its faults but are constrained by standards and practices.

Recommendations and Alternatives

- Reporting test results while acknowledging their limitations.
- Complementing significance tests with confidence intervals and other methods.
- Emphasizing a holistic understanding of data beyond just p-values and significance levels.

- Importance of critical and informed use of statistical methods.
- Significance testing has its place but should be used judiciously and in context.

The Basic Fallacy of Significance Testing

- Testing H_0 is inherently flawed because it's known a priori that H_0 is false.
- Example: For any real coin, $H_0 : p = 0.500000000...$ is practically impossible.
- This renders significance testing nonsensical from the start.

The Misleading Nature of "Significant"

- A coin with $p = 0.502$ is practically fair, but a large enough sample size could falsely identify it as "significantly" biased.
- The misuse of the term "significant" leads to misunderstanding the practical importance of the findings.

Problems with Significance Testing

- **Large sample sizes:** Tiny differences may be labeled as "significant."
- **Small sample sizes:** Could miss important differences, failing to identify actual significant effects.

- H_0 **is incorrectly specified:** The interest is in whether p is near 0.5, not exactly 0.5.
- **Misinterpretation of "significant":** It should not be confused with "important."

- Significance testing, while core to statistics, is noninformative or misleading.
- This perspective is recognized by statisticians and scientists, but the practice is deeply entrenched.

Critical Thinking in Statistics and Alternatives to Significance Testing

The Importance of Understanding Statistics

- Statistics is not just for learning but for practical use in various aspects of life.
- Apply critical thinking, especially about the problems of significance testing.
- Necessary for each individual to form their own opinion on the use of significance testing.

Alternatives to Significance Testing

- Making informed decisions is more important rather than relying solely on significance tests.
- Suggestion to set practical limits of fairness in hypothesis testing.
- **Example:** Testing $H_0 : 0.49 \leq p \leq 0.51$ for a coin's fairness.

Confidence Intervals: A Superior Approach

- Confidence intervals provide both accuracy (width) and fairness (location).
- Decision-making should not be based solely on whether a specific value is within the interval.
- **Example:** Accepting a coin as fair based on an interval close to 0.5, even if 0.5 is not included.
- The need for a thoughtful approach to statistical analysis.
- Understand the limitations of significance testing and explore more informative alternatives.

Decision-Making Based on Preponderance of Evidence

- Decision-making should be based on preponderance of evidence in statistical analysis.
- Significance Testing is analogous to the stringent proof standards in criminal trials.

Statistical Data as Evidence

- Statistical data should be viewed as evidence rather than conclusive proof.
- The width of a confidence interval indicates the likely accuracy of this evidence.

Integrating Evidence in Decision-Making

- Decision-making should integrate statistical evidence with other relevant information.
- The goal is to make a decision based on the overall preponderance of evidence.

- One should not rely solely on formulaic methods like significance testing.
- A more thoughtful and holistic approach to data analysis and decision-making is needed.

Conclusion

- Making informed decisions in statistics involves weighing all available evidence.
- Move beyond rigid statistical procedures to embrace a broader view of evidence and its interpretation.