Understanding Inferential Statistics: Confidence Intervals, Hypothesis Testing, and Logistic Regression

Ratnesh Prasad Srivastava, CSIT , GGV , Bilaspur, C.G.

1 Introduction to Statistical Inference

Statistical inference allows us to draw conclusions about population parameters based on sample statistics. Since we rarely have data for entire populations, we use samples to make educated guesses about population characteristics. The formulas presented here represent fundamental tools for this purpose.

2 Confidence Interval Estimation

Confidence intervals provide a range of plausible values for a population parameter. The confidence level (typically 95%) indicates how often the interval would contain the true parameter if we repeated the sampling process many times.

2.1 The Role and Importance of Confidence Intervals

Confidence intervals serve several crucial roles in statistical analysis:

- **Estimation Precision**: They provide a range of plausible values for a population parameter, giving a more complete picture than a single point estimate.
- **Uncertainty Quantification**: The width of the interval indicates the precision of our estimate. Narrow intervals suggest more precise estimates.
- **Hypothesis Testing**: If a confidence interval does not contain a particular value (like zero for differences or 1 for ratios), we can reject the null hypothesis that the parameter equals that value.
- **Clinical/Practical Significance**: Even when a result is statistically significant, the confidence interval shows whether the effect size is practically important.

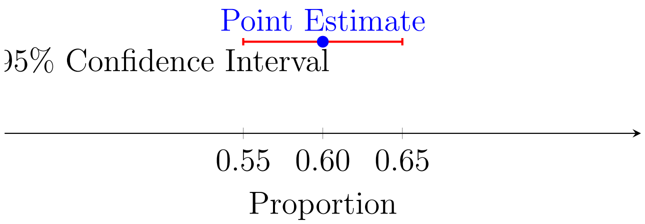

Figure 1: Visual representation of a confidence interval around a point estimate

Detailed Example of Confidence Interval Interpretation

Suppose a pharmaceutical company tests a new drug and finds it reduces symptoms in 60% of patients ($\hat{P} = 0.60$), with a 95% confidence interval of (0.55, 0.65) based on a sample of 400 patients.

Interpretation:

1. We are 95% confident that the true effectiveness of the drug in the population is between 55% and 65%.
2. If we were to repeat this study 100 times with different random samples, approximately 95 of the resulting confidence intervals would contain the true population parameter.
3. The interval does not mean there is a 95% probability that the true value is between 55% and 65% (a common misconception). The true value is fixed; the interval either contains it or not.
4. The margin of error is ±5%, which reflects the uncertainty in our estimate due to sampling variability.
5. If the company had hoped for at least 50% effectiveness, we can be confident the drug meets this threshold since the entire interval is above 50%.

2.2 Proportion Confidence Intervals

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Intuition and Explanation

This formula creates an interval estimate for a population proportion:

- $\hat{P}$ is the sample proportion (our best guess for the population proportion)

- $z_{\alpha/2}$ is the critical value from the standard normal distribution
- The expression under the square root is the standard error of the proportion

The "margin of error" (the part after the ±) accounts for sampling variability. A larger sample size (n) reduces the margin of error, making our estimate more precise.

Example

Suppose we survey 500 people and find that 60% ($\hat{P} = 0.60$) support a policy. For a 95% confidence interval ($z_{\alpha/2} = 1.96$):

$$0.60 \pm 1.96\sqrt{\frac{0.60(1-0.60)}{500}} = 0.60 \pm 1.96\sqrt{0.00048} \approx 0.60 \pm 0.043$$

We can be 95% confident that the true population proportion is between 55.7% and 64.3%.

2.3 Difference in Proportions

$$(\hat{P}_1 - \hat{P}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Intuition and Explanation

This formula estimates the difference between two population proportions:

- $\hat{P}_1$ and $\hat{P}_2$ are sample proportions from two groups
- n_1 and n_2 are the respective sample sizes
- The standard error now includes variability from both groups

If the confidence interval contains 0, we cannot conclude that there's a statistically significant difference between the groups.

Example

Group 1 (200 people): 40% success rate ($\hat{P}_1 = 0.40$)
Group 2 (250 people): 32% success rate ($\hat{P}_2 = 0.32$)

$$(0.40 - 0.32) \pm 1.96\sqrt{\frac{0.40(0.60)}{200} + \frac{0.32(0.68)}{250}} = 0.08 \pm 1.96\sqrt{0.0012 + 0.00087}$$

$$0.08 \pm 1.96\sqrt{0.00207} \approx 0.08 \pm 0.089$$

The 95% confidence interval is $(-0.009, 0.169)$. Since it contains 0, we cannot conclude with 95% confidence that there's a true difference between the groups.

3 Hypothesis Testing

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It allows us to make probabilistic conclusions about population parameters based on sample data.

3.1 Null and Alternative Hypotheses

- **Null Hypothesis (H)**: The default assumption that there is no effect, no difference, or no relationship. It represents the status quo.
- **Alternative Hypothesis (H or H)**: The claim we're trying to find evidence for. It represents a new effect, difference, or relationship.

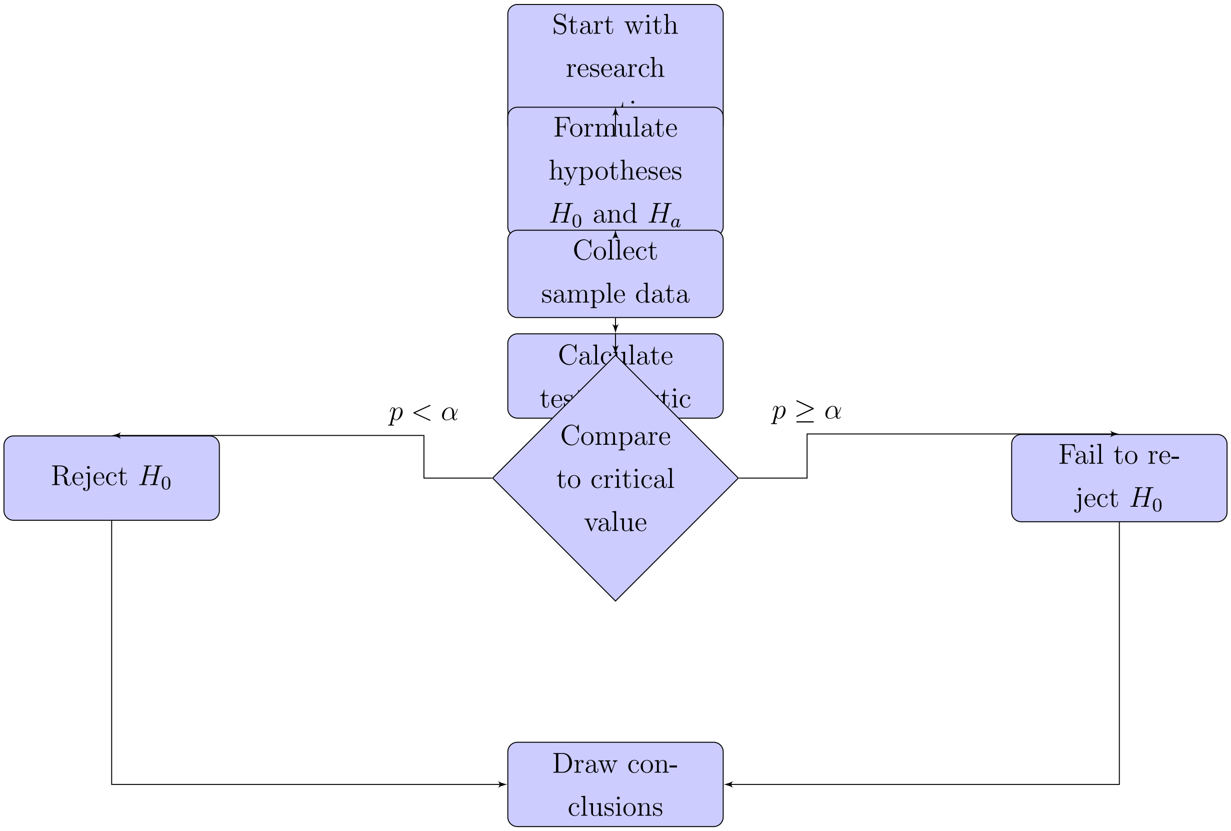

Figure 2: Flowchart of the hypothesis testing process

Example

Suppose we want to test whether a new teaching method improves exam scores compared to the traditional method.

- H: The new teaching method does not improve exam scores ($\mu_{\text{new}} \leq \mu_{\text{traditional}}$)
- H: The new teaching method improves exam scores ($\mu_{\text{new}} > \mu_{\text{traditional}}$)

3.2 p-value

The p-value is the probability of obtaining results at least as extreme as the observed results, assuming the null hypothesis is true.

- A small p-value (typically 0.05) indicates strong evidence against the null hypothesis.
- A large p-value (¿ 0.05) suggests weak evidence against the null hypothesis.

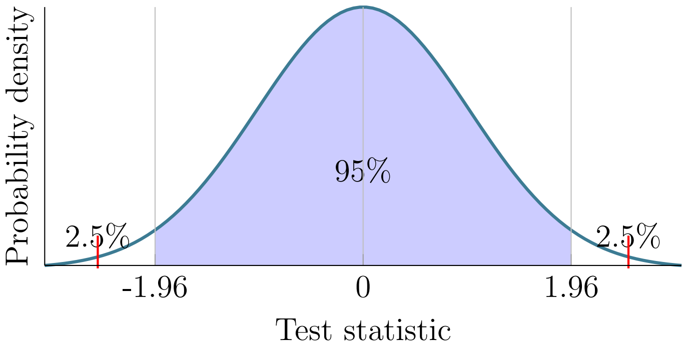

Figure 3: Visual representation of p-values in a standard normal distribution. The red lines represent critical values at =0.05.

Interpretation

- p-value ¡ 0.01: Very strong evidence against H
- 0.01 p-value ¡ 0.05: Strong evidence against H
- 0.05 p-value ¡ 0.10: Weak evidence against H
- p-value 0.10: Little to no evidence against H

Example

In our teaching method example, suppose we calculate a p-value of 0.03. This means that if the new teaching method were actually not effective (H true), we would observe a difference as large as we did only 3% of the time by random chance. This provides evidence to reject H in favor of H.

4 Regression Analysis

Regression models help us understand relationships between variables, particularly how independent variables affect a dependent variable.

4.1 Logistic Regression

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Intuition and Explanation

Logistic regression is used when the outcome variable is binary (Yes/No, Success/Failure):

- The left side is the log odds (logit) of the probability that $Y = 1$
- β_0 is the intercept (log odds when all predictors are 0)
- $\beta_1, \ldots, \beta_k$ are coefficients representing the change in log odds for a one-unit change in each predictor

We use the logistic function because probabilities must be between 0 and 1, while linear regression could predict values outside this range.

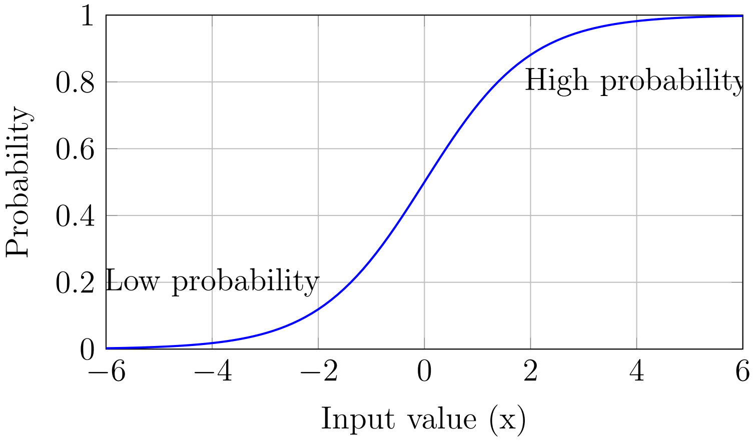

Figure 4: The logistic function (S-shaped curve) used in logistic regression

Example

Suppose we model the probability of passing an exam ($Y = 1$) based on hours studied (X_1) and prior GPA (X_2). After analysis, we might get:

$$\log\left(\frac{P(\text{Pass})}{1 - P(\text{Pass})}\right) = -3.5 + 0.8X_1 + 1.2X_2$$

Interpretation:

- For each additional hour studied, the log odds of passing increase by 0.8 (holding GPA constant)
- For each additional GPA point, the log odds of passing increase by 1.2 (holding study hours constant)

- A student who studies 10 hours with a GPA of 3.0 has predicted log odds of: $-3.5 + 0.8(10) + 1.2(3.0) = 5.1$
- Converting to probability: $P(\text{Pass}) = \frac{e^{5.1}}{1+e^{5.1}} \approx 0.994$ (99.4% chance of passing)

4.2 Confusion Matrix

A confusion matrix is a table that describes the performance of a classification model by comparing predicted labels to actual labels.

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Figure 5: Confusion matrix with clearly separated actual and predicted categories

Key Metrics

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$ - Overall correctness of the model
- **Precision**: $\frac{TP}{TP+FP}$ - When predicting positive, how often correct
- **Recall/Sensitivity**: $\frac{TP}{TP+FN}$ - How many actual positives were identified correctly
- **Specificity**: $\frac{TN}{TN+FP}$ - How many actual negatives were identified correctly
- **F1 Score**: $2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$ - Harmonic mean of precision and recall

Example

Suppose we have a medical test for a disease:

- True Positives (TP): 90 (correctly identified diseased patients)
- False Positives (FP): 10 (healthy patients incorrectly identified as diseased)
- False Negatives (FN): 5 (diseased patients incorrectly identified as healthy)

- True Negatives (TN): 95 (correctly identified healthy patients)

$$
\begin{aligned}
\text{Accuracy} &= \frac{90+95}{90+95+10+5} = \frac{185}{200} = 0.925 \quad (92.5\%) \\
\text{Precision} &= \frac{90}{90+10} = 0.90 \quad (90\%) \\
\text{Recall} &= \frac{90}{90+5} \approx 0.947 \quad (94.7\%) \\
\text{Specificity} &= \frac{95}{95+10} \approx 0.905 \quad (90.5\%) \\
\text{F1 Score} &= 2 \times \frac{0.90 \times 0.947}{0.90 + 0.947} \approx 0.923 \quad (92.3\%)
\end{aligned}
$$

Conclusion

These statistical tools allow researchers to:

1. Estimate population parameters with confidence intervals that quantify uncertainty
2. Test hypotheses about population parameters using formal procedures
3. Compare groups by examining differences in proportions with associated confidence intervals
4. Model relationships between variables, even with binary outcomes, using logistic regression
5. Evaluate classification models using confusion matrices and related metrics

Understanding these concepts provides a foundation for interpreting many research findings across various fields, from medicine to social sciences to business analytics.