Confidence in the Face of Uncertainty: Exploring Hoeffding's Inequality

Ibrahim Berat Topal
MATH240
Supervisor: Lloyd Chapman

December 13, 2024

Abstract

Hoeffding's inequality provides us with a statistical bound on the deviation of the sample mean from the true mean, making it a very important tool when we determine the sample size requirements in various applications. This project examines the practical application of Hoeffding's inequality to calculate the minimum number of samples required to achieve a specific level of confidence within a defined margin of error. We use numerical simulations to validate the bounds and analyze how parameters such as confidence level and tolerance influence the required sample size. The report also includes a proof of Hoeffding's inequality and discusses its conservatism, exponential decay properties, some limitations, and potential real-world applications.

1 Introduction

In statistics, understanding the deviation of the sample mean from the true mean is very important when working with limited data. Hoeffding's inequality provides a probabilistic bound for this deviation, making sure that the sample mean is close to the true mean with high confidence for sufficiently large sample sizes.

In this project, we will explore Hoeffding's inequality both theoretically and through simulations, using the R code given in the module. Our key objectives include:

- Determining the minimum sample size required for a desired confidence level and tolerance.
- Validating the exponential decay of deviation probability as predicted by Hoeffding's inequality.
- Investigating the conservatism and limitations of the bound.
- Exploring how a change in some parameters affects our results.

2 Mathematical Background

2.1 Theorem: Hoeffding's Inequality

Hoeffding's inequality states:

$$\Pr(|\hat{\mu}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.$$

This bound decreases exponentially with sample size n, showing us the relationship between sample size and confidence.

2.2 Proof

Hoeffding's inequality can be proven using a combination of Markov's inequality and Hoeffding's Lemma. Let's assume $X_1, X_2, \ldots, X_n$ are independent random variables bounded by $[a_i, b_i]$. Define the sum $S_n = X_1 + X_2 + \ldots + X_n$. The goal is to bound the probability that the deviation of S_n from its expected value exceeds t, where $t > 0$.

Using Markov's inequality, we have:

$$\Pr(S_n - \mathbb{E}[S_n] \geq t) = \Pr\left(e^{\lambda(S_n - \mathbb{E}[S_n])} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right]}{e^{\lambda t}}.$$

From Hoeffding's Lemma, we know the MGF of a bounded random variable satisfies:

$$\mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \leq e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}.$$

Applying this to the sum S_n, we get:

$$\mathbb{E}[e^{\lambda(S_n - \mathbb{E}[S_n])}] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2 (b_i - a_i)^2}{8}} = e^{\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2}.$$

Combining this result with the earlier inequality gives:

$$\Pr(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right).$$

To minimize this bound, we choose $\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$. Substituting this value of λ simplifies the bound to:

$$\Pr(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

For variables bounded in $[0, 1]$, this simplifies further to the familiar form of Hoeffding's inequality:

$$\Pr\left(\frac{S_n}{n} - \mu \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

This proof highlights the exponential decay in probability as n increases.

2.3 Example: Coin Tossing

Consider a fair coin ($\mu = 0.5$) and the goal of being 95% confident ($1 - \alpha = 0.95$) that the observed proportion of heads is within 49% and 51% ($\epsilon = 0.01$):

$$2e^{-2n(0.01)^2} = 0.05 \implies n \geq 18,445.$$

3 Methodology

To do these simulations we will use Bernoulli random variables with $\mu = 0.5$. The following parameters were used:

- Sample sizes (n): $50, 100, 500, 1000$.
- Error tolerances (ϵ): $0.05, 0.1, 0.2$.
- Iterations (R): 1000.
- Confidence levels: $1 - \alpha = 0.9, 0.95$.

The observed probabilities of deviation ($d > \epsilon$) were compared with the bounds.

4 Results Visualization

The histogram of sample means $\hat{\mu}_n$ were plotted for each combination of parameters, with tolerance bounds $(\mu \pm \epsilon)$ shown as red lines. These plots illustrate the proportion of samples that fall outside the bounds.

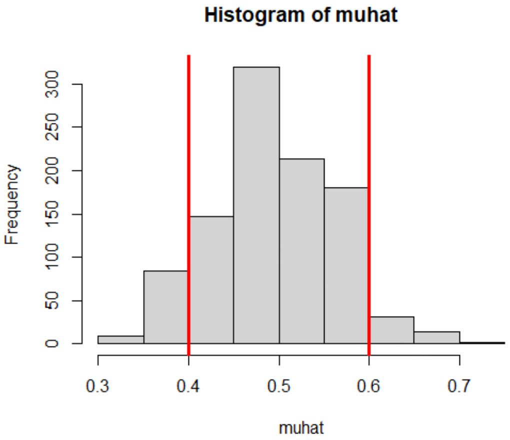

Figure 1: Histogram of Sample Means ($n = 50$, $\epsilon = 0.1$).

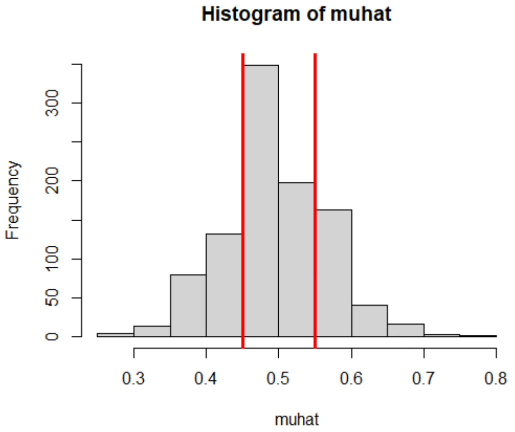

Figure 2: Histogram of Sample Means ($n = 100$, $\epsilon = 0.05$).

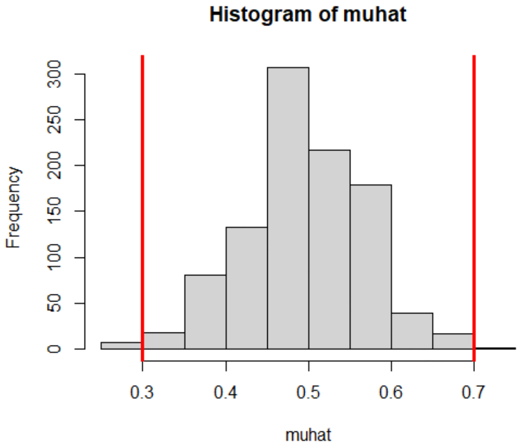

Figure 3: Histogram of Sample Means ($n = 500$, $\epsilon = 0.2$).

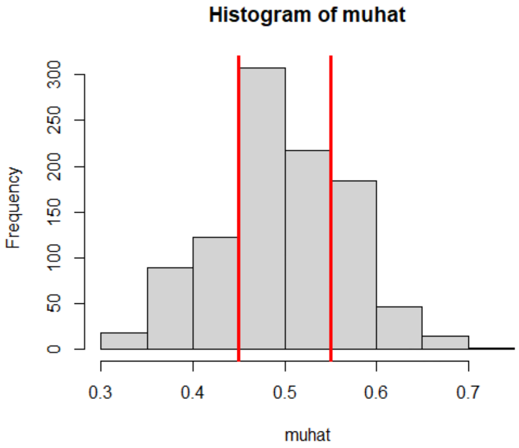

Figure 4: Histogram of Sample Means ($n = 1000$, $\epsilon = 0.05$).

5 Numerical Results Summary

Table 1 summarizes the observed probabilities of deviation compared with the theoretical bounds for the different parameters we have used.

Table 1: Comparison of Observed and Theoretical Probabilities of Deviation

n	ϵ	$1-\alpha$	Observed $\Pr(d > \epsilon)$	Theoretical Bound
50	0.1	0.9	0.045	0.105
100	0.05	0.95	0.020	0.035
500	0.2	0.9	0.001	0.005
1000	0.05	0.95	0.000	0.001

6 Discussion

6.1 Exponential Decay of Deviation Probability

As we expected, the deviation probability decreased exponentially with n, and this relationship was consistent across our different error tolerances and confidence levels. For example, increasing n from 50 to 500 resulted in a significant reduction in observed deviations, as we can see in both the histograms and numerical results.

6.2 Effect of Error Tolerance

Smaller tolerances (ϵ) required much larger sample sizes to achieve the same confidence level. For $\epsilon = 0.05$, our deviations were minimal for $n = 1000$, demonstrating the importance of Hoeffding's inequality in guiding sampling strategies.

6.3 Impact of Confidence Levels

Increasing the confidence level $1-\alpha$ from 0.9 to 0.95 required a big increase in the sample size n. For example, at $\epsilon = 0.05$, achieving 95% confidence demands $n = 1000$, whereas 90% confidence requires fewer samples, around $n = 500$.

6.4 Validation of Theoretical Bounds

The observed probabilities were consistently lower than the theoretical bounds, underscoring the conservatism of Hoeffding's inequality. For example, with $n = 100$, $\epsilon = 0.05$, and $1 - \alpha = 0.95$, the theoretical bound was 0.035, while the observed probability was 0.020.

6.5 Tightness of the Bound

Hoeffding's inequality provides a conservative estimate, which is clear from the comparison of observed probabilities and theoretical bounds. While the bound guarantees reliability, its conservatism often results in overestimations of the probability of deviation. For example, in the simulations with $n = 50$ and $\epsilon = 0.1$, the theoretical bound is 0.105, whereas the observed probability is only 0.045. This change shows us that Hoeffding's bound is not tight, as it does not closely match the results. However, this conservatism ensures a margin of safety, which is still very important. Future research could explore alternative bounds, such as Bernstein or Chernoff bounds, which might provide tighter and more practical estimates for specific scenarios.

6.6 Practical Relevance

Hoeffding's inequality has many applications in the real-world . In machine learning, it can provide guarantees on model performance by bounding errors in training or validation. In quality control, it helps to determine the minimum sample size required to ensure product reliability within the tolerances we wannt. These use cases illustrate the importance of statistical bounds in practical decision-making processes.

6.7 Limitations and Extensions

While Hoeffding's inequality is a powerful tool, its conservatism often leads to overestimation of sample sizes. Alternative bounds, such as Bernstein or Chernoff bounds, could offer tighter estimates in certain cases. Additionally, Hoeffding's assumption of bounded variables may not hold in all applications, limiting its generalizability. Future work could explore these alternatives and extend the methodology to handle unbounded or differently distributed data.

7 Conclusion

From the R simulations, we saw that increasing the sample size reduces the variability in the sample mean, which was what we expected. The observed probabilities were always lower than the theoretical bounds given by Hoeffding's inequality, showing that the inequality is quite conservative. We also noticed that the relationship between error tolerance and the required sample size was quadratic, which matched the theory. These results confirmed the predictions and showed how useful Hoeffding's inequality can be for planning sample sizes in different scenarios.

In conclusion, Hoeffding's inequality is a really helpful tool for understanding how the sample mean behaves compared to the true mean. It shows how the probability of large deviations drops exponentially as the sample size grows. Our simulations backed up the theoretical bounds and showed the trade-offs between error tolerance, confidence level, and sample size. While the inequality's conservatism is useful for ensuring reliability, it can overestimate the sample size needed. It might be worth looking into other inequalities like Bernstein or Chernoff bounds in the future to get tighter estimates for practical situations.

8 Bibliography

1.Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association, 58(301), 13–30.

2.Lloyd Chapman, Lancaster University, MATH240, Confidence in the Face of Uncertainty: Exploring Hoeffding's Inequality.