BONTA AALAYA
12340530
DSL253 - Statistical Programming

1

INTRODUCTION

In this assignment, we observe the impact of sample size and confidence level on the accuracy of estimated mean and standard deviation for a product which follows normal distribution

DATA

Product data is generated using in-built matlab function <u>normrnd</u> and parameters were selected accordingly

METHODOLOGY

Question 1

Random samples are drawn for different values of n and α. Confidence intervals for the mean and standard deviation are computed. The proportion of intervals capturing the true parameters is analyzed over m = 1000 repetitions. The effect of random measurement noise is tested by introducing uniform distribution noise in the range (-1,1)

$$CI_{\text{mean}} = \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

$$CI_{\text{std}} = \left[s \cdot \sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}},\ s \cdot \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}} \right]$$

$$\sigma^2_{\text{total}} = \sigma^2_{\text{original}} + \frac{1}{3} \quad \text{(Var of uniform noise)}.$$

RESULTS

Coverage probability for mean estimates:

	0.01	0.05	0.1
10	0.989	0.963	0.894
30	0.988	0.951	0.908
50	0.986	0.957	0.893
100	0.995	0.952	0.907
200	0.99	0.945	0.908

Coverage probability for std dev estimates:

	0.01	0.05	0.1
10	0.987	0.949	0.897
30	0.992	0.942	0.91
50	0.989	0.944	0.894
100	0.991	0.946	0.894
200	0.988	0.942	0.901

Average width of CI for mean estimates:

	0.01	0.05	0.1
10	30.152	21.064	16.779
30	14.982	11.105	9.3258
50	11.208	8.5505	7.0795
100	7.862	5.9147	4.9593
200	5.5051	4.1719	3.504

Average width of CI for std dev estimates:

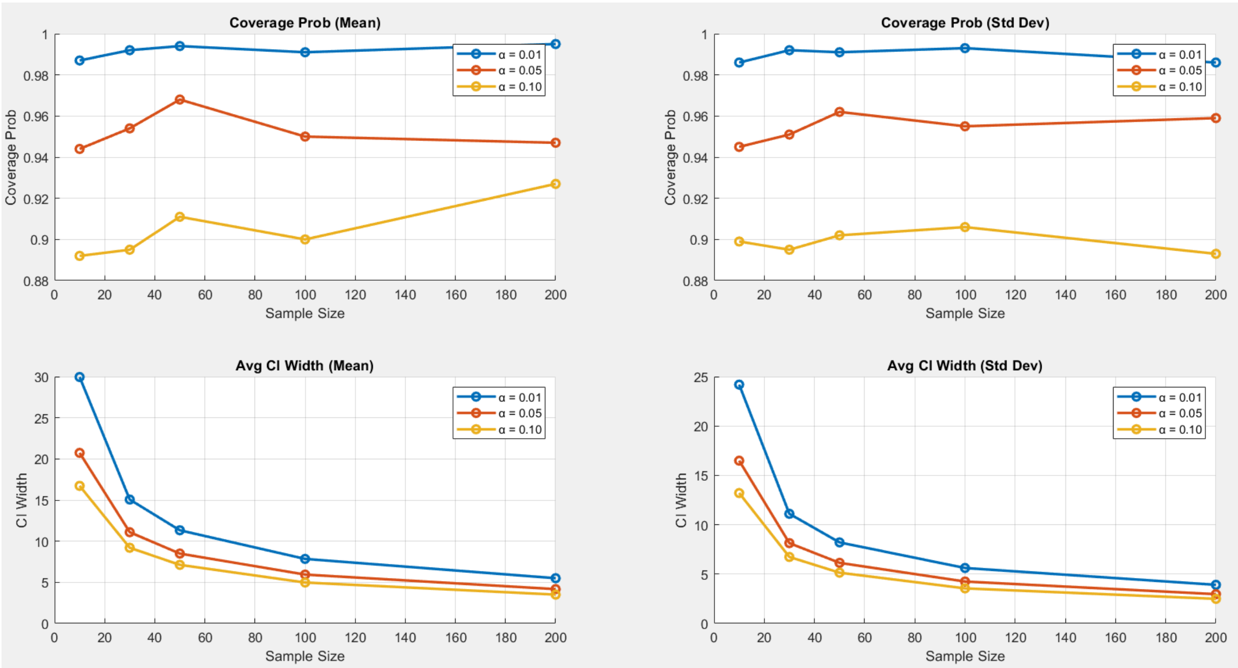
	0.01	0.05	0.1
10	24.351	16.751	13.255
30	11.049	8.1474	6.8273
50	8.1256	6.1798	5.1101
100	5.6286	4.2279	3.5427
200	3.9168	2.966	2.4903

Coverage for mean with noise:

	0.01	0.05	0.1
10	0.988	0.945	0.9
30	0.994	0.947	0.891
50	0.99	0.965	0.912
100	0.996	0.953	0.902
200	0.986	0.958	0.902

Coverage for std dev with noise:

	0.01	0.05	0.1
10	0.985	0.959	0.903
30	0.988	0.964	0.896
50	0.989	0.958	0.902
100	0.994	0.957	0.905
200	0.99	0.963	0.884



DISCUSSION

The results show that confidence intervals for the mean have high coverage probabilities at different sample sizes and confidence levels. As expected, larger sample sizes improved the reliability of coverage, particularly at 95% and 99% confidence levels. Smaller, had a slightly lower coverage, particularly at the 90% level, meaning there was a higher chance of failing to capture the true mean. The confidence intervals also performed well in estimating standard deviation, with the coverage probabilities being greater than 90% in almost all instances. Intervals for smaller sample sizes were more variable. Confidence interval width was always increasing with a pattern; wider intervals with higher confidence levels and with larger sample sizes, with more accurate estimates being produced by the larger sample sizes. With measurement noise , coverage probabilities for mean and standard deviation remained fairly stable. The moderate decrease in coverage of standard deviation estimates suggests that noise introduces additional variability, reducing confidence interval reliability

CONCLUSION

The study confirms that confidence intervals are a good method of estimating manufacturing parameters, with larger sample sizes improving precision. Smaller sample sizes reduce coverage by a minimal amount, especially in the estimation of standard deviation. Measurement noise reduces precision, and thus there is a need for regular machine calibration. Manufacturers need to use sample sizes of 50 to 100 and have strict quality control measures to ensure consistent manufacturing processes

2

INTRODUCTION

Pharmaceutical companies compare two formulations of a drug to determine their relative efficacy. This study compares confidence intervals for the difference in mean efficacy between two drug formulations based on data simulated to represent patients. We investigate the effects of sample size and variations in standard deviation on the accuracy of these intervals

DATA

Distributions are simulated using built-in functions and parameters are changed to observe differences in confidence intervals. Experiment is repeated to calculate success rate

METHODOLOGY

Question 2

Generate random samples from normal distributions for both drug formulations. Compute confidence intervals for the mean difference, where degrees of freedom are approximated using Welch's t-test formula. Repeat experiment 1000 times to calculate success rates

$$CI_{\text{diff}} = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

RESULTS

Results for parameter set 1:

mu1	sigma1	n1	mu2	sigma2	n2	Coverage	AvgCIWidth
120	10	30	115	12	30	0.942	11.382

Results for parameter set 2:

mu1	sigma1	n1	mu2	sigma2	n2	Coverage	AvgCIWidth
120	15	50	118	10	50	0.949	10.107

Results for parameter set 3:

mu1	sigma1	n1	mu2	sigma2	n2	Coverage	AvgCIWidth
100	20	100	105	20	100	0.947	11.14

DISCUSSION

Wider sample sizes result in more precise and stable confidence intervals, ensuring a higher proportion captures the true difference. Increasing variance in drug effects increases uncertainty, meaning wider samples should be used to provide reliable coverage. Confidence intervals widen and become less reliable if sample sizes are too small, risking failure to capture the true difference in drug effectiveness.

CONCLUSION

The simulation proves that sample size and variance have a significant effect on confidence interval accuracy in drug formulation comparison. Researchers should use proper sample sizes, account for variance levels in study design—greater variance necessitates larger samples, and interpret confidence intervals with care, especially for small samples where variability can confound results.

3

INTRODUCTION

Election polling aims to estimate voter support for candidates, but sample-based confidence intervals introduce uncertainty. This question examines the accuracy of Wilson score confidence intervals in estimating the proportion of voters supporting a candidate

DATA

Vaying parameters are chosen and the pdfs are generated using built-in functions, cdf is computed used wilson's

METHODOLOGY

Question 3

Random samples are generated on different population proportions and sample size is varied. Confidence intervals are calculated using updated wilson's formula below

$$CI_{\text{Wilson}} = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} \pm z_{\alpha/2} \cdot \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

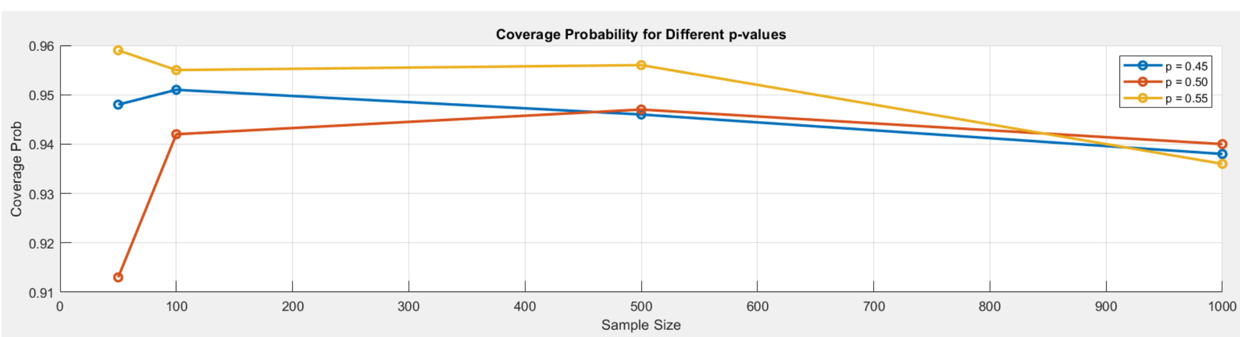
RESULTS

Cov Prob for Polling:

	50	100	500	1000
0.45	0.948	0.951	0.946	0.938
0.5	0.913	0.942	0.947	0.94
0.55	0.959	0.955	0.956	0.936

Avg Width of CI for Polling:

	50	100	500	1000
0.45	0.2632	0.19042	0.086819	0.061525
0.5	0.26433	0.19135	0.087232	0.06183
0.55	0.26343	0.19044	0.086791	0.061515



DISCUSSION

The outcomes show that the coverage probability is better with larger sample sizes, i.e., bigger samples give a more accurate estimate. For small samples (n = 50, 100), the confidence intervals were larger, resulting in more variability in coverage. This implies that smaller polls are more susceptible to sampling error and sometimes might not pick up the real population proportion within the CI. Also, for p = 0.50, the coverage probabilities tended to be higher than in the case of p = 0.45 or p = 0.55, implying that the precision of polls might be influenced by the true proportion of the population, especially if it is more towards the tails

CONCLUSION

The results highlight the importance of election sample size polling, as larger magnitude samples provided narrower confidence intervals and better coverage probabilities. The use of the adjusted Wilson score interval ensured that confidence intervals remained valid even under smaller samples, with less bias compared to previous practices. Outcomes obtained from surveys of smaller samples should be seriously taken into consideration owing to broader confidence intervals and poor coverage probability. These outcomes attest to the importance of properly large sample sizes in real polls in order to make reasonable election forecasts

---