# CSDS 313: Introduction to Data Analysis
# Homework 2: Statistical Inference

Camden Larson

Fall 2025

## Task 1: German Tank Problem

Recall: The Nazi Army assigns sequential integer IDs to tanks $1, 2, \ldots, M$. The Allied forces capture $n$ tanks with observed IDs $1 \leq x_1 \leq x_2 \leq \ldots \leq x_n$. The goal is to estimate $M$ based on uniform random sampling.

### (a) Maximum Likelihood Estimator

Compute the maximum likelihood estimator $\hat{M}_{MLE}$ for $M$.

The likelihood function is nonzero only for $M \geq x_n$, where $x_n$ is the largest observed tank ID, and it decreases as $M$ increases. Therefore, the maximum likelihood occurs at the smallest feasible value of $M$, which is $x_n$. Hence, the maximum likelihood estimator is:

$$\boxed{\hat{M}_{\text{MLE}} = x_n}$$

That is, the MLE for $M$ is simply the highest integer (tank ID) observed among the captured tanks.

## (b) Bias and Expected Value of Estimators

Let:

$$\hat{M}_{\text{MEAN}} = 2\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - 1, \quad \hat{M}_{\text{MVU}} = x_n\left(\frac{n+1}{n}\right) - 1$$

Derive and compare the expected values $E[\hat{M}_{MLE}]$, $E[\hat{M}_{MEAN}]$, and $E[\hat{M}_{MVU}]$. Comment on which estimators are biased or unbiased.

**Mean-based estimator.** $\hat{M}_{\text{MEAN}} = 2\bar{X} - 1$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Under random sampling without replacement from $\{1, \ldots, M\}$, each $X_i$ has marginal $\mathbb{P}(X_i = k) = 1/M$ for $k = 1, \ldots, M$, so $\mathbb{E}[X_i] = (M+1)/2$. By linearity of expectation, $\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = (M+1)/2$, and thus

$$\mathbb{E}[\hat{M}_{\text{MEAN}}] = 2\,\mathbb{E}[\bar{X}] - 1 = 2 \cdot \frac{M+1}{2} - 1 = M.$$

Therefore $\hat{M}_{\text{MEAN}}$ is unbiased for $M$.

**MVU estimator.** The minimum-variance unbiased (MVU) estimator is defined as

$$\hat{M}_{\text{MVU}} = \frac{n+1}{n} X_{(n)} - 1,$$

where $X_{(n)}$ is the largest observed tank ID in the sample of size $n$.

For samples drawn uniformly without replacement from $\{1, \ldots, M\}$, the expected value of the sample maximum is

$$\mathbb{E}[X_{(n)}] = \frac{n}{n+1}(M+1).$$

Substituting this into the estimator gives

$$\mathbb{E}[\hat{M}_{\text{MVU}}] = \frac{n+1}{n}\,\mathbb{E}[X_{(n)}] - 1 = \frac{n+1}{n} \cdot \frac{n}{n+1}(M+1) - 1 = M.$$

Thus, $\hat{M}_{\text{MVU}}$ is an **unbiased** estimator of $M$. It corrects the downward bias of $\hat{M}_{\text{MLE}}$ by scaling $X_{(n)}$ appropriately.

**MLE estimator.** The MLE is the sample maximum: $\hat{M}_{\text{MLE}} = X_{(n)}$. For uniform sampling without replacement from $\{1, \ldots, M\}$, the expected maximum is

$$\mathbb{E}[X_{(n)}] = \frac{n}{n+1}(M+1).$$

Hence

$$\mathbb{E}[\hat{M}_{\text{MLE}}] = \frac{n}{n+1}(M+1), \qquad \text{Bias}(\hat{M}_{\text{MLE}}) = -\frac{M-n}{n+1},$$

which is negative for $M > n$. Thus, the MLE is downward biased (it underestimates $M$ on average).

## (c) Simulation and Visualization

Simulate different instances for various $M$ (e.g., 100, 1000, 10,000) and varying $n$. For each $M$, repeat the sampling without replacement, compute the three estimators ($\hat{M}_{\text{MLE}}, \hat{M}_{\text{MEAN}}, \hat{M}_{\text{MVU}}$), and record their empirical means and variances across repetitions.

### Simulation setup

- For each $M \in \{100, 1000, 10{,}000\}$ and each $n$ in the given list, run `reps` simulations.
- In each simulation, draw $n$ unique IDs from $\{1, \ldots, M\}$, then compute the three estimators.
- Aggregate across repetitions to get the empirical mean and variance for each estimator at that $n$.

**Results for $M = 100$**

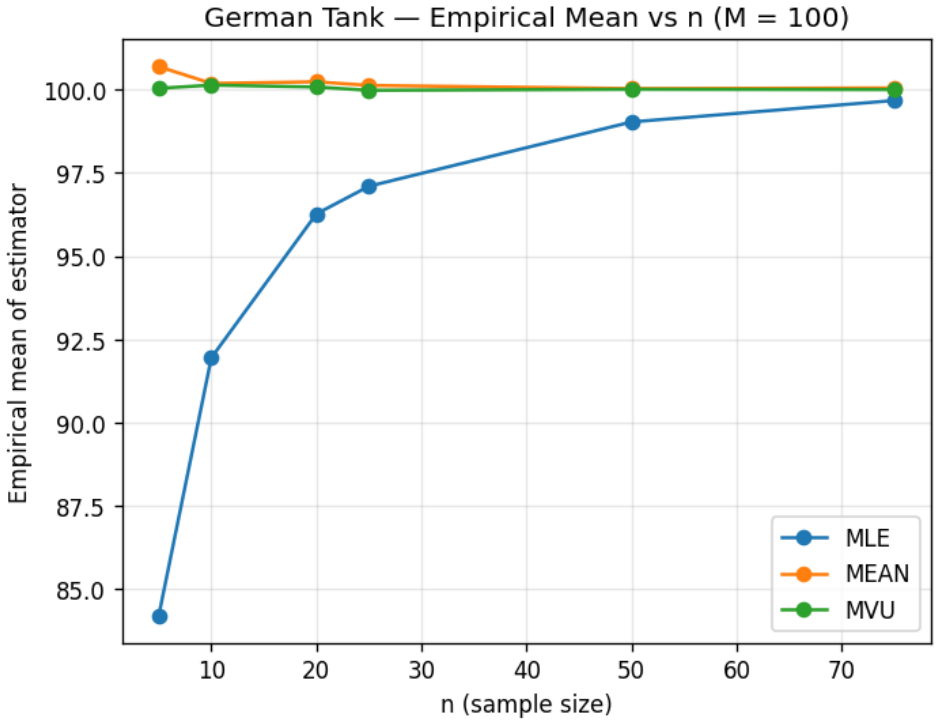


Figure 1: Empirical mean of each estimator vs. $n$ for $M = 100$.

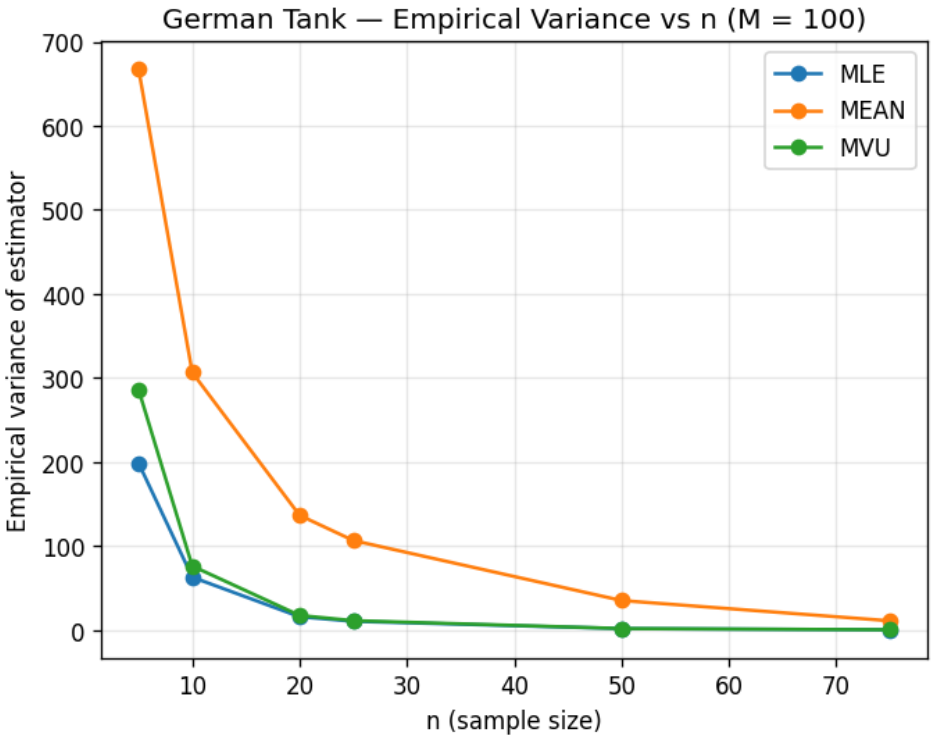


Figure 2: Empirical variance of each estimator vs. $n$ for $M = 100$.

**Results for $M = 1000$**

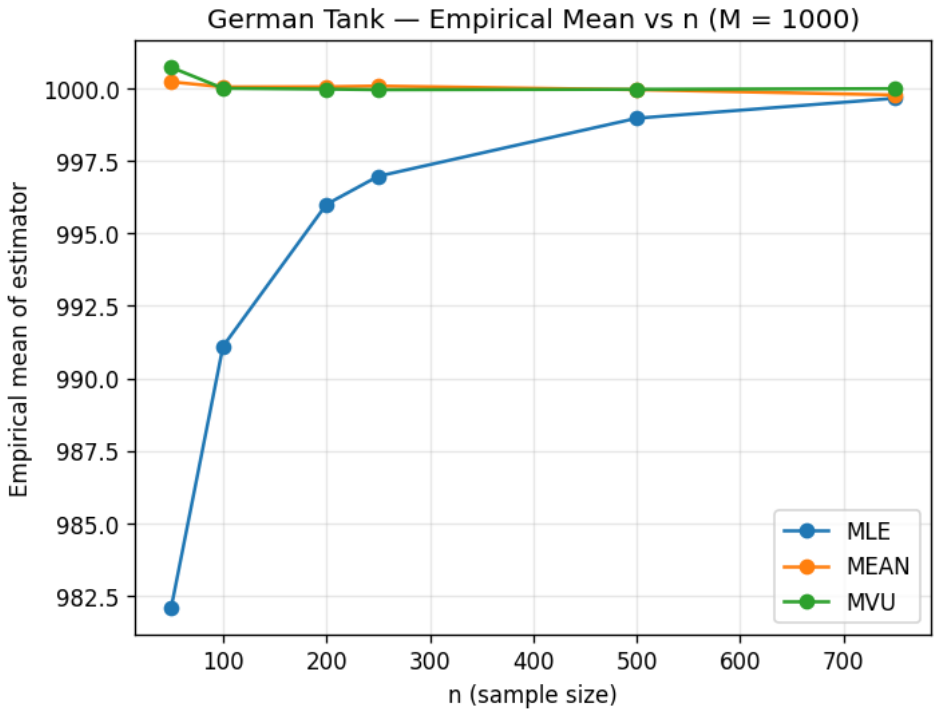


Figure 3: Empirical mean of each estimator vs. $n$ for $M = 1000$.

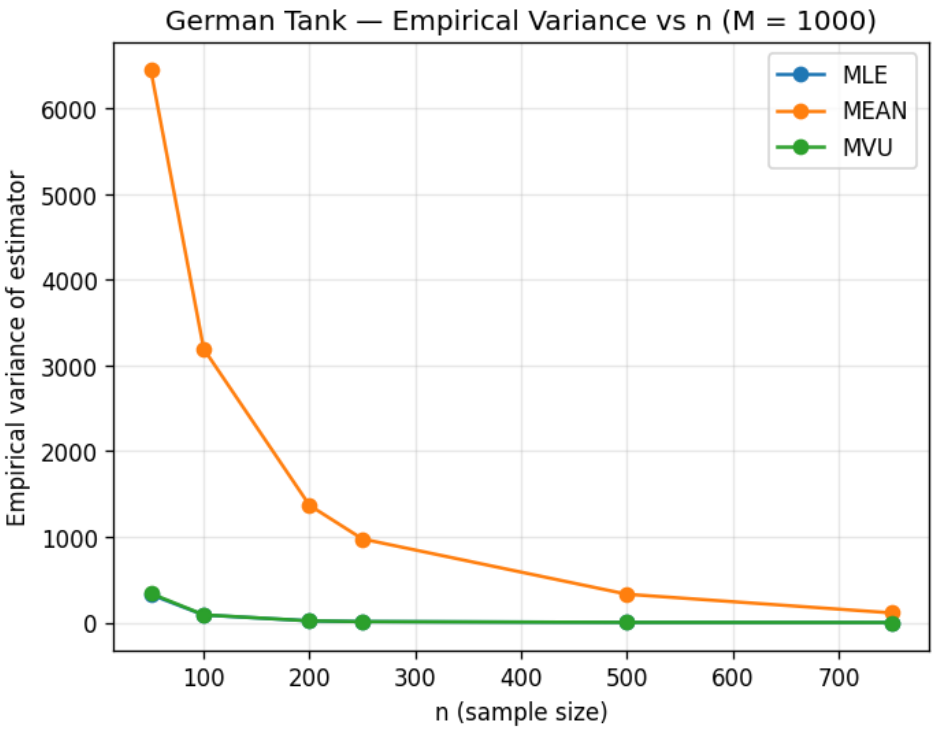


Figure 4: Empirical variance of each estimator vs. $n$ for $M = 1000$.

**Results for $M = 10{,}000$**

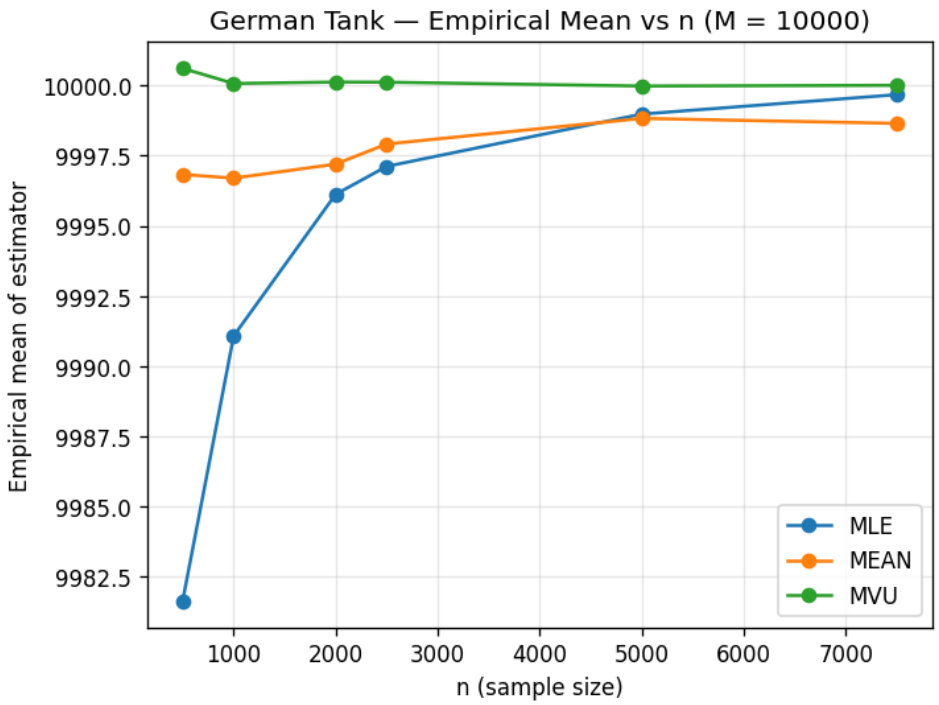


Figure 5: Empirical mean of each estimator vs. $n$ for $M = 10{,}000$.

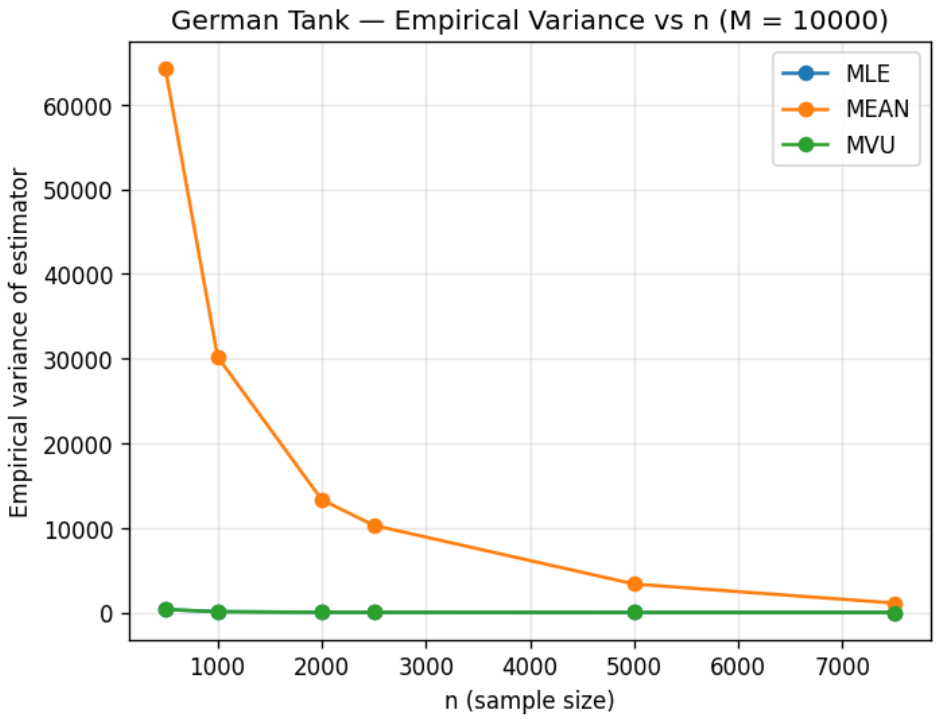


Figure 6: Empirical variance of each estimator vs. $n$ for $M = 10{,}000$.

**Discussion** ($M = 100, 1000, 10000$)**.**

- *Bias:* For both Mean and MVU, the empirical mean of the estimator hovered around M across all trials from n = 1/10th of M to n = 3/4 of m, while the MLE estimator underestimated the mean massively low n values and logarithmically approached 100 as n increased.
- *Variance:* For the variance at M, the MLE showed the least variance, rapidly decreasing to almost 0 at trials of low sample size, while MVU

followed a similar pattern starting with a slightly higher variance for low sample sizes. The variance of the Mean estimator was extremely high for small sample sizes, but exponentially decreased as the sample size increased. *Note For trials at M = 1000 and M = 10000, the variances of MLE and MVU were extremely similar causing the blue MLE line to be hidden behind the green MVU line.

# Task 2: Distribution of the Sample Mean

This task explores the Central Limit Theorem using two random variables:

- Uniform distribution on $[0.25, 1.25]$
- Power-law distribution with parameter $\alpha = 3$

## Part A: Data Generation

Generate datasets with sizes 10, 100, and 1000. Each replicate has 100 samples. Plot one replicate from each distribution to confirm shape.

**Uniform — Sample Distributions**

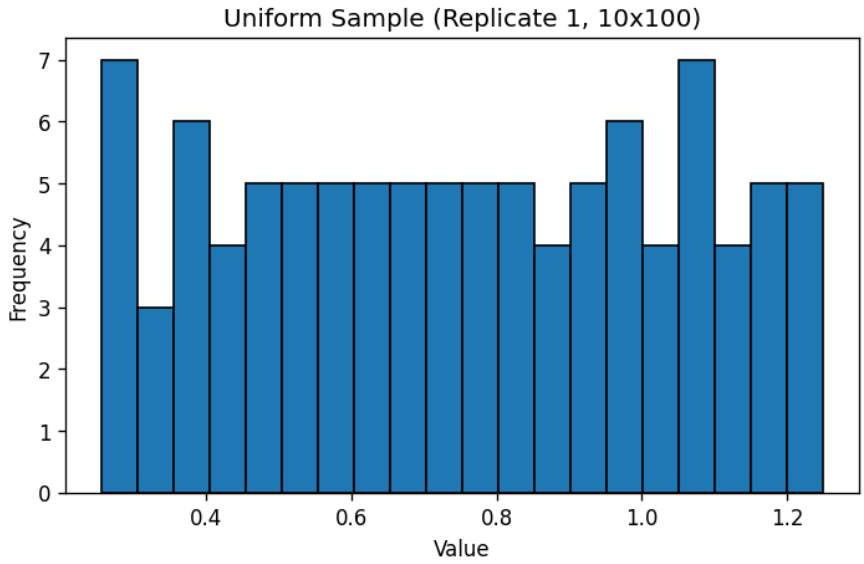


Figure 7: Uniform: Sample distribution (Replicate 1, $10 \times 100$).

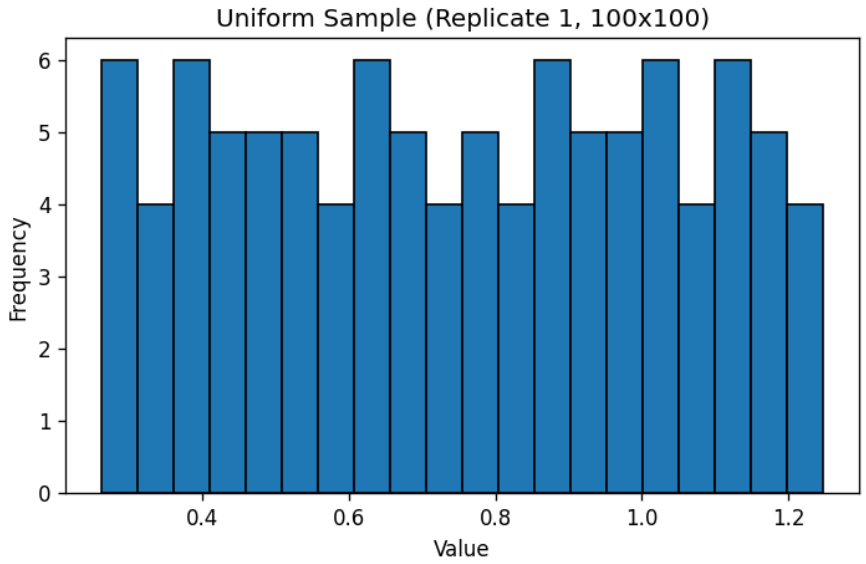


Figure 8: Uniform: Sample distribution (Replicate 1, $100 \times 100$).

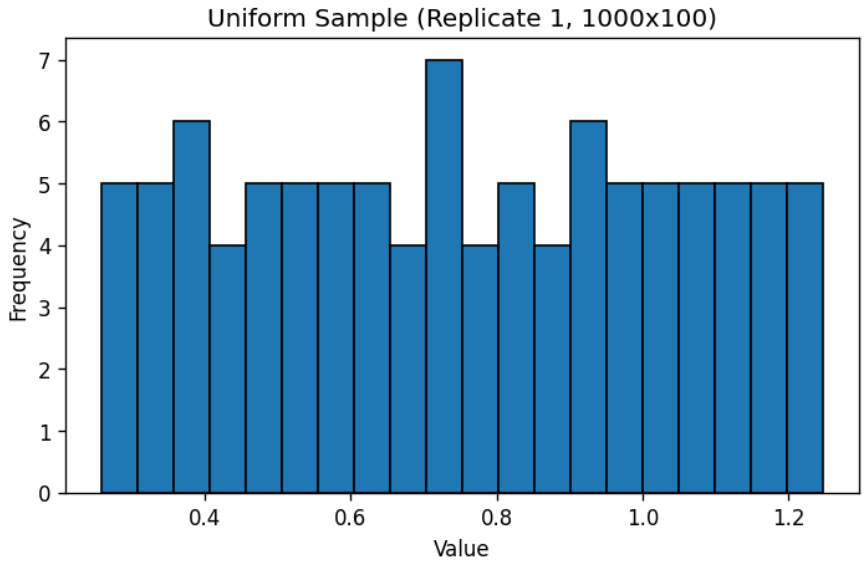


Figure 9: Uniform: Sample distribution (Replicate 1, $1000 \times 100$).

**Power-law — Sample Distributions**

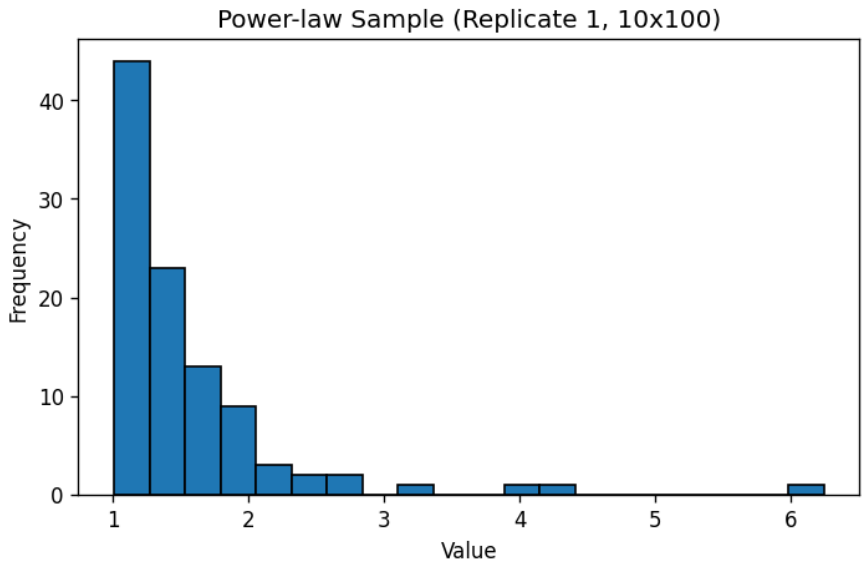


Figure 10: Power-law: Sample distribution (Replicate 1, $10 \times 100$).

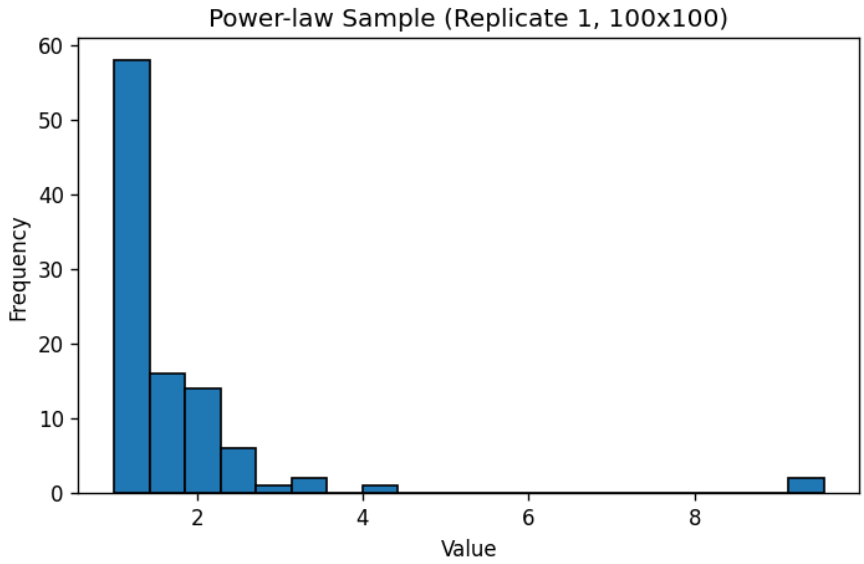


Figure 11: Power-law: Sample distribution (Replicate 1, $100 \times 100$).

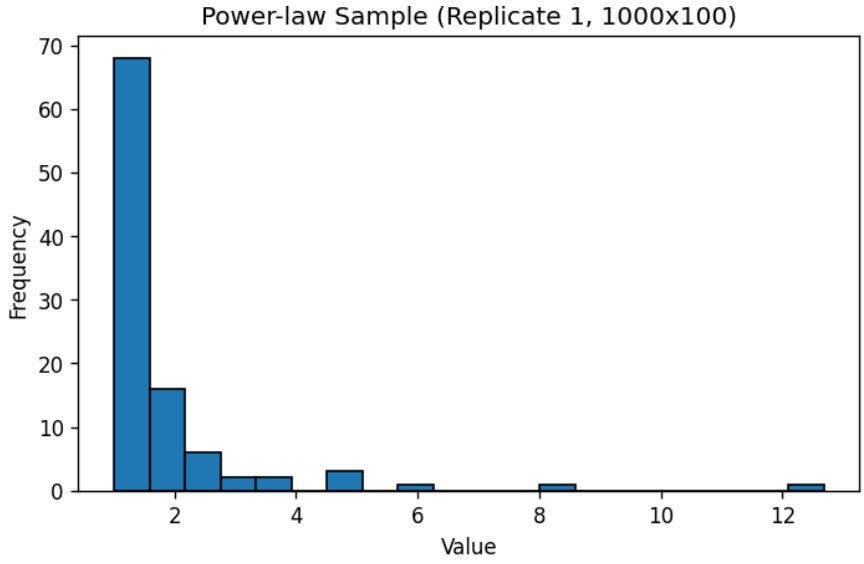


Figure 12: Power-law: Sample distribution (Replicate 1, $1000 \times 100$).

## Part B: Sample Means and Visualization

Compute and plot the sample means for each replicate and distribution.

### Uniform — Sample Mean Distributions

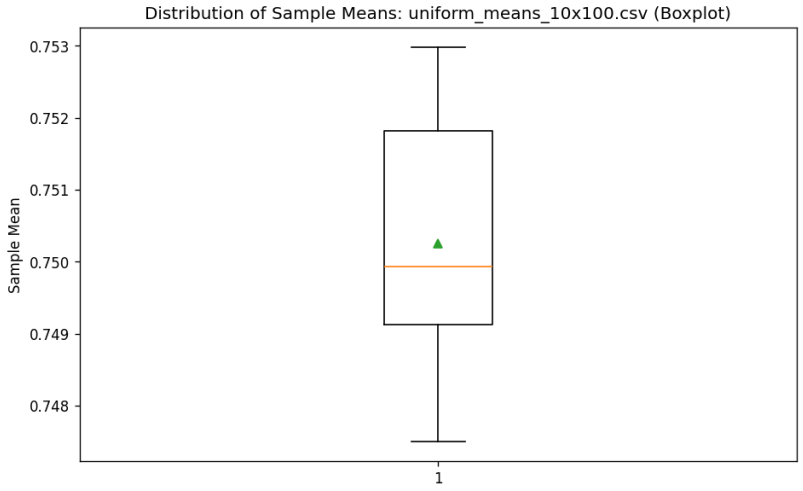


Figure 13: Uniform: Boxplot of sample means ($10 \times 100$).

*I chose a box plot for this sample because it had a minimal number of sample means (only 10) which makes the histogram not a great visual for the distribution as the frequency of the means was mostly a max of 1.*

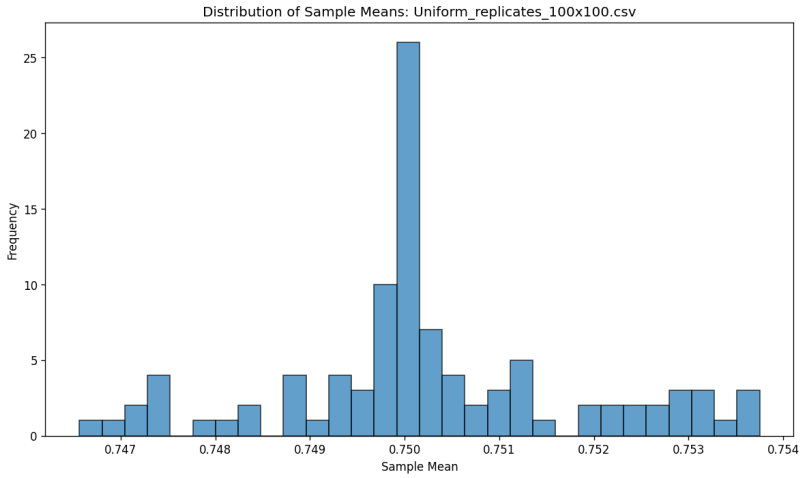


Figure 14: Uniform: Histogram of sample means ($100 \times 100$).

*I chose a histogram for this graph as it visually displayed the distribution very easily.*

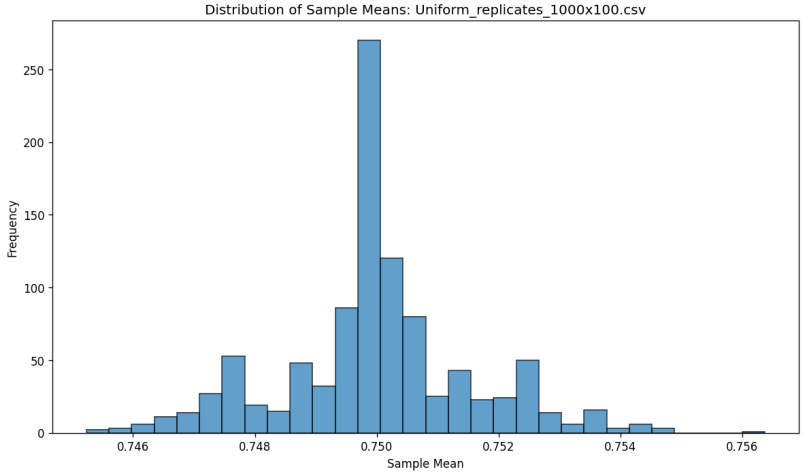


Figure 15: Uniform: Histogram of sample means ($1000 \times 100$).

*With an increased concentration, I updated the bin size a little but kept it as a histogram to display the distribution easily.*

**Power-law — Sample Mean Distributions**

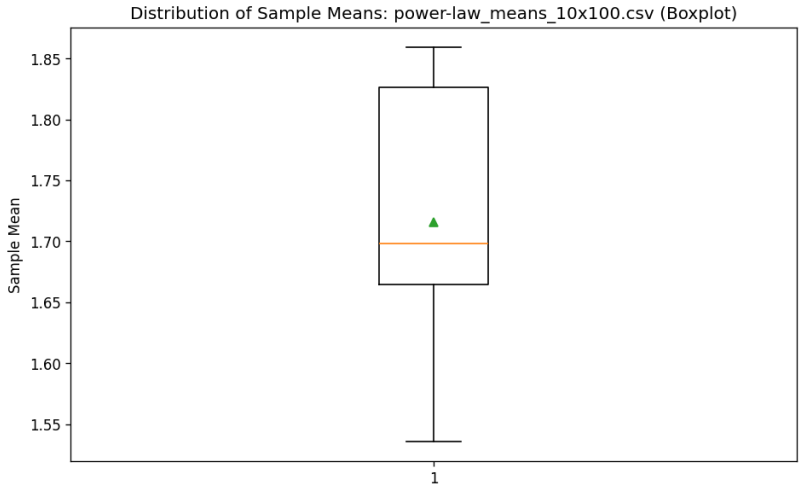


Figure 16: Power-law: Boxplot of sample means ($10 \times 100$).

*Similar to the uniform sample mean distribution, I decided to use a box and whisker plot to show the distribution as there were not enough samples to generate a histogram that would be appropriate. It also shows the outliers and skew pretty easily.:*

Figure 17: Power-law: Histogram of sample means ($100 \times 100$).

*For the 100 sample replicate, I decided to plot a histogram as there were a sufficient amount of samples to accurately visualize the distribution.*

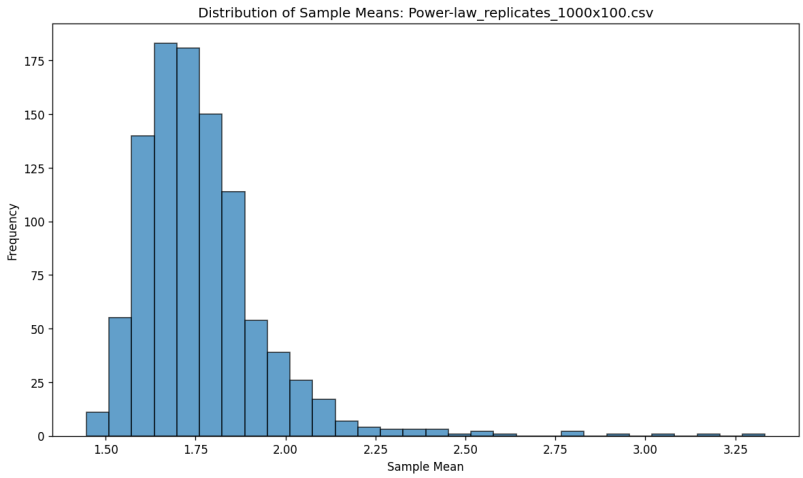


Figure 18: Power-law: Histogram of sample means ($1000 \times 100$).

*For the 1000 sample distribution plot, I created a histogram as well as you can see the clustering towards certain values as well as any significant outliers.*

## Part C: Normal Approximation via CLT

Approximate each sample-mean distribution with $\mathcal{N}(\hat{\mu}, \hat{\sigma})$ and overlay PDFs.

## Uniform — CLT Overlays

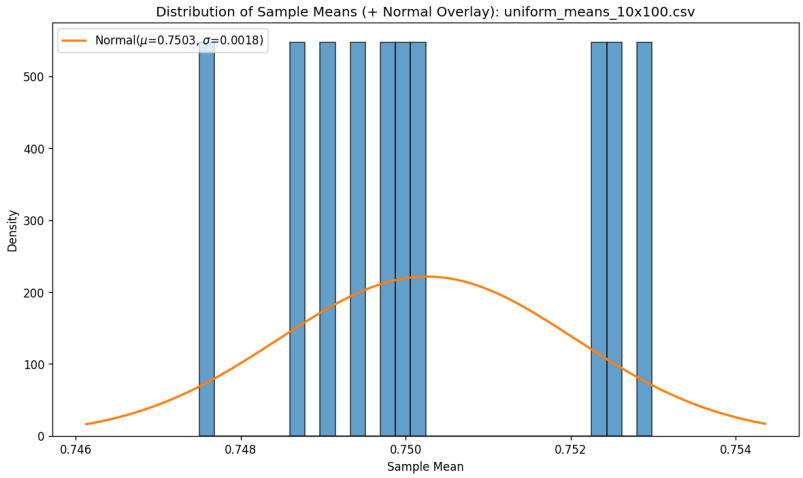


Figure 19: Uniform: Histogram + Normal overlay ($10 \times 100$).

$\hat{\mu} = 0.750258, \quad \hat{\sigma} = 0.001802.$

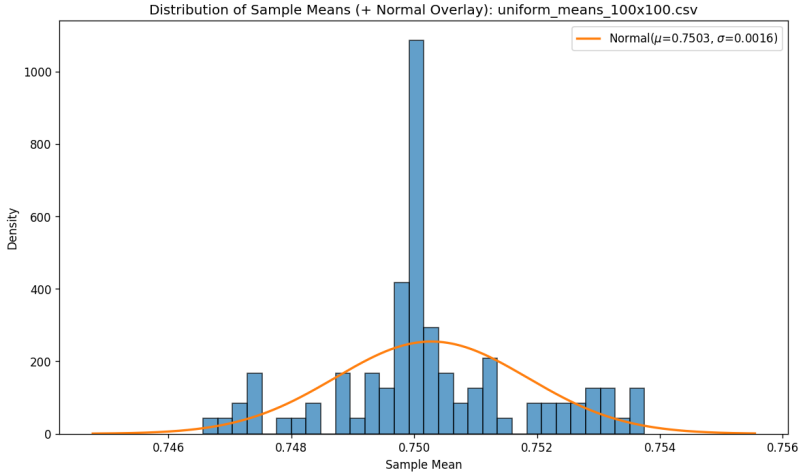


Figure 20: Uniform: Histogram + Normal overlay ($100 \times 100$).

$\hat{\mu} = 0.750281, \quad \hat{\sigma} = 0.001567.$

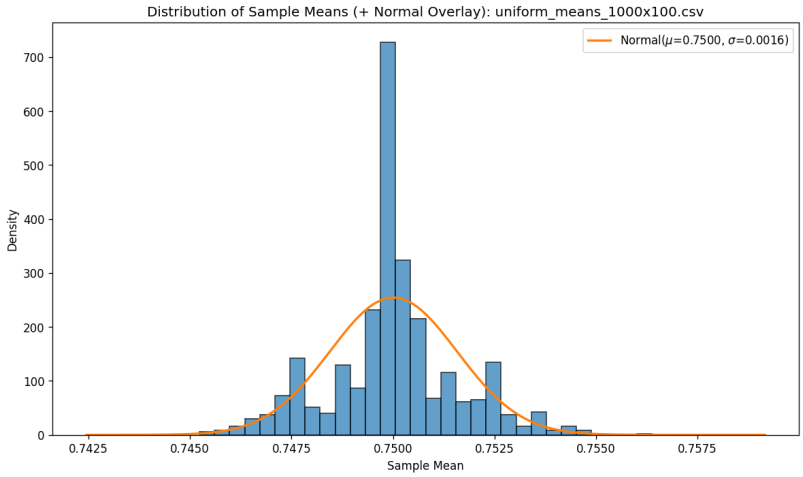


Figure 21: Uniform: Histogram + Normal overlay ($1000 \times 100$).

$\hat{\mu} = 0.750009, \quad \hat{\sigma} = 0.001569.$

## Power-law — CLT Overlays

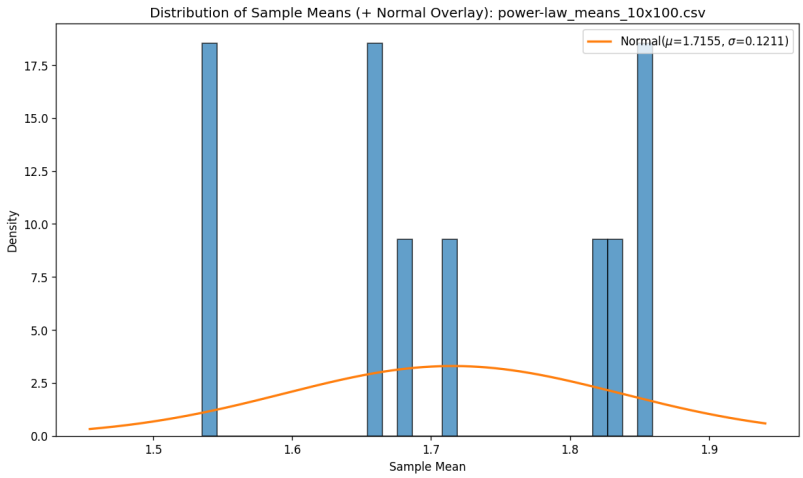


Figure 22: Power-law: Histogram + Normal overlay ($10 \times 100$).

$\hat{\mu} = 1.715484, \quad \hat{\sigma} = 0.121066.$

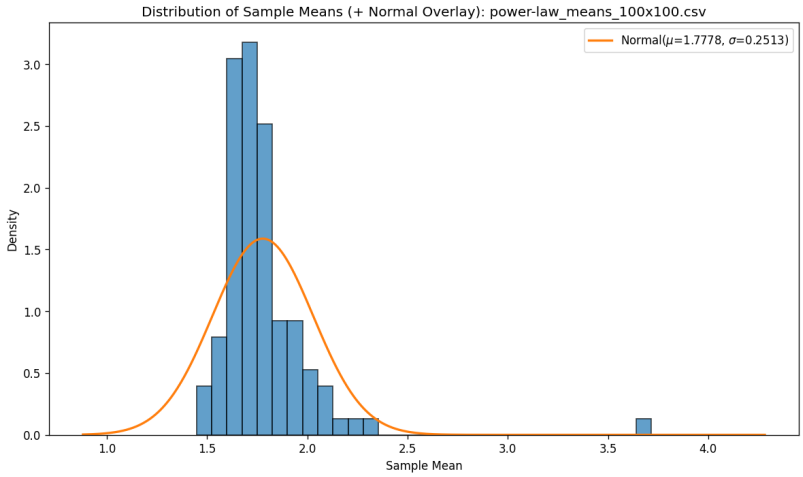


Figure 23: Power-law: Histogram + Normal overlay ($100 \times 100$).

$\hat{\mu} = 1.777796, \quad \hat{\sigma} = 0.251344.$

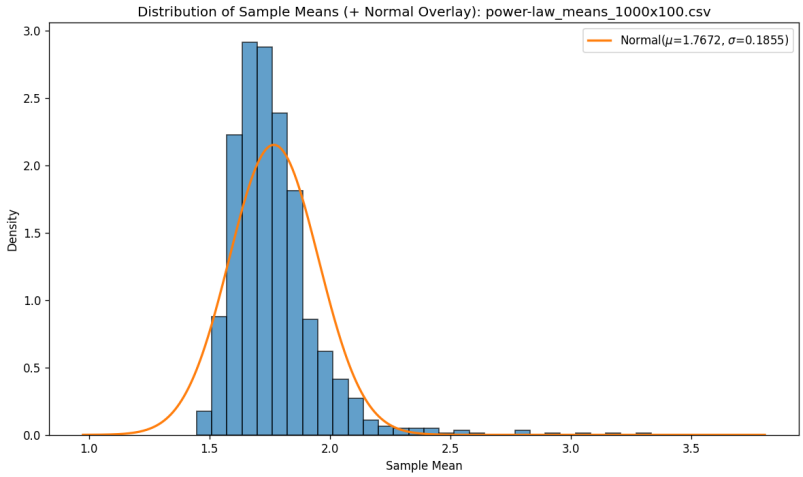


Figure 24: Power-law: Histogram + Normal overlay ($1000 \times 100$).

$\hat{\mu} = 1.767245, \quad \hat{\sigma} = 0.185465.$

**Discussion:** For the uniform distributions, the 100 x 100 and 1000 x 100 histograms appear to resemble normal distributions. The 10 x 100 does not, likely due to insufficient number of samples. None of the power law distributions appear to follow normal distributions very closely as they tend to have a wide range of values greater than the average, but only a small range of values less than the average leading to heavily tailed distributions.

## Part D: Distribution Comparison with KS-Test

Compare the Uniform vs Power-law sample-mean distributions using the two-sample KS test.

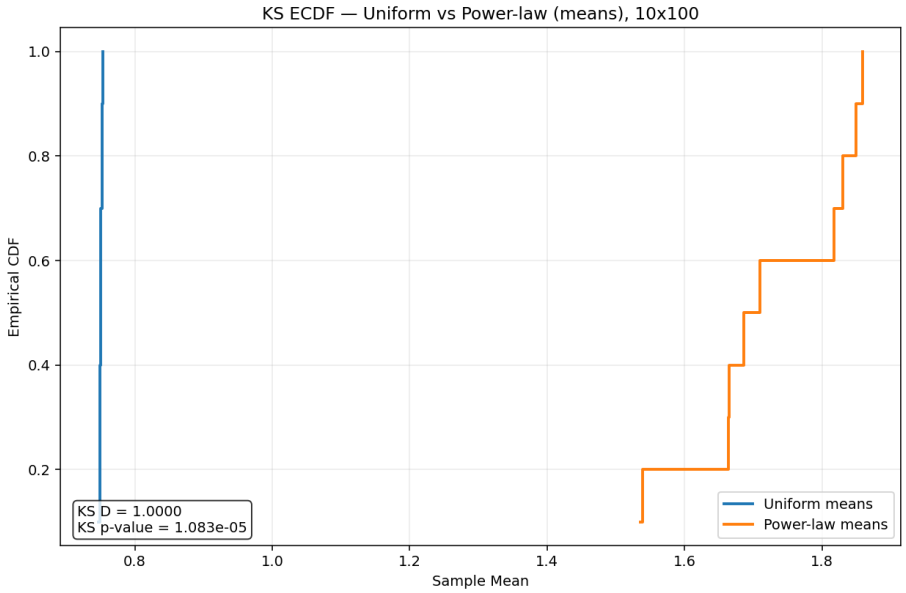


Figure 25: KS ECDF comparison (Uniform vs Power-law), $10 \times 100$.

$D = 1.0000, \quad p\text{-}value = 1.0825 \times 10^{-5}, \quad Reject\ @\ 0.05?\ \mathbf{Yes}.$

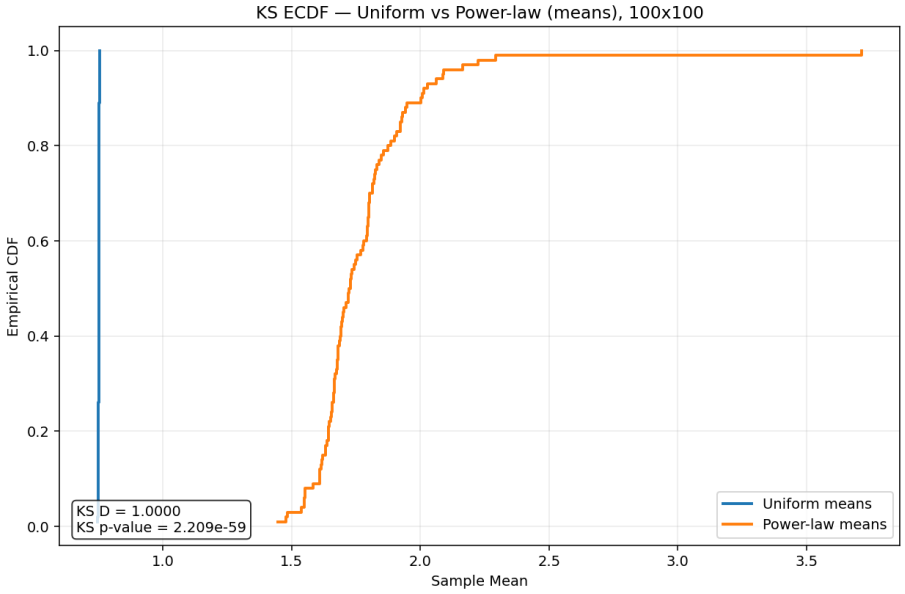


Figure 26: KS ECDF comparison (Uniform vs Power-law), $100 \times 100$.

$D = 1.0000, \quad p\text{-}value = 2.2088 \times 10^{-59}, \quad Reject\ @\ 0.05?\ \mathbf{Yes}.$

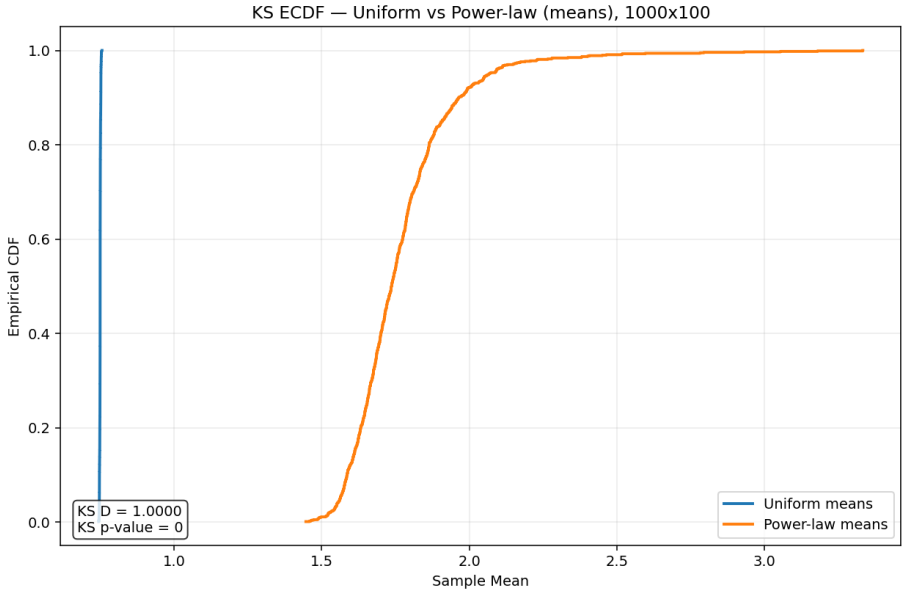


Figure 27: KS ECDF comparison (Uniform vs Power-law), $1000 \times 100$.

$D = 1.0000$, *p-value* $\approx 0$ *(underflow),* *Reject @ 0.05?* ***Yes***.

**KS Summary and Interpretation:**
Interpretation & CLT alignment: The CLT describes *each* distribution of sample means tending toward normality as the per-replicate sample size grows (here, $n = 100$ fixed). Our empirical overlays use $\mathcal{N}(\hat{\mu}, \hat{\sigma})$ and show good Gaussian fits within each distribution for the uniform tests with sufficient samples and ok (skewed slightly) fits for the power law means. The KS test compares the *two* mean distributions to each other. Because the power-law means are centered near $\sim 1.76$ (heavy-tailed Pareto) while uniform means are near 0.75, their ECDFs remain widely separated, yielding $D \approx 1$ and tiny $p$-values. As the number of means (replicates) increases from $10 \to 100 \to 1000$, KS power rises, so evidence against equality strengthens, consistent with theory.

# Task 3: Hypothesis Testing

**(a) Sample sizes.** US cars: $n_{\text{US}} = 248$ Japanese cars: $n_{\text{JP}} = 78$

**(b) Hypotheses.** Sales claim: Japanese mileage > US mileage.

$$H_0 : \ \mu_{\text{JP}} \leq \mu_{\text{US}} \qquad \text{vs} \qquad H_1 : \ \mu_{\text{JP}} > \mu_{\text{US}}.$$

**(c) Test statistic and reference distribution.** Welch's two-sample $t$-test (one-sided), which uses an approximately $t$ distribution with Satterthwaite (Welch) degrees of freedom.
Descriptive statistics:

$\bar{x}_{\text{US}} = 20.1532, \ s_{\text{US}} = 6.42622, \ n_{\text{US}} = 248; \qquad \bar{x}_{\text{JP}} = 30.5641, \ s_{\text{JP}} = 6.10214, \ n_{\text{JP}} = 78.$

Normality checks (Shapiro–Wilk):

$$W_{\text{US}} = 0.9380 \ (p = 9.96 \times 10^{-9}), \qquad W_{\text{JP}} = 0.9769 \ (p = 0.1685).$$

(Notes: Welch's test does not assume equal variances and is reasonably robust for large samples via the CLT.)

**(d) Result.** Welch $t$-statistic = 12.9741, df $\approx 134.966$, one-sided $p$-value $= 8.52572 \times 10^{-26}$.
Mean difference: $\bar{x}_{\text{JP}} - \bar{x}_{\text{US}} = 10.4109$.

**(e) Conclusion.** At $\alpha = 0.05$, reject $H_0$. The data provide strong statistical evidence that Japanese cars have higher average mileage than US cars in this sample.

Assumptions & remarks. Samples are treated as independent. While US normality is rejected by Shapiro–Wilk, the large sample size and use of Welch's test (which is robust to unequal variances) supports the inference.