# Analysis of Simulation Output

- Estimation Methods
- Simulation Run Statstics
- Replication of Runs
- Elimination of Initial Bias

# Estimation Methods

- **Estimates the range for the random variable so that the desired output can be achieved.**
  - In estimation, we aim to find a range where the random variable likely lies, based on the desired outcome. For example:

    Example:
  - Imagine a factory producing light bulbs. The lifetime of a light bulb is a random variable. Based on past data, we estimate that the lifetime of most bulbs is between 800 and 1200 hours. This range helps the factory maintain quality control and customer satisfaction.

# Estimation Methods....cont

- **Infinite population has a stationary probability distribution with a finite mean μ and finite variance $\sigma^2$.**

  This assumes that the population of data is so large (infinite) that its statistical properties (mean and variance) do not change over time.

  Example:

  Consider tossing a fair coin. If we keep tossing it, the proportion of heads and tails approaches 50% each, no matter how many tosses. The mean (expected value) and variance of outcomes remain constant.

  - Mean (μ): Probability of heads or tails = 0.5.
  - Variance ($\sigma^2$): A fixed value based on probabilities.

# Estimation Methods....cont

- **Sample variable and time do not affect population distribution :**

The sample we take and the time we measure it do not change the overall characteristics of the population.

Example:

The heights of adult males in a city can be represented by a population distribution. If you randomly sample people at different times of the day, the mean height of your sample will still be close to the mean height of the entire population, assuming the sampling is random.

# Estimation Methods....cont

- **Variables that meet all these conditions are called independently and identically distributed (i.i.d.).**
  - For variables to be considered i.i.d., they must be:
    1. Independent: The outcome of one does not influence another.
    2. Identically distributed: All variables have the same probability distribution.

  Example:
  - Rolling a fair six-sided die multiple times:
    - Each roll is independent: The outcome of one roll does not affect the next.
    - Each roll is identically distributed: Every roll has the same probabilities (1/6 for each number).

# Estimation Methods....cont

- **Central limit theorem must be invoked to rely upon normal distribution of infinite population.**

  The Central Limit Theorem (CLT) states that when we take a large number of samples (of sufficient size) from any population, the distribution of their means will approximate a normal distribution, even if the population itself is not normally distributed. This property allows us to use the normal distribution for estimation.

  Example:

  - Imagine you are measuring the weights of apples in a large orchard.
  - The population distribution of apple weights might not be normal (it could be skewed if most apples are small but a few are very large).
  - However, if you randomly select a large number of samples (say, 30 apples each) and calculate the mean weight for each sample, the distribution of these sample means will approximate a normal distribution, regardless of the original distribution of weights.

# Estimation Methods....cont

- **Only then we can apply estimation method to that variable taken from infinite population.**
- **A random variable is drawn from an infinite population that has a stationary probability distribution with a finite mean ($\mu$) and variance ($\sigma^2$).**
- **Random variables that meet all these conditions are said to be independently and identically distributed (i.i.d.).**
- **For i.i.d. variables, the central limit theorem can be applied.**
  - The Central Limit Theorem (CLT) states that if you add up or take the average of a large number of i.i.d. variables, their sum or average will approximate a normal distribution, even if the original variables are not normally distributed.

# Estimation Methods....cont

- Let $X_i$ $(i = 1, 2, ..., n)$ be the $n$ i.i.d. random variables. Then normal variate:

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$$

- $\sum_{i=1}^{n} x_i$: This is the sum of the $n$ random variables $(x_1, x_2, ..., x_n)$.
- $n\mu$: Represents the expected total sum of the $n$ variables if the mean of the population ($\mu$ ) is known.
- $\sqrt{n}\sigma$: This is the standard deviation of the total sum of $n$ variables. Since the standard deviation of each variable is $\sigma$, the sum's standard deviation scales as $\sqrt{n}\sigma$.
- $Z$: The standard score (z-score), which measures how far the observed total is from the expected total in units of standard deviation.

# Estimation Methods....cont

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- $\bar{x}$: The sample mean, which is the average of the $n$ random variables.
- $\mu$: The population mean.
- $\sigma / \sqrt{n}$: The standard deviation of the sample mean. It is smaller than the standard deviation of the population because averaging reduces variability.
- $Z$: The z-score for the sample mean, indicating how far the observed sample mean is from the population mean in units of standard deviation.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

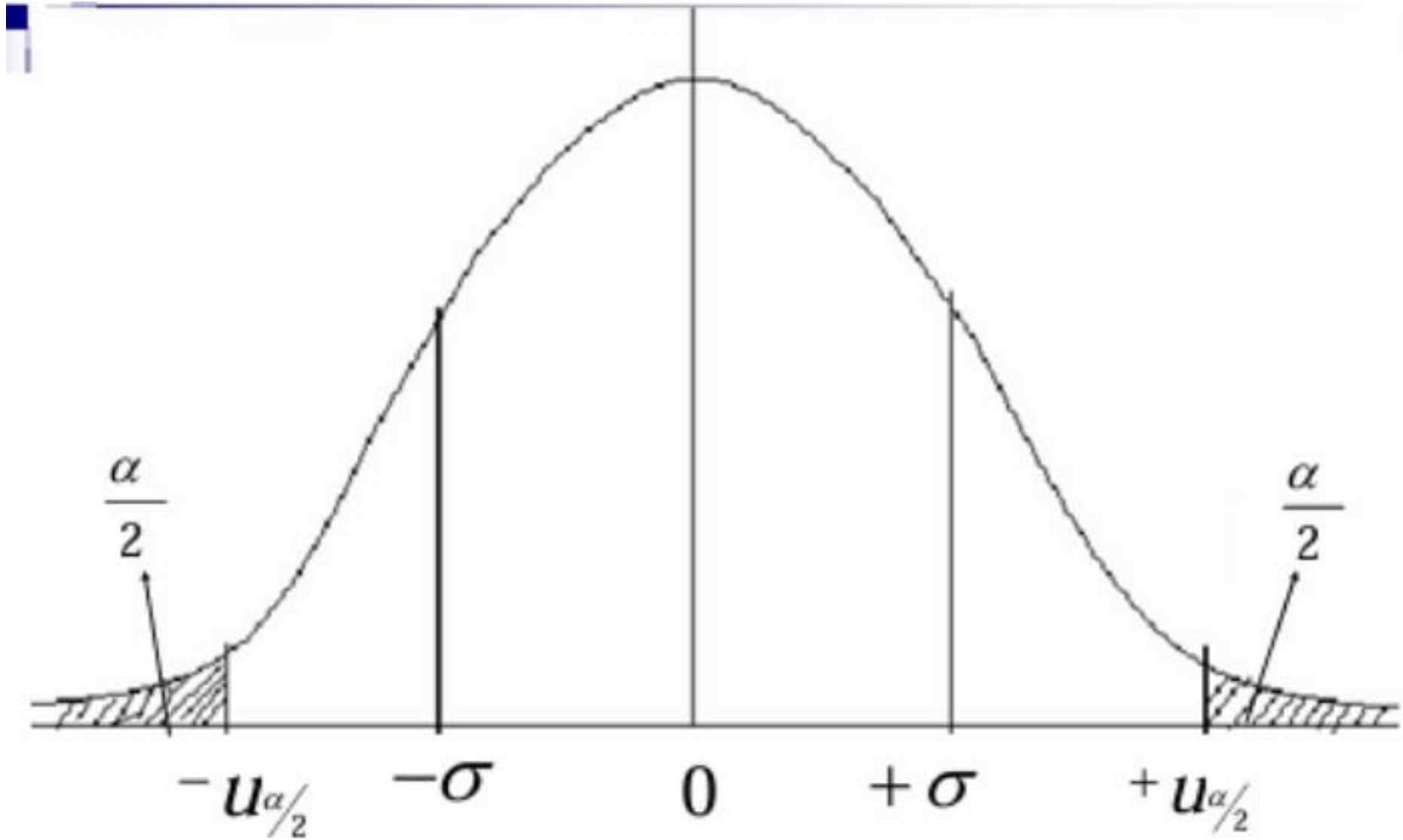


**Fig:Probability density function of standard normal variate**

# Estimation Methods

- The integral from -∞ to a value μ is the probability that z is less than or equal to μ. The integral is denoted by $\phi(u)$
- Suppose the value of μ is chosen so that $\phi(u) = 1 - \alpha / 2$ where α is some constant less than 1, and denote this value of u by $u_{\alpha/2}$.
- The normal distribution is symmetric about its mean, so the probability that z is less than $-u_{\alpha/2}$ is also α/2.

- The probability that $z$ lies between $-u_{\alpha/2}$ and $u_{\alpha/2}$ is $1-\alpha$.

  That is,

$$\text{Prob}\{-u_{\alpha/2} \leq z \leq u_{\alpha/2}\} = 1-\alpha$$

- In terms of sample mean, this probability statement can be written as:

$$\text{Prob}\left\{\bar{x} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2} \geq \mu \geq \bar{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\right\} = 1-\alpha$$

- The constant $1-\alpha$ is the confidence level, and the confidence interval is:

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$$

# Estimation Methods

- **Typically, the confidence level might be 90% in which case $u_{\alpha/2}$ is 1.65.**

- **The population variance σ ² is usually not known; in which case it is replaced by an estimate calculated from the formula**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- **The normalized random variable based on σ ² is replaced by a normalized random variable based on $s^2$.This has a Student-t distribution, with n-1 degrees of freedom.**

# Student-t Distribution

| | Confidence Probability | | | |
|---|---|---|---|---|
| $\nu$ | .80 | .90 | .96 | .98 |
| 1 | 3.078 | 6.314 | 15.895 | 31.821 |
| 2 | 1.886 | 2.920 | 4.849 | 6.965 |
| 3 | 1.638 | 2.353 | 3.482 | 4.541 |
| 4 | 1.533 | 2.132 | 2.999 | 3.747 |
| 5 | 1.476 | 2.015 | 2.757 | 3.365 |
| 6 | 1.440 | 1.943 | 2.612 | 3.143 |
| 7 | 1.415 | 1.895 | 2.517 | 2.998 |
| 8 | 1.397 | 1.860 | 2.449 | 2.896 |
| 9 | 1.383 | 1.833 | 2.398 | 2.821 |
| 10 | 1.372 | 1.812 | 2.359 | 2.764 |
| 11 | 1.363 | 1.796 | 2.328 | 2.718 |
| 12 | 1.356 | 1.782 | 2.303 | 2.681 |
| 13 | 1.350 | 1.771 | 2.282 | 2.650 |
| 14 | 1.345 | 1.761 | 2.264 | 2.624 |
| 15 | 1.341 | 1.753 | 2.249 | 2.602 |
| 16 | 1.337 | 1.746 | 2.235 | 2.583 |
| 17 | 1.333 | 1.740 | 2.224 | 2.567 |
| 18 | 1.330 | 1.734 | 2.214 | 2.552 |
| 19 | 1.328 | 1.729 | 2.205 | 2.539 |
| 20 | 1.325 | 1.725 | 2.197 | 2.528 |
| 25 | 1.316 | 1.708 | 2.167 | 2.485 |
| 30 | 1.310 | 1.697 | 2.147 | 2.457 |
| 40 | 1.303 | 1.684 | 2.123 | 2.423 |
| 50 | 1.299 | 1.676 | 2.109 | 2.403 |
| 75 | 1.293 | 1.665 | 2.090 | 2.377 |
| 100 | 1.290 | 1.660 | 2.081 | 2.364 |
| $\infty$ | 1.282 | 1.645 | 2.054 | 2.326 |

# Estimation Methods

- The quantity $u_{\alpha/2}$ used in the definition of a confidence interval given above, is replaced by a similar quantity, $t_{n-1,\alpha/2}$ based on the Student-t distribution.

  (which tables are also readily available)

- The Student-t distribution is strictly accurate only when the population from which the samples are drawn is normally distributed.

- Expressed in terms of the estimated variance $s^2$, the confidence interval for $\chi$ is defined by

$$\overline{x} \pm \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}$$

# CONCLUSION

- **Hence the estimation method gives the desired range of the sample variable taken from infinite population.**

# Simulation Run Statistics

- Consider a single-server system in which the arrivals occur with a Poisson distribution and the service time has an exponential distribution.
- Suppose the study objective is to measure the mean waiting time, defined as the time entities spend waiting to receive service and excluding the service time itself.
- This system is commonly denoted by M/M/1 which indicates; first , that the inter-arrival time is distributed exponentially; second that the service time is distributed exponentially; and, third, that there is one server. The M stands for Markovian, which implies an exponential distribution.

# Simulation Run Statistics

- In a simulation run, the simplest approach is to estimate the mean waiting time by accumulating the waiting time of n successive entities and dividing by n.
- This measure, the sample mean, is denoted by $\bar{x}(n)$ to emphasize the fact that its value depends upon the number of observations taken.
- If $x_i$ (i=1,2,….,n) are the individual waiting times (including the value 0 for those entities that do not have to wait), then

$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

# Simulation Run Statistics

- **Whenever a waiting line forms, the waiting time of each entity on the line clearly depends upon the waiting time of its predecessors.**

- **Any series of data that has this property of having one value affect other values is said to be autocorrelated.**

- **The sample mean of autocorrelated data can be shown to approximate a normal distribution as the sample size increases.**

# Simulation Run Statistics

- **The equation** $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ **remains a satisfactory estimate for the mean of autocorrelated data.**

- **A simulation run is started with the system in some initial state, frequently the idle state, in which no service is being given and no entities are waiting.**

- **The early arrivals then have a more than normal probability of obtaining service quickly, so a sample mean that includes the early arrivals will be biased.**

# Simulation Run Statistics

- For a given sample size starting from a given initial condition, the sample mean distribution is stationary; but , if the distributions could be compared for different sample sizes, the distribution would be slightly different.

- The following figure is based on theoretical results, which shows how the expected value of sample mean depends upon the sample length, for the M/M/1 system, starting from an initial empty state, with a server utilization of 0.9.

# Simulation Run Statistics

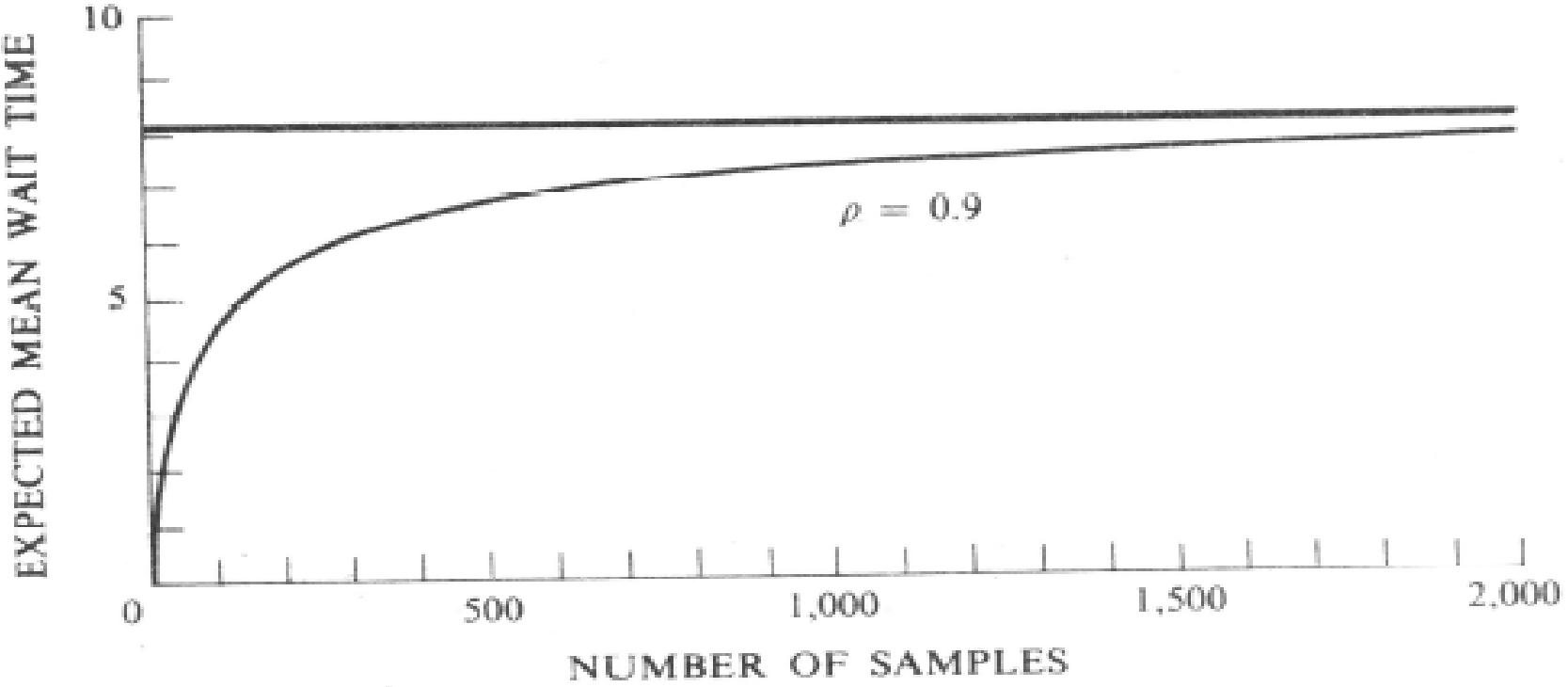


**Figure 14-2.** Mean wait time in M/M/1 system for different sample sizes.

# Replications of Runs

- The precision of results of a dynamic stochastic can be increased by repeating the experiment with different random numbers strings.

- For each replication of a small sample size, the sample mean is determined.

- The sample means of the independent runs can be further used to estimate the variance of distribution. Let $X_{ij}$ be the $i^{th}$ observation in $j^{th}$ run, then the sample mean and variance for the $j^{th}$ run are:

# Replications of Runs

$$\overline{x}_j(n) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} [x_{ij} - \overline{x}_j(n)]^2$$

# Replications of Runs

- **When we have similar means and variances for m independent measurements, then by combining them, the mean and variance for the population can be obtained as:**

# Replications of Runs

$$\bar{\bar{x}} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_j(n)$$

$$S^2 = \frac{1}{p}\sum_{j=1}^{p} s_j^2(n)$$

# Replications of Runs

- **The following figure shows the result of applying the procedure to experimental results for the M/M/1 system.**

# Replications of Runs

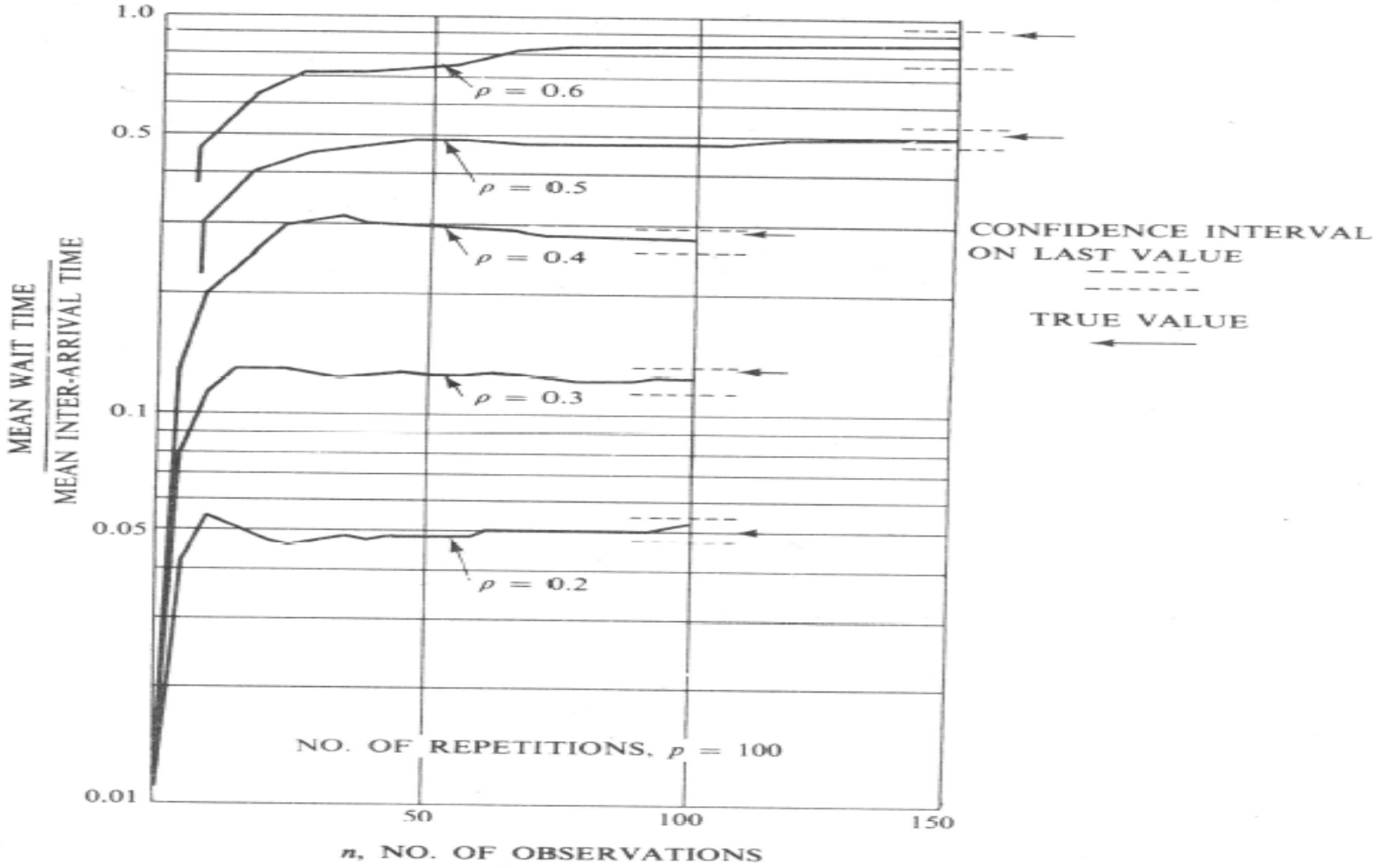


**Figure 14-3.** Experimentally measured wait time in M/M/1 system for different sample sizes.

# Replications of Runs

- This variance can further be used to establish the confidence interval for p-1 degrees of freedom.
- The length of run of replications is so selected that all combined it comes to the sample size N. i.e. p.n=N.
- By increasing the number of replications and shortening their length of run, the confidence interval can be narrowed.
- But due to shortening of length of replication the effect of starting conditions will increase.
- The results obtained will not be accurate, especially when the initiazation of the runs is not proper.
- Thus, a compromise has to be made.
- There is no established procedure of dividing the sample size N into replications.
- However, it is suggested that the number of replications should not be very large, and that the sample means should approximate a normal distribution.

# Elimination of Initial Bias

- Two general approaches can be taken to remove the bias: the system can be started in a more representative state that the empty state, or the first part of the simulation can be ignored.
- The ideal situation is to know the steady state distribution for the system, and select the initial condition from that distribution.
- In the study previously discussed, repeated the experiments on the M/M/1 system, supplying an initial waiting line for each run, selected at random from the known steady state distribution of waiting line.

# Elimination of Initial Bias

- The case of 40 repetitions of 320 samples, which previously resulted in a coverage of only 9% was improved to coverage of 88%.
- The more common approach to removing the initial bias is to eliminate an initial section of the run.
- The run is started from an idle state and stopped after a certain period of time.

# Elimination of Initial Bias

- The run is then restarted with statistics being gathered from the point of restart.
- It is usual to program the simulation so that statistics are gathered from the beginning, and simply wipe out the statistics gathered up to the point of restart.
- No simple rules can be given to decide how long an interval should be eliminated.

# Elimination of Initial Bias

- **The disadvantage of eliminating the first part of a simulation run is that the estimate of the variance, needed to establish a confidence limit, must be based on less information.**

- **The reduction in bias, therefore, is obtained at the price of increasing the confidence interval size.**

# Reference

- Geoffrey Gordon, System Simulation, Chapter 14, analysis of simulation output