S. Gopalakrishnan

- **Population** – Large collection of similar items which have some measurable property / value associated with them.

- **Population** – Large collection of similar items which have some measurable property / value associated with them.
- **Sample** – A small subset (sampling) of the population.

- **Population** – Large collection of similar items which have some measurable property / value associated with them.
- **Sample** – A small subset (sampling) of the population.
- Sample data is used to make inferences about the whole population. In order to do this one has to make some assumptions about the relationship between the sample and the population.

- **Population** – Large collection of similar items which have some measurable property / value associated with them.
- **Sample** – A small subset (sampling) of the population.
- Sample data is used to make inferences about the whole population. In order to do this one has to make some assumptions about the relationship between the sample and the population.
- **Assumption** – The population has a probability distribution such that the measurable values of the items in the population can be thought of as being independent random variables having this distribution.
- If the sample data are then chosen in a random fashion, then it is reasonable to suppose that they too are independent values from the distribution.

- **Population** – Large collection of similar items which have some measurable property / value associated with them.
- **Sample** – A small subset (sampling) of the population.
- Sample data is used to make inferences about the whole population. In order to do this one has to make some assumptions about the relationship between the sample and the population.
- **Assumption** – The population has a probability distribution such that the measurable values of the items in the population can be thought of as being independent random variables having this distribution.
- If the sample data are then chosen in a random fashion, then it is reasonable to suppose that they too are independent values from the distribution.
- **Definition** – If $X_1, \ldots, X_n$ are independent random variables having a common distribution F, then we say that they constitute a sample (sometimes called a random sample) from the distribution F.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.
- If the distribution F is specified up to some unknown parameters, For e.g.,
  - F is a normal distribution with unknown mean and variance.
  - F is a Poisson distribution with unknown mean.

  Then problems on such nature are called *parametric inference* problems.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.
- If the distribution F is specified up to some unknown parameters, For e.g.,
  - F is a normal distribution with unknown mean and variance.
  - F is a Poisson distribution with unknown mean.

  Then problems on such nature are called *parametric inference* problems.
- If nothing is known about F then the problems are called *nonparametric inference* problems.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.
- If the distribution F is specified up to some unknown parameters, For e.g.,
  - F is a normal distribution with unknown mean and variance.
  - F is a Poisson distribution with unknown mean.

  Then problems on such nature are called *parametric inference* problems.
- If nothing is known about F then the problems are called *nonparametric inference* problems.
- **Statistic** is a random variable whose value is determined by the sample data.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.
- If the distribution F is specified up to some unknown parameters, For e.g.,
  - F is a normal distribution with unknown mean and variance.
  - F is a Poisson distribution with unknown mean.

  Then problems on such nature are called *parametric inference* problems.
- If nothing is known about F then the problems are called *nonparametric inference* problems.
- **Statistic** is a random variable whose value is determined by the sample data.
- **Sample mean** and **Sample variance** are two most commonly used statistics.

- In statistics, generally complete knowledge of the underlying population distribution function F is not available.
- One will use the sample data to make inferences about the the distribution F.
- If the distribution F is specified up to some unknown parameters, For e.g.,
  - F is a normal distribution with unknown mean and variance.
  - F is a Poisson distribution with unknown mean.

  Then problems on such nature are called *parametric inference* problems.
- If nothing is known about F then the problems are called *nonparametric inference* problems.
- **Statistic** is a random variable whose value is determined by the sample data.
- **Sample mean** and **Sample variance** are two most commonly used statistics.
- The quantities μ and σ^2 are called the **population mean** and **population variance** respectively.

Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Since the value of the sample mean $\bar{X}$ is determined by the values of the random variables in the sample, it follows that $\bar{X}$ is also a random variable. Its expected value and variance are obtained as follows:

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n}(E[X_1] + \ldots E[X_n]) \\ &= \mu \end{aligned}$$

Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

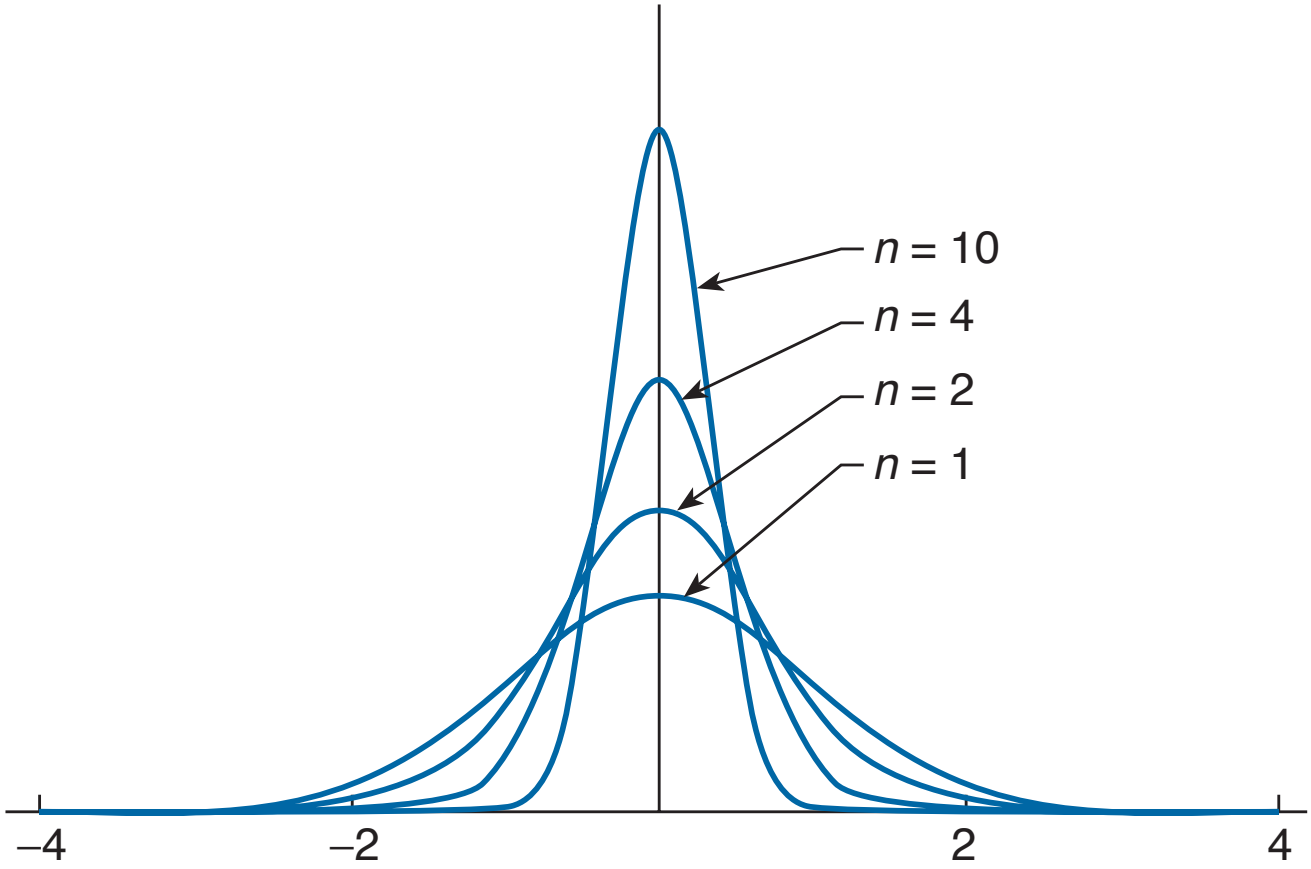
$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Since the value of the sample mean $\bar{X}$ is determined by the values of the random variables in the sample, it follows that $\bar{X}$ is also a random variable. Its expected value and variance are obtained as follows:

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n}(E[X_1] + \ldots E[X_n]) \\ &= \mu \end{aligned}$$

And

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{X_1 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}[Var(X_1) + \cdots + Var(X_n)] \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$



Densities of sample means from a standard normal population

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with a mean $n\mu$ and variance $n\sigma^2$.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with a mean $n\mu$ and variance $n\sigma^2$. From Central Limit theorem, it follows,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable.

Example

An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

Example

An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

Solution: Let X denote the total yearly claim. Number the policy holders, and let X_i denote yearly claim of policy holder i.
From central limit theorem , $X = \sum_{i=1}^{n} X_i$ where $n = 25,000$. Therefore,

$$n\mu = 320 \times 25000 = 8 \times 10^6 \qquad \sigma\sqrt{n} = 540\sqrt{25000} = 85381$$

Therefore

$$P\{X > 8.3 \times 10^6\} = P\left\{\frac{X - 8 \times 10^6}{85381} > \frac{8.3 \times 10^6 - 8 \times 10^6}{85381}\right\}$$
$$= P\{Z > 3.51\} \approx 0.00023$$

Thus, there are only 2.3 chances out of 10,000 that the total yearly claim will exceed 8.3 million

Example

Civil engineers believe that W, the amount of weight (in units of 1,000 pounds) that a certain span of a bridge can withstand without structural damage resulting, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (again, in units of 1,000 pounds) of a car is a random variable with mean 3 and standard deviation .3. How many cars would have to be on the bridge span for the probability of structural damage to exceed .1?

Example

Civil engineers believe that W, the amount of weight (in units of 1,000 pounds) that a certain span of a bridge can withstand without structural damage resulting, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (again, in units of 1,000 pounds) of a car is a random variable with mean 3 and standard deviation .3. How many cars would have to be on the bridge span for the probability of structural damage to exceed .1?

Solution: Let P_n denote the probability of structural damage when there are n cars on the bridge. That is,

$$\begin{aligned} P_n &= P\{X_1 + \cdots + X_n \geq W\} \\ &= P\{X_1 + \cdots + X_n - W \geq 0\} \end{aligned}$$

where X_i is the weight of the ith car, $i = 1, \ldots, n$. Now it follows from the central limit theorem that $\sum_{i=1}^{n} X_i$ is approximately normal with mean $3n$ and variance $.09n$. Hence, since W is independent of the $X_i, i = 1, \ldots, n$, and is also normal, it follows that $\sum_{i=1}^{n} X_i - W$ is approximately normal, with mean and variance given by.

$$E\left[\sum_{i=1}^{n} X_i - W\right] = 3n - 400$$

$$Var\left(\sum_{i=1}^{n} X_i - W\right) = Var\left(\sum_{i=1}^{n} X_i\right) + Var(W) = .09n + 1600$$

If,

$$Z = \frac{\sum_{i=1}^{n} X_i - W - (3n - 400)}{\sqrt{0.09n + 1600}}$$

then

$$P_n = P\left\{ Z \geq \frac{-(3n - 400)}{\sqrt{0.09n + 1600}} \right\}$$

where Z is approximately a standard normal random variable.

If,

$$Z = \frac{\sum_{i=1}^{n} X_i - W - (3n - 400)}{\sqrt{0.09n + 1600}}$$

then

$$P_n = P\left\{ Z \geq \frac{-(3n - 400)}{\sqrt{0.09n + 1600}} \right\}$$

where Z is approximately a standard normal random variable.
Now $P\{Z \geq 1.28\} \approx 0.1$, and so if the number of cars n is such that

$$\frac{-(3n - 400)}{\sqrt{0.09n + 1600}} \leq 1.28$$

or

$$n \geq 117$$

then there is at least 1 chance in 10 that structural damage will occur.

Application of Central Limit Theorem to Binomial Random Variables

A binomial random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, we can express it as

$$X = \frac{X_1 + \cdots + X_n}{n}$$

where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

Application of Central Limit Theorem to Binomial Random Variables

A binomial random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, we can express it as

$$X = \frac{X_1 + \cdots + X_n}{n}$$

where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

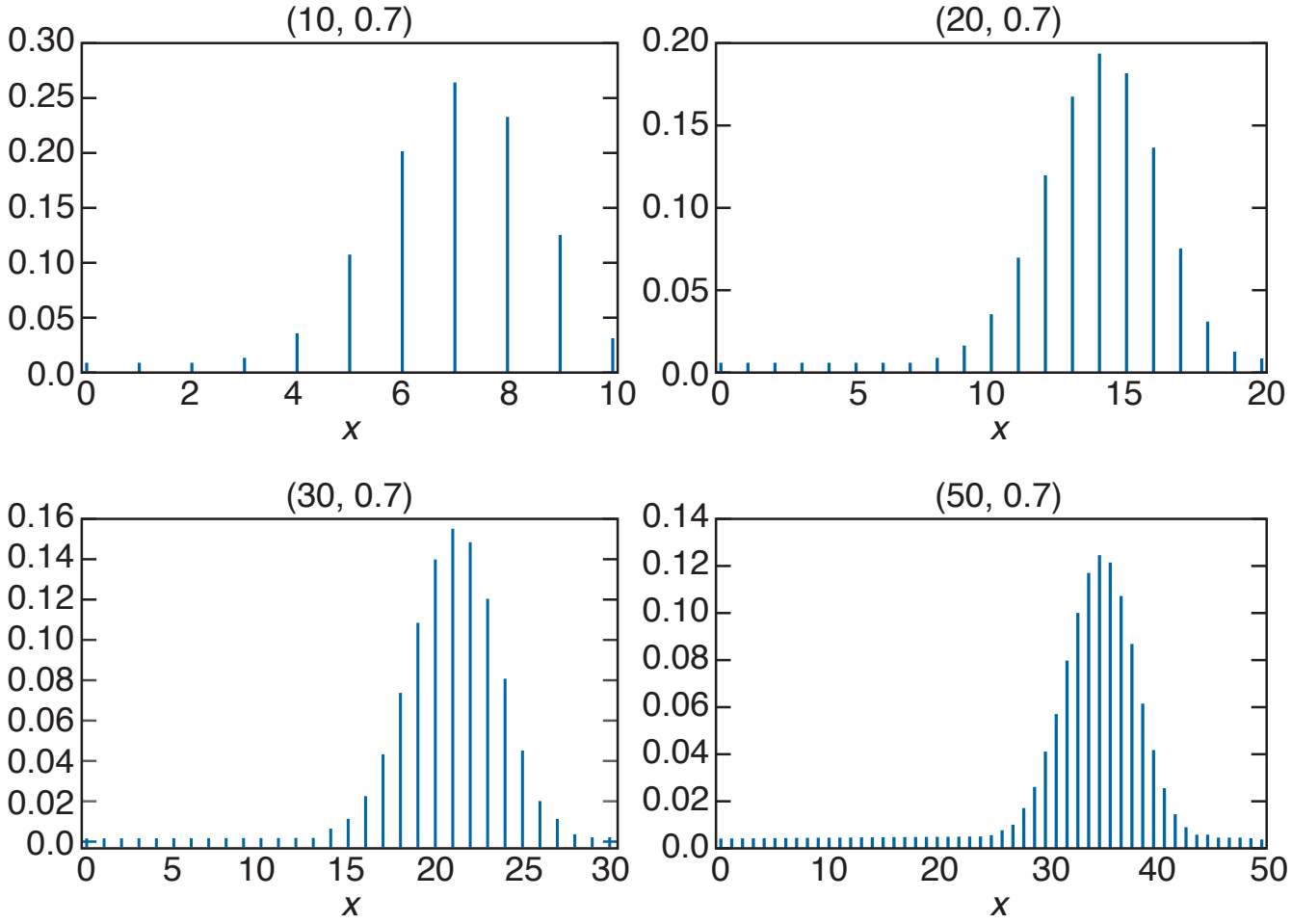
Since,

$$E[X_i] = p, \qquad Var(X_i) = p(1-p)$$

it follows from the central limit theorem that for n large

$$\frac{X - np}{\sqrt{np(1-p)}}$$

will approximately be a standard normal random variable



Binomial probability mass functions converging to the normal density.

Example

The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Example

The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Solution: Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random variable with parameters $n = 450$ and $p = .3$.

Example

The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Solution: Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random variable with parameters $n = 450$ and $p = .3$.
Since the binomial is a discrete and the normal a continuous distribution,it is best to compute $P\{X = i\}$ as $P\{i - 0.5 < X < i + 0.5\}$ when applying the normal approximation (this is called the continuity correction).

$$P\{X > 150.5\} = P\left\{\frac{X - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}} \geq \frac{150.5 - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}}\right\} \approx P\{Z > 1,59\} = 0.06$$

Hence, only 6 percent of the time do more than 150 of the first 450 accepted actually attend.

Note on Binomial Random Variable

It should be noted that we now have two possible approximations to binomial probabilities: The Poisson approximation, which yields a good approximation when n is large and p small, and the normal approximation, which can be shown to be quite good when $np(1-p)$ is large. [The normal approximation will, in general, be quite good for values of n satisfying $np(1-p) \geq 10$.]

Approximate Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2 . The central limit theorem can be used to approximate the distribution of the sample mean

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

Approximate Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2 . The central limit theorem can be used to approximate the distribution of the sample mean

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

Since a constant multiple of a normal random variable is also normal, it follows from the central limit theorem that $\bar{X}$ will be approximately normal when the sample size n is large. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has approximately a standard normal distribution.

Example

An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within ± 0.5 light years?

Example

An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within ± 0.5 light years?

Solution: If the astronomer makes n measurements, then $\bar{X}$, the sample mean of these measurements, will be approximately a normal random variable with mean d and standard deviation $2/\sqrt{n}$. Thus, the probability that it will lie between $d \pm 0.5$ is obtained as follows:

$$P\{-0.5 < \bar{X} - d < 0.5\} = P\left\{\frac{-0.5}{2/\sqrt{n}} < \frac{\bar{X} - d}{2/\sqrt{n}} < \frac{0.5}{2/\sqrt{n}}\right\}$$
$$\approx P\{-\sqrt{n}/4 < Z < \sqrt{n}/4\} = 2\{Z < \sqrt{n}/4\} - 1$$

where Z is a standard normal random variable.

Thus, the astronomer should make n measurements, where n is such that

$$2P\{Z < \sqrt{n}/4\} - 1 \geq 0.95$$

or

$$P\{Z < \sqrt{n}/4\} \geq 0.975$$

Since $P\{Z < 1.96\} = .975$, it follows that n should be chosen so that

$$\sqrt{n}/4 \geq 1.96$$

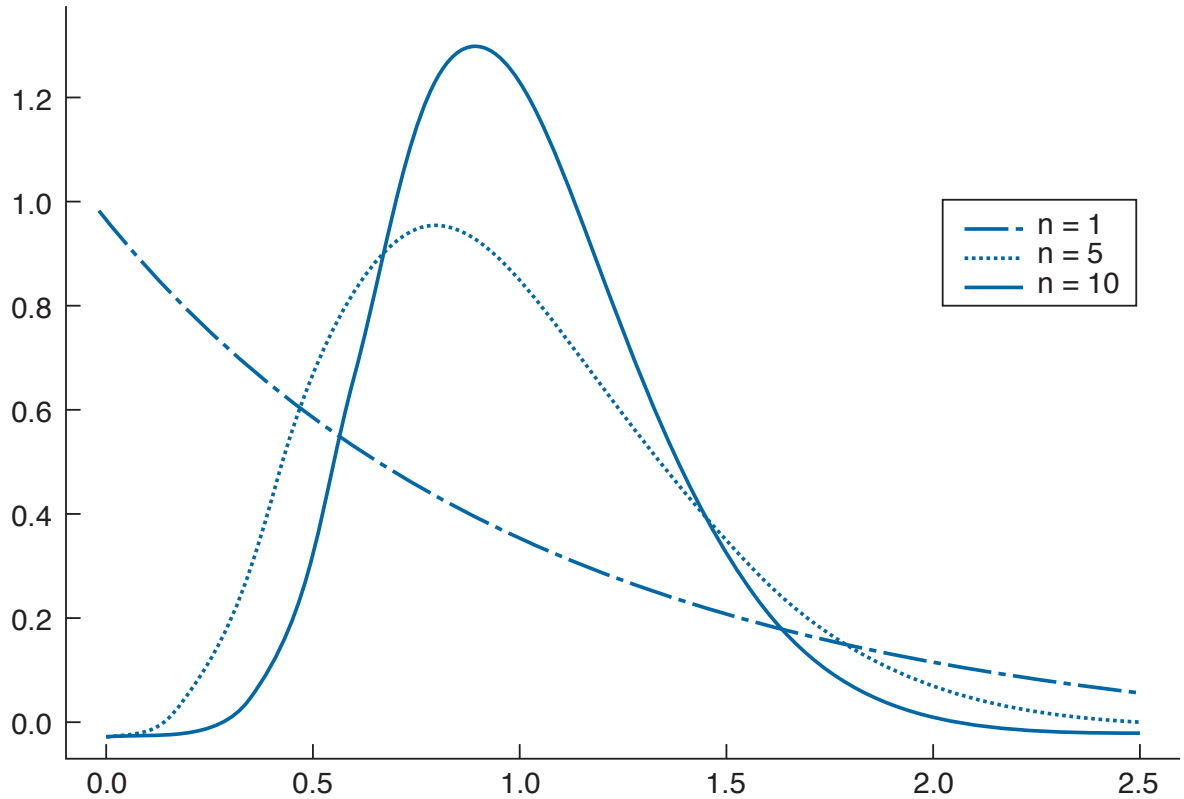
That is, at least 62 observations are necessary.

Minimum Sample size

If the underlying population distribution is normal, then the sample mean $\bar{X}$ will also be normal regardless of the sample size.

Minimum Sample size

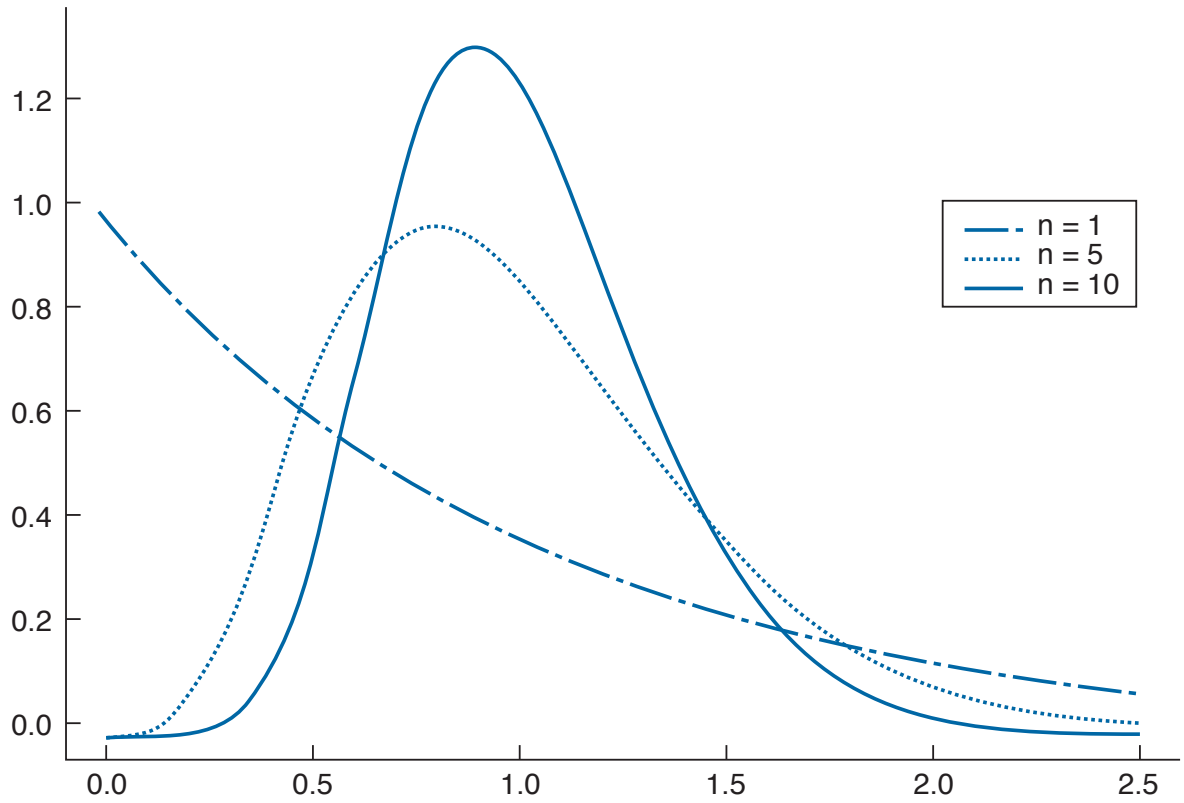
If the underlying population distribution is normal, then the sample mean $\bar{X}$ will also be normal regardless of the sample size.



Densities of the average of n exponential random variables having mean 1

Minimum Sample size

If the underlying population distribution is normal, then the sample mean $\bar{X}$ will also be normal regardless of the sample size.



Densities of the average of n exponential random variables having mean 1

A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30.

Sample Variance

Let $X_1, X_2, \ldots, X_n$ be a sample of values for a population with a mean μ and variance σ^2.
The statistic S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

is called the sample variance. $S = \sqrt{S^2}$ is called the sample standard deviation.

Sample Variance

Let $X_1, X_2, \ldots, X_n$ be a sample of values for a population with a mean μ and variance σ^2.
The statistic S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

is called the sample variance. $S = \sqrt{S^2}$ is called the sample standard deviation.
To compute $E[S^2]$, we use the identity

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad \text{where } \bar{x} = \frac{\sum_{i=1}^{n}}{n}$$

Therefore

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Taking expectations on both sides,

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\bar{X}^2] \\ &= nE[X_1^2] - nE[\bar{X}^2] \\ &= nVar(X_1) + n(E[X_1])^2 - nVar(\bar{X}) - n(E[\bar{X}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

or

$$E[S^2] = \sigma^2$$

That is, the expected value of the sample variance S^2 is equal to the population variance σ^2