Proability and Statistics

MA20205

Bibhas Adhikari

Autumn 2022-23, IIT Kharagpur

Lecture 16
October 31, 2022

Descriptive statistics

Review

Descriptive statistics

Review

1. We have studied two important functions associated with a random variable: pdf and cdf

Descriptive statistics

Review

1. We have studied two important functions associated with a random variable: pdf and cdf
2. These functions help us predicting the behavior of the random variable corresponding to a real life problem

Descriptive statistics

Review

1. We have studied two important functions associated with a random variable: pdf and cdf
2. These functions help us predicting the behavior of the random variable corresponding to a real life problem
3. There are many parameters, mainly the mean (μ) and variance σ^2 characterize the random variable

Descriptive statistics

Review

1. We have studied two important functions associated with a random variable: pdf and cdf
2. These functions help us predicting the behavior of the random variable corresponding to a real life problem
3. There are many parameters, mainly the mean (μ) and variance σ^2 characterize the random variable
4. In practice, however the pdf or cdf are not known and hence the parameters are also not known

Descriptive statistics

Review

1. We have studied two important functions associated with a random variable: pdf and cdf
2. These functions help us predicting the behavior of the random variable corresponding to a real life problem
3. There are many parameters, mainly the mean (μ) and variance σ^2 characterize the random variable
4. In practice, however the pdf or cdf are not known and hence the parameters are also not known
5. Goal: how to determine a reasonable pdf and approximate the values for the distribution parameters from a data set

Sampling distributions

Statistical problem Consider predicting life expectation of an object (such as bulbs, batteries etc.) produced by a company. The problem is then is to determine the mean effective life span of these objects so that a limited warranty period can be placed.

Sampling distributions

Statistical problem Consider predicting life expectation of an object (such as bulbs, batteries etc.) produced by a company. The problem is then is to determine the mean effective life span of these objects so that a limited warranty period can be placed.

Let X denote the random variable for the life span of a battery. Then the distribution of X can never be obtained precisely because there will be future objects that are not produced yet, and the existing number of object is quite large for a inspection.

Sampling distributions

Statistical problem Consider predicting life expectation of an object (such as bulbs, batteries etc.) produced by a company. The problem is then is to determine the mean effective life span of these objects so that a limited warranty period can be placed.

Let X denote the random variable for the life span of a battery. Then the distribution of X can never be obtained precisely because there will be future objects that are not produced yet, and the existing number of object is quite large for a inspection.

Population The (large) set of objects about which some inferences are to be made is called the population.

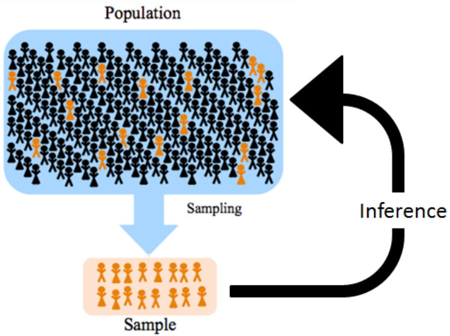
Sampling distributions

Statistical problem Consider predicting life expectation of an object (such as bulbs, batteries etc.) produced by a company. The problem is then is to determine the mean effective life span of these objects so that a limited warranty period can be placed.

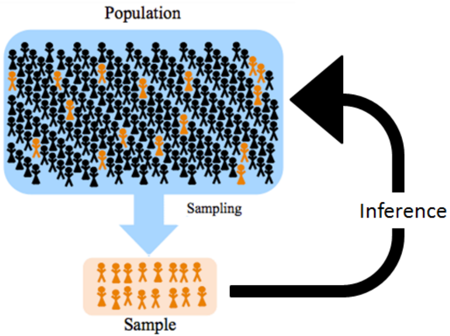
Let X denote the random variable for the life span of a battery. Then the distribution of X can never be obtained precisely because there will be future objects that are not produced yet, and the existing number of object is quite large for a inspection.

Population The (large) set of objects about which some inferences are to be made is called the population. there should be at least one random variable relative to the population whose behavior is to be studied. Since we can not inspect all the elements of the population, we have to select a sample of objects from the population.

Sampling distributions



Sampling distributions



Random sampling Select n objects from the population such that the selection of one object neither ensures nor precludes the selection of any other. Thus the selection of one object is independent of selection of any other. This set is called random sample.

Sampling distributions

Sampling with replacement

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.
2. Repeat the experiment n times

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.
2. Repeat the experiment n times
3. Let X_k denote the rv associated with kth repetition

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.
2. Repeat the experiment n times
3. Let X_k denote the rv associated with kth repetition
4. Then the rvs $\{X_1, X_2, \ldots, X_n\}$ is called a random sample of size n

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.
2. Repeat the experiment n times
3. Let X_k denote the rv associated with kth repetition
4. Then the rvs $\{X_1, X_2, \ldots, X_n\}$ is called a random sample of size n
5. Note that $X_1, \ldots, X_n$ are iids with density function $f(x)$

Sampling distributions

Sampling with replacement

1. Let X be a random variable associated with an experiment with pdf $f(x)$.
2. Repeat the experiment n times
3. Let X_k denote the rv associated with kth repetition
4. Then the rvs $\{X_1, X_2, \ldots, X_n\}$ is called a random sample of size n
5. Note that $X_1, \ldots, X_n$ are iids with density function $f(x)$

Random sample

A random of size n from the distribution of X is a collection of n independent random variables, each with the same distribution of X

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
2. Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
2. Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

The statistic $S = \sqrt{S^2}$ is called the sample standard deviation

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$
2. Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$

The statistic $S = \sqrt{S^2}$ is called the sample standard deviation

Question Why $1/(n-1)$ instead of $1/n$ in the definition of S^2?

Sampling distributions

For practical purposes, the sample constitutes 5% of the population.

Question What is a statistic?

A statistic is a random variable whose numerical value can be determined from a random sample $X_1, \ldots, X_n$. Some of the important statistics for statisticians are $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\max_i X_i$, $\min_i X_i$, sample mean, sample variance etc.

Statistics

1. Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
2. Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$

The statistic $S = \sqrt{S^2}$ is called the sample standard deviation
Question Why $1/(n-1)$ instead of $1/n$ in the definition of S^2?
Then S^2 becomes an unbiased (will be defined later) statistic for σ_X^2.

Sampling distributions

We denote $\overline{x}$ and s^2 as realizations from a particular sample. We denote X as the population random variable. Then note that μ_x and $\overline{X}$ need not be same.

Sampling distributions

We denote $\overline{x}$ and s^2 as realizations from a particular sample. We denote X as the population random variable. Then note that μ_x and $\overline{X}$ need not be same.

We are mainly interested finding the probability distributions of the sample mean $\overline{X}$ and sample variance S^2, that is the distribution of the statistics of samples.

Sampling distributions

We denote $\overline{x}$ and s^2 as realizations from a particular sample. We denote X as the population random variable. Then note that μ_x and $\overline{X}$ need not be same.

We are mainly interested finding the probability distributions of the sample mean $\overline{X}$ and sample variance S^2, that is the distribution of the statistics of samples.

Problems..

Sampling distributions

If $X_1, \ldots, X_n$ are mutually independent random variables with respective means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ then the mean and variance of $Y = \sum_{i=1}^{n} a_i X_i$, $a_i \in \mathbb{R}$ are given by

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \text{ and } \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Sampling distributions

If $X_1, \ldots, X_n$ are mutually independent random variables with respective means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ then the mean and variance of $Y = \sum_{i=1}^n a_i X_i$, $a_i \in \mathbb{R}$ are given by

$$\mu_Y = \sum_{i=1}^n a_i \mu_i \text{ and } \sigma_Y^2 = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Proof

$$\mu_Y = E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i \mu_i$$

Sampling distributions

If $X_1, \ldots, X_n$ are mutually independent random variables with respective means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ then the mean and variance of $Y = \sum_{i=1}^n a_i X_i$, $a_i \in \mathbb{R}$ are given by

$$\mu_Y = \sum_{i=1}^n a_i \mu_i \text{ and } \sigma_Y^2 = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Proof

$$\mu_Y = E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i \mu_i$$

Since X_is are mutually independent, $Cov(X_i, X_j) = 0$ when $i \neq j$. Thus

$$\sigma_Y^2 = Var(Y) = \sum_{i=1}^n Var(a_i X_i) = \sum_{i=1}^n a_i^2 \, Var(X_i) = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Sampling distributions

If $X_1, \ldots, X_n$ are independent random variables with respective mgfs $M_{X_i}(t)$, $i = 1, \ldots, n$ then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

Sampling distributions

If $X_1, \ldots, X_n$ are independent random variables with respective mgfs $M_{X_i}(t)$, $i = 1, \ldots, n$ then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

Proof

$$M_Y(t) = M_{\sum_{i=1}^{n} a_i X_i}(t) = \prod_{i=1}^{n} M_{a_i X_i}(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

Sampling distributions

If $X_1, \ldots, X_n$ are independent random variables with respective mgfs $M_{X_i}(t)$, $i = 1, \ldots, n$ then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

Proof

$$M_Y(t) = M_{\sum_{i=1}^{n} a_i X_i}(t) = \prod_{i=1}^{n} M_{a_i X_i}(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

If $X_1, \ldots, X_n$ are independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ then the random variable $Y = \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, where

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \text{ and } \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

(the proof from above result)

Sampling distributions

If $X_1, \ldots, X_n$ is a random sample of size n from a normal distribution with mean and variance σ^2 then $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (proof follows from above result)

Sampling distributions

If $X_1, \ldots, X_n$ is a random sample of size n from a normal distribution with mean and variance σ^2 then $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (proof follows from above result)

If $X_1, \ldots, X_n$ are independent random variables with respective distributions $\chi^2(r_1), \ldots, \chi^2(r_n)$ then

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} r_i\right)$$

Sampling distributions

If $X_1, \ldots, X_n$ is a random sample of size n from a normal distribution with mean and variance σ^2 then $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (proof follows from above result)

If $X_1, \ldots, X_n$ are independent random variables with respective distributions $\chi^2(r_1), \ldots, \chi^2(r_n)$ then

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} r_i\right)$$

Proof Since $X_i \sim \chi^2(r_i)$, the mgf of X_i is

$$M_{X_i}(t) = (1-2t)^{-\frac{r_i}{2}}.$$

Then from the above result regarding mgf,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n}(1-2t)^{-\frac{r_i}{2}} = (1-2t)^{-\frac{1}{2}\sum_{i-1}^{n} r_i}$$

Sampling distributions

If $X_1, \ldots, X_n$ are iid with standard normal distribution then

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$$

Sampling distributions

If $X_1, \ldots, X_n$ are iid with standard normal distribution then

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$$

Before we discuss our next result, let us denote

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Sampling distributions

If $X_1, \ldots, X_n$ is a random sample of size n from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ then

1. $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ and
2. $\overline{X}_n$ and S_n^2 are independent

Sampling distributions

If $X_1, \ldots, X_n$ is a random sample of size n from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ then

1. $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ and
2. $\overline{X}_n$ and S_n^2 are independent

Proof The proof follows by induction. let us prove it for $n = 2$. Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ then $X_1 + X_2 \sim \mathcal{N}(2\mu, 2\sigma^2)$ and $X_1 - X_2 \sim \mathcal{N}(0, 2\sigma^2)$. Hence

$$\frac{X_1 - X_2}{2\sigma^2} \sim \mathcal{N}(0, 1)$$

and thus

$$\frac{(X_1 - X_2)^2}{2\sigma^2} \sim \chi^2(1)$$

which proves $S_2^2 \sim \chi^2(1)$.

Sampling distributions

Now, since X_1 and X_2 are independent,

$$\begin{aligned} & Cov(X_1 + X_2, X_1 - X_2) \\ = & \; Cov(X_1, X_1) + Cov(X_1, X_2) - Cov(X_2, X_1) - Cov(X_2, X_2) \\ = & \; \sigma^2 + 0 - 0 - \sigma^2 = 0 \end{aligned} \quad (1)$$

Thus $X_1 + X_2$ and $X_1 - X_2$ are uncorrelated bivariate normal random variables. This yields that they are independent. Therefore $\frac{1}{2}(X_1 + X_2)$ and $\frac{1}{2}(X_1 - X_2)^2$ are independent i.e. X_2 and S_2^2 are independent.

Sampling distributions

Now we will sow that the sample mean gets close (in terms of probability) to the population mean when the sample size is large.

Sampling distributions

Now we will sow that the sample mean gets close (in terms of probability) to the population mean when the sample size is large.First we recall Chebychev inequality as follows.

Suppose X is a nonnegative random variable with variance σ^2. Then

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

for all $t > 0$.

Sampling distributions

Now we will sow that the sample mean gets close (in terms of probability) to the population mean when the sample size is large.First we recall Chebychev inequality as follows.

Suppose X is a nonnegative random variable with variance σ^2. Then

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

for all $t > 0$.

Definition

Suppose $X_1, X_2, \ldots$ is a sequence of random variables on a sample space. Then the sequence converges in probability to the random variable X if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

Sampling distributions

Weak law of large numbers (WLLN)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $1 = 1, 2, \ldots$ Then

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0 \text{ i.e. } \lim_{n\to\infty} P(|\overline{X}_n - \mu| < \epsilon) = 1.$$

Sampling distributions

Weak law of large numbers (WLLN)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $1 = 1, 2, \ldots$ Then

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0 \text{ i.e. } \lim_{n\to\infty} P(|\overline{X}_n - \mu| < \epsilon) = 1.$$

Proof We proved before that $E(\overline{X}_n) = \mu$ and $Var(\overline{X}_n) = \frac{\sigma^2}{n}$. Then by Chebyshev's inequality,

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

for > 0. Then the result follows by setting $n \to \infty$ both sides.

Sampling distributions

Question What is the meaning of WLLN?

Sampling distributions

Question What is the meaning of WLLN?

It means that the sequence of sample means from a population stays close to the population mean most of the time.

Sampling distributions

Question What is the meaning of WLLN?

It means that the sequence of sample means from a population stays close to the population mean most of the time.

For instance, consider an experiment of tossing a coin infinitely many times. Let X_i be 1 if the ith toss is head, and 0 otherwise. Then WLLN says

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \frac{1}{2} \text{ as } n \to \infty$$

Sampling distributions

Question What is the meaning of WLLN?

It means that the sequence of sample means from a population stays close to the population mean most of the time.

For instance, consider an experiment of tossing a coin infinitely many times. Let X_i be 1 if the ith toss is head, and 0 otherwise. Then WLLN says

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \frac{1}{2} \text{ as } n \to \infty$$

However, there can exists sequence of coin tosses link

$$HHHHHHHHHH\ldots$$

but the WLLN says that probability of occurrence of such a sequence is zero.

Sampling distributions

Definition

Let $X - 1, X_2, \ldots$ be a sequence of random variables on a sample space S. Then the sequence $X_n(\omega)$ converges almost surely to $X(\omega)$ if

$$P\left(\left\{\omega \in S | \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

where X is a random variable on the sample space S.

Sampling distributions

Definition

Let $X-1, X_2, \ldots$ be a sequence of random variables on a sample space S. Then the sequence $X_n(\omega)$ converges almost surely to $X(\omega)$ if

$$P\left(\left\{\omega \in S \middle| \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

where X is a random variable on the sample space S.

String law of large numbers (SLLN)

Let $X_1, X_2, \ldots$ be a sequence of iids with $\mu = E(X_i), i = 1, 2, \ldots$ Then

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

Sampling distributions

Recall: Let $X_1, \ldots, X_n$ be a random sample. Then these are iids with a common pdf which is the pdf of the population. Further, if the population pdf is normal then the the sample mean is normal i.e. if X_i is from the distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

then

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Sampling distributions

Recall: Let $X_1, \ldots, X_n$ be a random sample. Then these are iids with a common pdf which is the pdf of the population. Further, if the population pdf is normal then the the sample mean is normal i.e. if X_i is from the distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

then

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Sampling distributions

Central limit theorem/Lindeberg-Levy theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Then the limiting distribution of

$$Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is standard normal, i.e. Z_n converges in distribution to the standard normal random variable.