# Random Sample

**Definition:** The random variables

$$X_1, X_2, \ldots, X_n$$

are said to form a **random sample of size** $n$ if:

- Each $X_i$ is independent of the others
- Each $X_i$ has the same probability distribution

# Random Sample

**Definition:** The random variables

$$X_1, X_2, \ldots, X_n$$

are said to form a **random sample of size** $n$ if:

- Each $X_i$ is independent of the others
- Each $X_i$ has the same probability distribution

These conditions also imply that the expected values and variances of the random variables are the same:

- $E(X_i) = \mu$ for each $X_i$
- $V(X_i) = \sigma^2$ for each $X_i$

# Random Sample

Regardless of what the exact probability distribution is, the following is true:

If

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{ the sample mean}$$

then

$$E(\overline{x}) = \mu = E(X_i) \quad \text{for all } i$$

# Random Sample

Regardless of what the exact probability distribution is, the following is true:

If

$$\overline{x} \;=\; \frac{\sum_{i=1}^{n} x_i}{n} \;=\; \text{the sample mean}$$

then

$$E(\overline{x}) \;=\; \mu \;= E(X_i) \quad \text{for all } i$$

That is, the expected value of the sample mean $E(\overline{x})$ is the same as the expected value of each of the $X_i$.

# Random Sample

Regardless of what the exact probability distribution is, the following is true:

If

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{ the sample mean}$$

then

$$E(\overline{x}) = \mu = E(X_i) \quad \text{for all } i$$

That is, the expected value of the sample mean $E(\overline{x})$ is the same as the expected value of each of the $X_i$.

The above statement is true for a random sample from a distribution for which $E(X_i)$ exists. There are pathological distributions for which $E(X_i)$ does not exist but they are rare.

# Random Sample

Again regardless of what the exact probability distribution is, the following is true:

If

$$V(\overline{x}) \; = \; \text{variance of the sample mean}$$

then

$$V(\overline{x}) \; = \; \sigma_{\overline{x}}^2 \; = \frac{\sigma^2}{n}$$

# Random Sample

Again regardless of what the exact probability distribution is, the following is true:

If

$$V(\overline{x}) = \text{variance of the sample mean}$$

then

$$V(\overline{x}) = \sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$$

That is, the variance of the sample mean $V(\overline{x})$ is $1/n$ times the variance of each of the $X_i$.

# Random Sample

Again regardless of what the exact probability distribution is, the following is true:

If

$$V(\overline{x}) \;=\; \text{variance of the sample mean}$$

then

$$V(\overline{x}) \;=\; \sigma^2_{\overline{x}} \;=\frac{\sigma^2}{n}$$

That is, the variance of the sample mean $V(\overline{x})$ is $1/n$ times the variance of each of the $X_i$.

Note that $V(\overline{x})$ gets smaller as the sample size $n$ increases.

# Random Sample

Again regardless of what the exact probability distribution is, the following is true:

If

$$V(\overline{x}) \;=\; \text{variance of the sample mean}$$

then

$$V(\overline{x}) \;=\; \sigma^2_{\overline{x}} \;=\frac{\sigma^2}{n}$$

That is, the variance of the sample mean $V(\overline{x})$ is $1/n$ times the variance of each of the $X_i$.

Note that $V(\overline{x})$ gets smaller as the sample size $n$ increases.

As before, we note that the variance $V(X_i)$ of each $X_i$ must exists for this to hold.

# Normal Population Random Sample

**Proposition:** If the random sample is from a normal population with mean $\mu$ and standard deviation $\sigma$,

$$X_i \sim N(\mu, \sigma), \quad i = 1, 2, \ldots, n$$

That is, if each $X_i$ has a normal or bell curve distribution, then the sample mean $\overline{x}$ also has a normal distribution:

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

# Normal Population Random Sample

**Proposition:** If the random sample is from a normal population with mean $\mu$ and standard deviation $\sigma$,

$$X_i \sim N(\mu, \sigma), \quad i = 1, 2, \ldots, n$$

That is, if each $X_i$ has a normal or bell curve distribution, then the sample mean $\overline{x}$ also has a normal distribution:

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

It should be emphasized that mean and standard deviation of $\overline{x}$ are:

- $E(\overline{x}) = \mu$
- $\sigma_{\overline{x}} = \sigma/\sqrt{n}$

# Central Limit Theorem

What if the underlying population for the random sample does not have a normal distribution?

It turns out that regardless of the underlying distribution, $\overline{x}$ is approximately normally distributed.

# Central Limit Theorem

What if the underlying population for the random sample does not have a normal distribution?

It turns out that regardless of the underlying distribution, $\overline{x}$ is approximately normally distributed.

The larger the sample, the closer the actual distribution of $\overline{x}$ is to a normal distribution.

# Central Limit Theorem

What if the underlying population for the random sample does not have a normal distribution?

It turns out that regardless of the underlying distribution, $\overline{x}$ is approximately normally distributed.

The larger the sample, the closer the actual distribution of $\overline{x}$ is to a normal distribution.

The practical importance of this theorem cannot be overstated.

It means that instead of having to develop separate tables and procedures for each possible distribution, we can treat the mean of any sufficiently large random sample as if it were that of a sample from a normal population.

# Central Limit Theorem

Like many things in statistics, opinions vary on how large a sample is needed to treat the sample mean as a normal random variable.

# Central Limit Theorem

Like many things in statistics, opinions vary on how large a sample is needed to treat the sample mean as a normal random variable.

Some authors say $n = 30$, others (DeVore) say $n = 50$. There really is no hard and fast number.

# Central Limit Theorem

Like many things in statistics, opinions vary on how large a sample is needed to treat the sample mean as a normal random variable.

Some authors say $n = 30$, others (DeVore) say $n = 50$. There really is no hard and fast number.

Practically speaking, it is not a big issue. A study with enough subjects to produce a statistically meaningful result usually has enough to justify the use of the central limit theorem.

# Central Limit Theorem

We will note in passing that the Central Limit Theorem is actually considerably more general than what is stated in the text.

# Central Limit Theorem

We will note in passing that the Central Limit Theorem is actually considerably more general than what is stated in the text.

The stated properties (independent, identically distributed random variables in the sample) are those of a random sample.

# Central Limit Theorem

We will note in passing that the Central Limit Theorem is actually considerably more general than what is stated in the text.

The stated properties (independent, identically distributed random variables in the sample) are those of a random sample.

Although we will not make use of it in its full generality, the Central Limit Theorem actually does not require identically distributed random variables.

The least restrictive sufficient conditions for the theorem to hold are rather technical so most authors state a less general version of the theorem that is adequate for most purposes.

# Example

An IQ test is standardized to have a mean of $100$ and a standard deviation of $15$. Suppose a random sample of $100$ subjects is given the test.

Assuming individual scores are normally distributed with mean $100$ and standard deviation $15$, what is the distribution of the sample mean IQ score, $\overline{x}$?

# Example

An IQ test is standardized to have a mean of $100$ and a standard deviation of $15$. Suppose a random sample of $100$ subjects is given the test.

Assuming individual scores are normally distributed with mean $100$ and standard deviation $15$, what is the distribution of the sample mean IQ score, $\overline{x}$?

An earlier proposition states that

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(100, \frac{15}{\sqrt{100}}\right) = N(100, 1.5)$$

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is less than $103$?

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is less than $103$?

This is just the probability that a normal random variable with mean 100 and standard deviation 1.5 is less than (or equal to, equality doesn't really matter) 103, which we can calculate as

- $= NORMDIST(103, 100, 1.5, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 1.5)$ using R

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is less than $103$?

This is just the probability that a normal random variable with mean 100 and standard deviation 1.5 is less than (or equal to, equality doesn't really matter) 103, which we can calculate as

- $=NORMDIST(103, 100, 1.5, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 1.5)$ using R

Either way, the result is:

$$P(\overline{x} \leq 103) = 0.9772$$

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is between $98$ and $102$?

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is between $98$ and $102$?

This is just the probability that a normal random variable with mean 100 and standard deviation 1.5 falls between $98$ and $102$.

- $= NORMDIST(102, 100, 1.5, TRUE) - NORMDIST(98, 100, 1.5, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 1.5) - pnorm(98, 100, 1.5)$ using R

# Example

In the previous example, what is the probability that the sample mean $\overline{x}$ is between $98$ and $102$?

This is just the probability that a normal random variable with mean 100 and standard deviation 1.5 falls between $98$ and $102$.

- $= NORMDIST(102, 100, 1.5, TRUE) - NORMDIST(98, 100, 1.5, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 1.5) - pnorm(98, 100, 1.5)$ using R

Either way, the result is:

$$P(98 \leq \overline{x} \leq 102) = 0.8176$$

# Example

In the previous example, what is the probability that an individual IQ score is between $98$ and $102$?

# Example

In the previous example, what is the probability that an individual IQ score is between $98$ and $102$?

This is just the probability that a normal random variable with mean 100 and standard deviation 15 falls between $98$ and $102$.

- $= NORMDIST(102, 100, 15, TRUE) - NORMDIST(98, 100, 15, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 15) - pnorm(98, 100, 15)$ using R

# Example

In the previous example, what is the probability that an individual IQ score is between $98$ and $102$?

This is just the probability that a normal random variable with mean 100 and standard deviation 15 falls between $98$ and $102$.

- $= NORMDIST(102, 100, 15, TRUE) - NORMDIST(98, 100, 15, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 15) - pnorm(98, 100, 15)$ using R

Either way, the result is:

$$P(98 \leq X \leq 102) = 0.1061$$

# Example

In the previous example, what is the probability that an individual IQ score is between $98$ and $102$?

This is just the probability that a normal random variable with mean 100 and standard deviation 15 falls between $98$ and $102$.

- $= NORMDIST(102, 100, 15, TRUE) - NORMDIST(98, 100, 15, TRUE)$ using a spreadsheet
- $pnorm(103, 100, 15) - pnorm(98, 100, 15)$ using R

Either way, the result is:

$$P(98 \leq X \leq 102) = 0.1061$$

You should compare this to the previous example and understand why they are different.