Statistics and Their Distributions

Definition

The random variables $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size $n$ if

1. The $X_i$s are independent random variables.
2. Every $X_i$ has the same probability distribution.

In words, $X_1, X_2, \ldots, X_n$ forms a random sample if the $X_i$'s are independent and identically distributed (iid).
Definition
The random variables $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size $n$ if

1. The $X_i$s are independent random variables.
2. Every $X_i$ has the same probability distribution.
Definition
The random variables $X_1, X_2, \ldots, X_n$ are said to form a (simple) random sample of size $n$ if

1. The $X_i$'s are independent random variables.
2. Every $X_i$ has the same probability distribution.

In words, $X_1, X_2, \ldots, X_n$ forms a random sample if the $X_i$'s are independent and identically distributed (iid).
Remark:
When sampling with replacement or from an infinite (conceptual) population, the two conditions are satisfied and the result can be regarded as a random sample.

For sampling without replacement from a finite population, although consecutive observations are not independent and identically distributed, we can still regard the result as a random sample if the sample size $n$ is much smaller than the population size $N$. In practice, if $n/N \leq 0.05$ (at most 0.05% of the population is sampled), we can regard the sample as a random sample.
Remark:
When sampling with replacement or from an infinite (conceptual) population, the two conditions are satisfied and the result can be regarded as a random sample.
Remark:
When sampling with replacement or from an infinite (conceptual) population, the two conditions are satisfied and the result can be regarded as a random sample.
For sampling WITHOUT replacement from a finite population, although consecutive observations are not independent and identically distributed, we can still regard the result as a random sample if the sample size $n$ is much smaller than the population size $N$. 

In practice, if $\frac{n}{N} \leq 0.05$ (at most 0.05% of the population is sampled), we can regard the sample as a random sample.
Remark:
When sampling with replacement or from an infinite (conceptual) population, the two conditions are satisfied and the result can be regarded as a random sample.
For sampling WITHOUT replacement from a finite population, although consecutive observations are not independent and identically distributed, we can still regard the result as a random sample if the sample size \( n \) is much smaller than the population size \( N \).
In practice, if \( n/N \leq .05 \) (at most .05% of the population is sampled), we can regard the sample as a random sample.
Statistics and Their Distributions

Example (Problem 38)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>0.2</td>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma^2 = 0.49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1, X_2$ is a random sample of size $n = 2$.

a. Let $X = (X_1 + X_2)/2$. Find the probability distribution of $X$.

b. Calculate $P(X \leq 1)$.

c. Calculate $\mu_X$. How does it relate to $\mu$, the population mean?

d. Calculate $\sigma^2_X$. How does it relate to $\sigma^2$, the population variance?
Statistics and Their Distributions

Deriving Sampling Distributions

Example (Problem 38)
There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma^2 = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1, X_2$ is a random sample of size $n = 2$. 

a. Let $X = \frac{X_1 + X_2}{2}$. Find the probability distribution of $X$.

b. Calculate $P(X \leq 1)$.

c. Calculate $\mu_X$. How does it relate to $\mu$, the population mean?

d. Calculate $\sigma^2_X$. How does it relate to $\sigma^2$, the population variance?
Deriving Sampling Distributions

Example (Problem 38)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma^2 = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$.

a. Let $\overline{X} = (X_1 + X_2)/2$. Find the probability distribution of $\overline{X}$. 
Statistics and Their Distributions

Deriving Sampling Distributions

Example (Problem 38)
There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>$p(x_1)$</th>
<th>.2</th>
<th>.5</th>
<th>.3</th>
</tr>
</thead>
</table>

$\mu = 1.1$, $\sigma^2 = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1, X_2$ is a random sample of size $n = 2$.

a. Let $\bar{X} = (X_1 + X_2)/2$. Find the probability distribution of $\bar{X}$.

b. Calculate $P(\bar{X} \leq 1)$. 

Deriving Sampling Distributions

Example (Problem 38)
There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1, \sigma^2 = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1, X_2$ is a random sample of size $n = 2$.

a. Let $\bar{X} = (X_1 + X_2)/2$. Find the probability distribution of $\bar{X}$.

b. Calculate $P(\bar{X} \leq 1)$.

c. Calculate $\mu_{\bar{X}}$. How does it relate to $\mu$, the population mean?
Example (Problem 38)
There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma^2 = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$.

a. Let $\bar{X} = (X_1 + X_2)/2$. Find the probability distribution of $\bar{X}$.

b. Calculate $P(\bar{X} \leq 1)$.

c. Calculate $\mu_{\bar{X}}$. How does it relate to $\mu$, the population mean?

d. Calculate $\sigma^2_{\bar{X}}$. How does it relate to $\sigma^2$, the population variance?
Distribution for Sample Mean

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $\mathbb{E}(X) = \mu$
2. $\text{Var}(X) = \frac{\sigma^2}{n}$ and $\sigma_X = \frac{\sigma}{\sqrt{n}}$

In words, the expected value of the sample mean equals the population mean, which is called the unbiased property. And the variance of the sample mean equals $\frac{1}{n}$ of the population variance.
Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $E(X) = \mu_X = \mu$

2. $V(X) = \sigma^2_X = \sigma^2 / n$ and $\sigma_X = \sigma / \sqrt{n}$
Distribution for Sample Mean

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $E(\bar{X}) = \mu_{\bar{X}} = \mu$
2. $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

In words, the expected value of the sample mean equals the population mean, which is called the \textbf{unbiased} property. And the variance of the sample mean equals $\frac{1}{n}$ of the population variance.
Example (Problem 38 revisit)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$p(x_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma = 0.49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1, X_2$ is a random sample of size $n = 2$.

Let $X = (X_1 + X_2)/2$ denote the average stops.

a. Calculate $\mu_X$.

b. Calculate $\sigma^2_X$. 
Example (Problem 38 revisit)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$. Let $\bar{X} = (X_1 + X_2)/2$ denote the average stops.
Example (Problem 38 revisit)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$. Let $\bar{X} = (X_1 + X_2)/2$ denote the average stops.

a. Calculate $\mu_{\bar{X}}$. 
Example (Problem 38 revisit)

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop, and suppose that the distribution of $X_1$ is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_1)$</td>
<td>.2</td>
<td>.5</td>
<td>.3</td>
</tr>
</tbody>
</table>

$\mu = 1.1$, $\sigma = .49$

Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$.

Let $\overline{X} = (X_1 + X_2)/2$ denote the average stops.

a. Calculate $\mu_{\overline{X}}$.

b. Calculate $\sigma^2_{\overline{X}}$. 
Distribution for Sample Mean

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Define $T_0 = X_1 + X_2 + \cdots + X_n$, then $E(T_0) = n\mu$, $V(T_0) = n\sigma^2$ and $\sigma_{T_0} = \sqrt{n}\sigma$. 
Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Define $T_0 = X_1 + X_2 + \cdots + X_n$, then

$$E(T_0) = n\mu, \quad V(T_0) = n\sigma^2 \quad \text{and} \quad \sigma_{T_0} = \sqrt{n}\sigma$$
Proposition

Let \( X_1, X_2, \ldots, X_n \) be a random sample from a normal distribution with mean value \( \mu \) and standard deviation \( \sigma \). Then for any \( n \), \( \bar{X} \) is normally distributed (with mean value \( \mu \) and standard deviation \( \sigma / \sqrt{n} \)), as is \( T_0 \) (with mean value \( n \mu \) and standard deviation \( \sqrt{n} \sigma \)).
Distribution for Sample Mean

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean value $\mu$ and standard deviation $\sigma$. Then for any $n$, $\bar{X}$ is normally distributed (with mean value $\mu$ and standard deviation $\sigma/\sqrt{n}$), as is $T_0$ (with mean value $n\mu$ and standard deviation $\sqrt{n\sigma}$).
Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation .85 (suggested in “Modeling Sediment and Water Column Interactions for Hydrophobic Pollutants”, Water Research, 1984: 1169-1174).

a. If a random sample of 25 specimens is selected, what is the probability that the sample average sediment density is at most 3.00?

b. How large a sample size would be required to ensure that the above probability is at least .99?
Example (Problem 54)
Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation .85 (suggested in “Modeling Sediment and Water Column Interactions for Hydrophobic Pollutants”, Water Research, 1984: 1169-1174).
Example (Problem 54)
Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation .85 (suggested in “Modeling Sediment and Water Column Interactions for Hydrophobic Pollutants”, Water Research, 1984: 1169-1174).

a. If a random sample of 25 specimens is selected, what is the probability that the sample average sediment density is at most 3.00?
Example (Problem 54)
Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation .85 (suggested in “Modeling Sediment and Water Column Interactions for Hydrophobic Pollutants”, Water Research, 1984: 1169-1174).

a. If a random sample of 25 specimens is selected, what is the probability that the sample average sediment density is at most 3.00?

b. How large a sample size would be required to ensure that the above probability is at least .99?
Distribution for Sample Mean

The Central Limit Theorem (CLT)

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then if $n$ is sufficiently large, $X$ has approximately a normal distribution with mean value $\mu$ and standard deviation $\sigma/\sqrt{n}$, and $T_0$ also has approximately a normal distribution with mean value $n\mu$ and standard deviation $\sqrt{n}\sigma$. The larger the value of $n$, the better the approximation.

Remark:
1. As long as $n$ is sufficiently large, CLT is applicable no matter $X_i$'s are discrete random variables or continuous random variables.
2. How large should $n$ be such that CLT is applicable? Generally, if $n > 30$, CLT can be used.
Distribution for Sample Mean

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then if $n$ is sufficiently large, $\bar{X}$ has approximately a normal distribution with mean value $\mu$ and standard deviation $\sigma/\sqrt{n}$, and $T_0$ also has approximately a normal distribution with mean value $n\mu$ and standard deviation $\sqrt{n}\sigma$. The larger the value of $n$, the better the approximation.
Distribution for Sample Mean

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then if $n$ is sufficiently large, $\bar{X}$ has approximately a normal distribution with mean value $\mu$ and standard deviation $\sigma/\sqrt{n}$, and $T_0$ also has approximately a normal distribution with mean value $n\mu$ and standard deviation $\sqrt{n}\sigma$. The larger the value of $n$, the better the approximation.

Remark:

1. As long as $n$ is sufficiently large, CLT is applicable no matter $X_i$’s are discrete random variables or continuous random variables.
The Central Limit Theorem (CLT)

Let \( X_1, X_2, \ldots, X_n \) be a random sample from a distribution with mean value \( \mu \) and standard deviation \( \sigma \). Then if \( n \) is sufficiently large, \( \overline{X} \) has approximately a normal distribution with mean value \( \mu \) and standard deviation \( \sigma / \sqrt{n} \), and \( T_0 \) also has approximately a normal distribution with mean value \( n\mu \) and standard deviation \( \sqrt{n}\sigma \). The larger the value of \( n \), the better the approximation.

Remark:

1. As long as \( n \) is sufficiently large, CLT is applicable no matter \( X_i \)'s are discrete random variables or continuous random variables.

2. How large should \( n \) be such that CLT is applicable?
The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then if $n$ is sufficiently large, $\bar{X}$ has approximately a normal distribution with mean value $\mu$ and standard deviation $\sigma/\sqrt{n}$, and $T_0$ also has approximately a normal distribution with mean value $n\mu$ and standard deviation $\sqrt{n}\sigma$. The larger the value of $n$, the better the approximation.

Remark:

1. As long as $n$ is sufficiently large, CLT is applicable no matter $X_i$’s are discrete random variables or continuous random variables.

2. How large should $n$ be such that CLT is applicable? Generally, if $n > 30$, CLT can be used.
Example (Problem 49)

There are 40 students in an elementary statistics class. On the basis of years of experience, the instructor knows that the time needed to grade a randomly chosen first examination paper is a random variable with an expected value of 6 min and a standard deviation of 6 min.

(a) If grading times are independent and the instructor begins grading at 6:50pm and grades continuously, what is the (approximate) probability that he is through grading before the 11:00pm TV news begins?

(b) If the sports report begins at 11:10pm, what is the probability that he misses part of the report if he waits until grading is done before turning on the TV?
Example (Problem 49)
There are 40 students in an elementary statistics class. On the basis of years of experience, the instructor knows that the time needed to grade a randomly chosen first examination paper is a random variable with an expected value of 6 min and a standard deviation of 6 min.
Example (Problem 49)
There are 40 students in an elementary statistics class. On the basis of years of experience, the instructor knows that the time needed to grade a randomly chosen first examination paper is a random variable with an expected value of 6 min and a standard deviation of 6 min.

a. If grading times are independent and the instructor begins grading at 6:50pm and grades continuously, what is the (approximate) probability that he is through grading before the 11:00pm TV news begins?
Example (Problem 49)
There are 40 students in an elementary statistics class. On the basis of years of experience, the instructor knows that the time needed to grade a randomly chosen first examination paper is a random variable with an expected value of 6 min and a standard deviation of 6 min.

a. If grading times are independent and the instructor begins grading at 6:50pm and grades continuously, what is the (approximate) probability that he is through grading before the 11:00pm TV news begins?

b. If the sports report begins at 11:10pm, what is the probability that he misses part of the report if he waits until grading is done before turning on the TV?
Distribution for Sample Mean

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables from a distribution with mean value $\mu$ and standard deviation $\sigma$. Define random variables $Y_n = \sum_{i=1}^{n} X_i - n\mu \sqrt{n}$ for $n = 1, 2, \ldots$

Then as $n \to \infty$, $Y_n$ has approximately a normal distribution.
Distribution for Sample Mean

The original version of CLT

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables from a distribution with mean value $\mu$ and standard deviation $\sigma$. Define random variables

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma}}$$ for $n = 1, 2, \ldots$

Then as $n \to \infty$, $Y_n$ has approximately a normal distribution.
Distribution for Sample Mean

Corollary

Let \( X_1, X_2, \ldots, X_n \) be a random sample from a distribution for which only positive values are possible \([P(X_i > 0) = 1]\). Then if \( n \) is sufficiently large, the product \( Y = X_1 X_2 \cdots X_n \) has approximately a lognormal distribution.
Corollary

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution for which only positive values are possible [$P(X_i > 0) = 1$]. Then if $n$ is sufficiently large, the product $Y = X_1X_2 \cdots X_n$ has approximately a lognormal distribution.