The (Weak) Law of Large Numbers

Steve Ziskind

References

- Hoel, Port, Stone, *Introduction to Probability Theory*, Houghton Mifflin, 1971
- Papoulis, *Probability, Random Variables, and Stochastic Processes*, MacGraw-Hill, 1965.
- Lamperti, *Probability*, W.A.Benjamin, 1966
- Feller, *An Introduction to Probability Theory and its Applications, 3rd Ed.*, Wiley, 1968

Overview

- The theorem that we will discuss is best understood in the context of coin flipping. Coin flipping illustrates the Central Limit Theorem and the Law of Large Numbers: the two most important non-trivial results in probability theory.

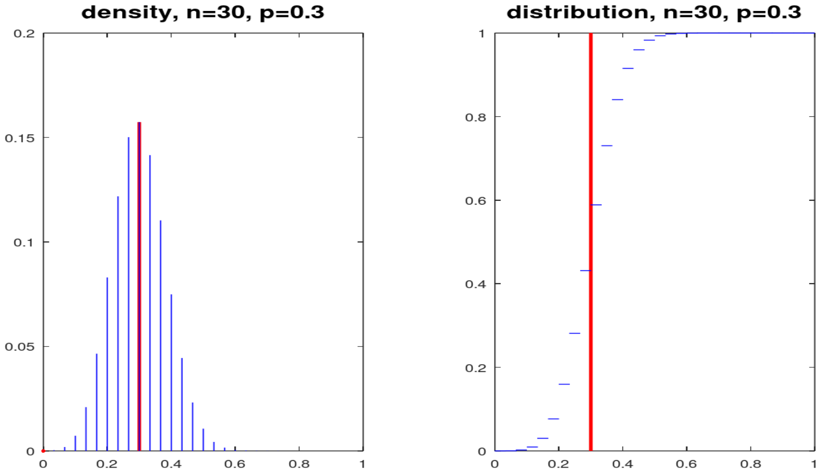
Overview

- The theorem that we will discuss is best understood in the context of coin flipping. Coin flipping illustrates the Central Limit Theorem and the Law of Large Numbers: the two most important non-trivial results in probability theory.
- When a coin is flipped it can result in a head or a tail, and we will say that a tail has value 0, while a head has value 1. If the coin is flipped n times, count the total of the values and divide by n, giving an average of the number of heads. This average can range from 0 for all tails to 1 for all heads, altogether n+1 different possibilities. We can plot both the probability of the possible values (density), and the probability that the average is $\leq$ various values (distribution). Let us suppose the probability of a head is p and a tail is 1-p. The probability of getting precisely k heads is $\binom{n}{k} * p^k * (1-p)^{n-k}$.

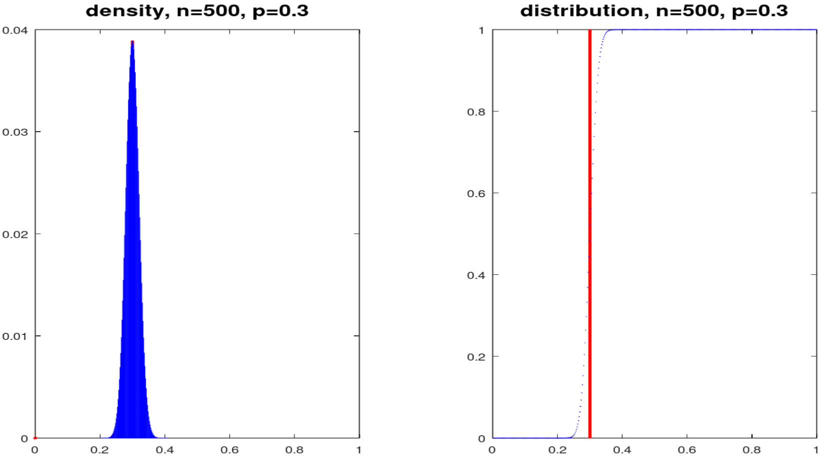
Overview

- The theorem that we will discuss is best understood in the context of coin flipping. Coin flipping illustrates the Central Limit Theorem and the Law of Large Numbers: the two most important non-trivial results in probability theory.
- When a coin is flipped it can result in a head or a tail, and we will say that a tail has value 0, while a head has value 1. If the coin is flipped n times, count the total of the values and divide by n, giving an average of the number of heads. This average can range from 0 for all tails to 1 for all heads, altogether n+1 different possibilities. We can plot both the probability of the possible values (density), and the probability that the average is $\leq$ various values (distribution). Let us suppose the probability of a head is p and a tail is 1-p. The probability of getting precisely k heads is $\binom{n}{k} * p^k * (1-p)^{n-k}$.
- When p=0.3, here are the results for n=30 and n=500.

Binomial Distribution



Binomial Distribution



Two Observations

- Two things should be noted from these plots.

Two Observations

- Two things should be noted from these plots.
- First, as the value of n increases the likelihood increases that the result will be close to p. This is an example of the Law of Large Numbers, which is true for more situations than coin flipping.

Two Observations

- Two things should be noted from these plots.
- First, as the value of n increases the likelihood increases that the result will be close to p. This is an example of the Law of Large Numbers, which is true for more situations than coin flipping.
- Second, the density plots have a definite shape, which becomes more and more like the function e^{-x^2}, the famous Bell Shaped Curve. Also true in very general situations, this second note is the Central Limit Theorem.

Two Observations

- Two things should be noted from these plots.
- First, as the value of n increases the likelihood increases that the result will be close to p. This is an example of the Law of Large Numbers, which is true for more situations than coin flipping.
- Second, the density plots have a definite shape, which becomes more and more like the function e^{-x^2}, the famous Bell Shaped Curve. Also true in very general situations, this second note is the Central Limit Theorem.
- In this talk we will prove a version of the Law of Large Numbers.

Two Observations

- Two things should be noted from these plots.
- First, as the value of n increases the likelihood increases that the result will be close to p. This is an example of the Law of Large Numbers, which is true for more situations than coin flipping.
- Second, the density plots have a definite shape, which becomes more and more like the function e^{-x^2}, the famous Bell Shaped Curve. Also true in very general situations, this second note is the Central Limit Theorem.
- In this talk we will prove a version of the Law of Large Numbers.
- In order to state things properly, we need to define some terms.

Some Basic Definitions

Some Basic Definitions

- A Sample Space (Ω) is any set, considered as the collection of possible outcomes of an experiment.

Some Basic Definitions

- A Sample Space (Ω) is any set, considered as the collection of possible outcomes of an experiment.
- Events are subsets of Ω that form a set closed under complementation, countable intersections, and countable unions.

Some Basic Definitions

- A Sample Space (Ω) is any set, considered as the collection of possible outcomes of an experiment.
- Events are subsets of Ω that form a set closed under complementation, countable intersections, and countable unions.
- A probability measure (P) is a function that assigns to every event a number between 0 and 1, satisfying these properties:
  1. $P(\Omega) = 1$
  2. P is countably additive. i.e. if A_n are mutually disjoint events, then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$

A Few Fundamental Examples

A Few Fundamental Examples

- Coin Flipping: $\Omega = \{H, T\}$, and all possible subsets are events. $P(H) = P(T) = 1/2$

A Few Fundamental Examples

- Coin Flipping: $\Omega = \{H, T\}$, and all possible subsets are events. $P(H) = P(T) = 1/2$
- Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, and all possible subsets are events. $P(1) = ...P(6) = 1/6$

A Few Fundamental Examples

- Coin Flipping: $\Omega = \{H, T\}$, and all possible subsets are events. $P(H) = P(T) = 1/2$
- Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, and all possible subsets are events. $P(1) = ...P(6) = 1/6$
- $\Omega = [0, 1]$. The measure of an event is its length (whatever that means). Thus the probability of an event occurring, i.e. of a point being picked within the set, is the size of the set. Unfortunately, not all subsets can be events, but most are.

A Few Fundamental Examples

- Coin Flipping: $\Omega = \{H, T\}$, and all possible subsets are events. $P(H) = P(T) = 1/2$
- Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, and all possible subsets are events. $P(1) = ...P(6) = 1/6$
- $\Omega = [0, 1]$. The measure of an event is its length (whatever that means). Thus the probability of an event occurring, i.e. of a point being picked within the set, is the size of the set. Unfortunately, not all subsets can be events, but most are.
- Let f be a non-negative function on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} f = 1$ Then $P(A) = \int_A f$.

A Few Fundamental Examples

- Coin Flipping: $\Omega = \{H, T\}$, and all possible subsets are events. $P(H) = P(T) = 1/2$
- Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, and all possible subsets are events. $P(1) = ...P(6) = 1/6$
- $\Omega = [0, 1]$. The measure of an event is its length (whatever that means). Thus the probability of an event occurring, i.e. of a point being picked within the set, is the size of the set. Unfortunately, not all subsets can be events, but most are.
- Let f be a non-negative function on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} f = 1$ Then $P(A) = \int_A f$.
- On the sample space [0,1] let the probability of the event [0.5,1] be 1/2. This is equivalent to flipping a fair coin, where the event [0,0.5] is a tail and [0.5,1] is a head.

A Confession

A Confession

- Some of the examples on the preceding slide are not well defined or explained.

A Confession

- Some of the examples on the preceding slide are not well defined or explained.
- Carefully defining which subsets of the line are events requires Measure Theory.

A Confession

- Some of the examples on the preceding slide are not well defined or explained.
- Carefully defining which subsets of the line are events requires Measure Theory.
- The integral in the last example is taken in the sense of Lebesgue, not Riemann.

A Confession

- Some of the examples on the preceding slide are not well defined or explained.
- Carefully defining which subsets of the line are events requires Measure Theory.
- The integral in the last example is taken in the sense of Lebesgue, not Riemann.
- Measure Theory and the Lebesgue Integral, while very interesting topics, are “Beyond the Scope of this Talk”.

A Confession

- Some of the examples on the preceding slide are not well defined or explained.
- Carefully defining which subsets of the line are events requires Measure Theory.
- The integral in the last example is taken in the sense of Lebesgue, not Riemann.
- Measure Theory and the Lebesgue Integral, while very interesting topics, are "Beyond the Scope of this Talk".
- If demanded, a separate talk might address them.

More Definitions

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as:
  $F_X(t) = P(X \leq t)$

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as: $F_X(t) = P(X \leq t)$
- Two r.v.s are Identically Distributed if they have equal distribution functions.

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as: $F_X(t) = P(X \leq t)$
- Two r.v.s are Identically Distributed if they have equal distribution functions.
- If the distribution of the r.v. X satisfies $F_X(t) = \int_{-\infty}^{t} f dx$, then f is called the Density of X.

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as: $F_X(t) = P(X \leq t)$
- Two r.v.s are Identically Distributed if they have equal distribution functions.
- If the distribution of the r.v. X satisfies $F_X(t) = \int_{-\infty}^{t} f dx$, then f is called the Density of X.
- If $\Omega = \mathbb{R}$ and X has the density f, The Mean Value (a.k.a. Expected Value) of an r.v. X is $E(X) = \int_{\mathbb{R}} X(t) f(t) dt$

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as: $F_X(t) = P(X \leq t)$
- Two r.v.s are Identically Distributed if they have equal distribution functions.
- If the distribution of the r.v. X satisfies $F_X(t) = \int_{-\infty}^{t} fdx$, then f is called the Density of X.
- If $\Omega = \mathbb{R}$ and X has the density f, The Mean Value (a.k.a. Expected Value) of an r.v. X is $E(X) = \int_{\mathbb{R}} X(t)f(t)dt$
- For discrete sample spaces (coins, dice, etc.), E(X) is defined as $\sum t * P(X = t)$

More Definitions

- A Random Variable (r.v.) is a real valued function on the sample space. (More or less)
- The Distribution Function of the r.v. X is defined as: $F_X(t) = P(X \leq t)$
- Two r.v.s are Identically Distributed if they have equal distribution functions.
- If the distribution of the r.v. X satisfies $F_X(t) = \int_{-\infty}^{t} f dx$, then f is called the Density of X.
- If $\Omega = \mathbb{R}$ and X has the density f, The Mean Value (a.k.a. Expected Value) of an r.v. X is $E(X) = \int_{\mathbb{R}} X(t) f(t) dt$
- For discrete sample spaces (coins, dice, etc.), E(X) is defined as $\sum t * P(X = t)$
- We will generally denote mean value with μ. Be careful: sometimes the mean value is not defined.

More Definitions (Sorry, these will be the last)

More Definitions (Sorry, these will be the last)

- Two r.v.s, X and Y, both defined on a common sample space, have a Joint Distribution defined by $F_{s,t} = P(X \leq s, Y \leq t)$

More Definitions (Sorry, these will be the last)

- Two r.v.s, X and Y, both defined on a common sample space, have a Joint Distribution defined by $F_{s,t} = P(X \leq s, Y \leq t)$
- X and Y are Independent if $P(a < X \leq b, c < Y \leq d) = P(a < X \leq b) * P(c < Y \leq d)$. If each has a density, this is equivalent to saying that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

More Definitions (Sorry, these will be the last)

- Two r.v.s, X and Y, both defined on a common sample space, have a Joint Distribution defined by $F_{s,t} = P(X \leq s, Y \leq t)$
- X and Y are Independent if $P(a < X \leq b, c < Y \leq d) = P(a < X \leq b) * P(c < Y \leq d)$. If each has a density, this is equivalent to saying that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$
- If X has an expected value of μ, the the Variance of X (generally denoted σ^2) is $E((X - \mu)^2)$. Some r.v.s can have expectations but not have variances. The square root of the variance is called the Standard Deviation.

More Definitions (Sorry, these will be the last)

- Two r.v.s, X and Y, both defined on a common sample space, have a Joint Distribution defined by $F_{s,t} = P(X \leq s, Y \leq t)$
- X and Y are Independent if $P(a < X \leq b, c < Y \leq d) = P(a < X \leq b) * P(c < Y \leq d)$. If each has a density, this is equivalent to saying that $f_{X,Y}(x, y) = f_X(x) f_Y(y)$
- If X has an expected value of μ, the the Variance of X (generally denoted σ^2) is $E((X - \mu)^2)$. Some r.v.s can have expectations but not have variances. The square root of the variance is called the Standard Deviation.
- A handy fact: $\sigma^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$

More Definitions (Sorry, these will be the last)

- Two r.v.s, X and Y, both defined on a common sample space, have a Joint Distribution defined by $F_{s,t} = P(X \leq s, Y \leq t)$
- X and Y are Independent if $P(a < X \leq b, c < Y \leq d) = P(a < X \leq b) * P(c < Y \leq d)$. If each has a density, this is equivalent to saying that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$
- If X has an expected value of μ, the the Variance of X (generally denoted σ^2) is $E((X-\mu)^2)$. Some r.v.s can have expectations but not have variances. The square root of the variance is called the Standard Deviation.
- A handy fact: $\sigma^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$
- Even handier: $E(X+Y) = E(X) + E(Y)$ and, when X and Y are independent, $Var(X+Y) = Var(X) + Var(Y)$. Also, for any constant k, $Var(kX) = k^2 Var(X)$

Examples

- If X is the r.v. corresponding to the value of a rolled die (i.e. 6 members of the sample space, each equally likely), then the expected value of X is $1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6} = \frac{7}{2}$

Examples

- If X is the r.v. corresponding to the value of a rolled die (i.e. 6 members of the sample space, each equally likely), then the expected value of X is $1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6} = \frac{7}{2}$
- The variance of the rolled die is $((1 - \frac{7}{2})^2 + (2 - \frac{7}{2})^2 + \cdots + (6 - \frac{7}{2})^2) * \frac{1}{6} = \frac{35}{12}$

Examples

- If X is the r.v. corresponding to the value of a rolled die (i.e. 6 members of the sample space, each equally likely), then the expected value of X is $1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6} = \frac{7}{2}$
- The variance of the rolled die is $((1 - \frac{7}{2})^2 + (2 - \frac{7}{2})^2 + \cdots + (6 - \frac{7}{2})^2) * \frac{1}{6} = \frac{35}{12}$
- If X assigns a value of -1 to a tail and 1 to a head when tossing a fair coin, then $\mu = 0$ and $\sigma^2 = 1$.

Examples

- If X is the r.v. corresponding to the value of a rolled die (i.e. 6 members of the sample space, each equally likely), then the expected value of X is $1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6} = \frac{7}{2}$
- The variance of the rolled die is $((1 - \frac{7}{2})^2 + (2 - \frac{7}{2})^2 + \cdots + (6 - \frac{7}{2})^2) * \frac{1}{6} = \frac{35}{12}$
- If X assigns a value of -1 to a tail and 1 to a head when tossing a fair coin, then $\mu = 0$ and $\sigma^2 = 1$.
- If X is the value of a point chosen from [-1,1], with all choices equally likely, then the density is $f(t) = \frac{1}{2}$. Then $\mu = 0$ and $\sigma^2 = \int_{-1}^{1} (t - 0)^2 * \frac{1}{2} dt = \frac{1}{3}$

The Law of Large Numbers

The Law of Large Numbers

- If $X_1, X_2, \ldots$ are independent, identically distribution random variables (iidrv's), define the n^{th} sample sum as $S_n = X_1 + X_2 + \cdots + X_n$

The Law of Large Numbers

- If $X_1, X_2, \ldots$ are independent, identically distribution random variables (iidrv's), define the n^{th} sample sum as $S_n = X_1 + X_2 + \cdots + X_n$
- (Weak Law of Large Numbers) Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and let $\{S_k\}_{k=1}^{\infty}$ be the corresponding sample sums. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

The Law of Large Numbers

- If $X_1, X_2, \dots$ are independent, identically distribution random variables (iidrv's), define the n^{th} sample sum as $S_n = X_1 + X_2 + \cdots + X_n$
- (Weak Law of Large Numbers) Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and let $\{S_k\}_{k=1}^{\infty}$ be the corresponding sample sums. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

- We will not prove this result in this generality. However, there is a slightly weaker version that can be proved simply, using Chebyshev's Inequality.

The Law of Large Numbers

- If $X_1, X_2, \ldots$ are independent, identically distribution random variables (iidrv's), define the n^{th} sample sum as $S_n = X_1 + X_2 + \cdots + X_n$
- (Weak Law of Large Numbers) Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and let $\{S_k\}_{k=1}^{\infty}$ be the corresponding sample sums. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

- We will not prove this result in this generality. However, there is a slightly weaker version that can be proved simply, using Chebyshev's Inequality.
- The simpler version still has enough power to prove a surprising theorem of Weierstrass.

Chebyshev's Inequality

Chebyshev's Inequality

- Lemma: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$

Chebyshev's Inequality

- Lemma: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$
- Proof: Define a new r.v. Y, by $Y = 0$ when $X < t$ and $Y = t$ when $X \geq t$. Note that $X \geq Y$, so $E(X) \geq E(Y)$

Chebyshev's Inequality

- Lemma: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$
- Proof: Define a new r.v. Y, by $Y = 0$ when $X < t$ and $Y = t$ when $X \geq t$. Note that $X \geq Y$, so $E(X) \geq E(Y)$
- Y has only values of 0 and t, so $E(Y) = 0 * P(Y = 0) + t * P(Y = t) = t * P(X \geq t)$ QED

Chebyshev's Inequality

- Lemma: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$
- Proof: Define a new r.v. Y, by $Y = 0$ when $X < t$ and $Y = t$ when $X \geq t$. Note that $X \geq Y$, so $E(X) \geq E(Y)$
- Y has only values of 0 and t, so
  $E(Y) = 0 * P(Y = 0) + t * P(Y = t) = t * P(X \geq t)$ QED
- (Chebyshev's Inequality) Let X be an r.v. with $E(X) = \mu$ and Variance=σ^2. The for any real $t > 0$, $P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$

Chebyshev's Inequality

- Lemma: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$
- Proof: Define a new r.v. Y, by $Y = 0$ when $X < t$ and $Y = t$ when $X \geq t$. Note that $X \geq Y$, so $E(X) \geq E(Y)$
- Y has only values of 0 and t, so
  $E(Y) = 0 * P(Y = 0) + t * P(Y = t) = t * P(X \geq t)$ QED
- (Chebyshev's Inequality) Let X be an r.v. with $E(X) = \mu$ and Variance$=\sigma^2$. The for any real $t > 0$, $P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$
- Proof: Apply the lemma to the r.v $(X - \mu)^2$ and the number t^2. QED

Chebyshev's Inequality

- <u>Lemma</u>: If $X \geq 0$ is a r.v. with expected value $E(X)$ and $t > 0$, then $P(X \geq t) \leq E(X)/t$
- <u>Proof</u>: Define a new r.v. Y, by $Y = 0$ when $X < t$ and $Y = t$ when $X \geq t$. Note that $X \geq Y$, so $E(X) \geq E(Y)$
- Y has only values of 0 and t, so $E(Y) = 0 * P(Y = 0) + t * P(Y = t) = t * P(X \geq t)$ QED
- (Chebyshev's Inequality) Let X be an r.v. with $E(X) = \mu$ and Variance$=\sigma^2$. The for any real $t > 0$, $P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$
- <u>Proof</u>: Apply the lemma to the r.v $(X - \mu)^2$ and the number t^2. QED
- Feller: A small variance indicates that large deviations from the mean are improbable.

The Very Weak Law of Large Numbers

- Very Weak Law of Large Numbers Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and finite variances of σ^2. Let $\{S_n\} = X_1 + X_2 + \cdots + X_n$ be the corresponding partial sums. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

The Very Weak Law of Large Numbers

- Very Weak Law of Large Numbers Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and finite variances of σ^2. Let$\{S_n\} = X_1 + X_2 + \cdots + X_n$ be the corresponding partial sums. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

- The difference from the full Weak Law is that we assume finite variances. The theorem is true without this assumption, but the proof is harder. [Feller] has a nice proof of the full result.

The Very Weak Law of Large Numbers

- Very Weak Law of Large Numbers Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ, and finite variances of σ^2.
  Let $\{S_n\} = X_1 + X_2 + \cdots + X_n$ be the corresponding partial sums.
  Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0$$

- The difference from the full Weak Law is that we assume finite variances. The theorem is true without this assumption, but the proof is harder. [Feller] has a nice proof of the full result.
- Proof: $E(S_n) = n\mu$. Because of independence, $\mathrm{Var}(S_n) = n\sigma^2$. Apply Chebyshev to find that

$$P(|S_n - n\mu| > n\epsilon) \leq \frac{n\sigma^2}{(n\epsilon)^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

The Weierstrass Polynomial Approximation Theorem

The Weierstrass Polynomial Approximation Theorem

- Using the VWLLN, we can provide a relatively short proof of a remarkable theorem of Weierstrass: Every continuous function on a closed bounded interval can be approximated uniformly by a polynomial. WLOG we will take the interval to be [0,1].

The Weierstrass Polynomial Approximation Theorem

- Using the VWLLN, we can provide a relatively short proof of a remarkable theorem of Weierstrass: Every continuous function on a closed bounded interval can be approximated uniformly by a polynomial. WLOG we will take the interval to be [0,1].
- We need a fact proved in advanced calculus: Any continuous function f on a closed bounded interval is Uniformly Continuous. This means that in the δ, ϵ definition of continuity a single δ can be chosen that will satisfy the $|f(x) - f(a)| < \epsilon$ requirement everywhere on the interval. If the interval is not closed, then each point might require its own special δ for the given ϵ.

The Weierstrass Polynomial Approximation Theorem

- Using the VWLLN, we can provide a relatively short proof of a remarkable theorem of Weierstrass: Every continuous function on a closed bounded interval can be approximated uniformly by a polynomial. WLOG we will take the interval to be [0,1].
- We need a fact proved in advanced calculus: Any continuous function f on a closed bounded interval is Uniformly Continuous. This means that in the δ, ϵ definition of continuity a single δ can be chosen that will satisfy the $|f(x) - f(a)| < \epsilon$ requirement everywhere on the interval. If the interval is not closed, then each point might require its own special δ for the given ϵ.
- The use of uniform continuity is this: by breaking [0,1] into sufficiently small intervals, we can treat the continuous function f as if it were nearly constant on each piece. Then we can approximate f by approximating a single value in each piece, turning a continuous problem into a discrete problem.

Lagrange Interpolation

- Given any two points $(a_1, b_1), (a_2, b_2)$, with $a_1 \neq a_2$, there is a line (linear function) passing through the two points. If we are given three points then there is a quadratic passing through the three points. In general, given n+1 points, (a_i, b_i) there is a polynomial of degree $\leq n$ whose graph passes through these points.

Lagrange Interpolation

- Given any two points $(a_1, b_1), (a_2, b_2)$, with $a_1 \neq a_2$, there is a line (linear function) passing through the two points. If we are given three points then there is a quadratic passing through the three points. In general, given n+1 points, (a_i, b_i) there is a polynomial of degree $\leq n$ whose graph passes through these points.
- The simple idea is to build a personal polynomial for each of the points, having the desired value of b_i at a_i, but which is zero at the other a_k values.

Lagrange Interpolation

- Given any two points $(a_1, b_1), (a_2, b_2)$, with $a_1 \neq a_2$, there is a line (linear function) passing through the two points. If we are given three points then there is a quadratic passing through the three points. In general, given n+1 points, (a_i, b_i) there is a polynomial of degree $\leq n$ whose graph passes through these points.
- The simple idea is to build a personal polynomial for each of the points, having the desired value of b_i at a_i, but which is zero at the other a_k values.
- The i^{th} polynomial is $p_i(x) = b_i * \prod_{k \neq i} \frac{x - a_k}{a_i - a_k}$

Lagrange Interpolation

- Given any two points $(a_1, b_1), (a_2, b_2)$, with $a_1 \neq a_2$, there is a line (linear function) passing through the two points. If we are given three points then there is a quadratic passing through the three points. In general, given n+1 points, (a_i, b_i) there is a polynomial of degree $\leq n$ whose graph passes through these points.
- The simple idea is to build a personal polynomial for each of the points, having the desired value of b_i at a_i, but which is zero at the other a_k values.
- The i^{th} polynomial is $p_i(x) = b_i * \prod_{k \neq i} \frac{x - a_k}{a_i - a_k}$
- Now the full solution is $\sum_{i=1}^{n+1} p_i(x)$, the Lagrange Interpolating Polynomial

Lagrange Interpolation

- Given any two points $(a_1, b_1), (a_2, b_2)$, with $a_1 \neq a_2$, there is a line (linear function) passing through the two points. If we are given three points then there is a quadratic passing through the three points. In general, given n+1 points, (a_i, b_i) there is a polynomial of degree $\leq n$ whose graph passes through these points.
- The simple idea is to build a personal polynomial for each of the points, having the desired value of b_i at a_i, but which is zero at the other a_k values.
- The i^{th} polynomial is $p_i(x) = b_i * \prod_{k \neq i} \frac{x - a_k}{a_i - a_k}$
- Now the full solution is $\sum_{i=1}^{n+1} p_i(x)$, the Lagrange Interpolating Polynomial
- The important idea is to deal with each point separately, and in such a way that each does not interfere with the others.

The Weierstrass Polynomial Approximation Theorem

- Let $t \in [0,1]$, and let $X_1 \dots X_n$ be i.i.d.r.v's satisfying $P(X_k = 1) = t, P(X_k = 0) = (1-t)$. Then define

$$B_n(t;f) = E\left(f\left(\frac{X_1 + \dots X_n}{n}\right)\right) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k} t^k (1-t)^{n-k}$$

The Weierstrass Polynomial Approximation Theorem

- Let $t \in [0,1]$, and let $X_1 \dots X_n$ be i.i.d.r.v's satisfying $P(X_k = 1) = t, P(X_k = 0) = (1-t)$. Then define

$$B_n(t;f) = E\left(f\left(\frac{X_1 + \dots X_n}{n}\right)\right) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k} t^k (1-t)^{n-k}$$

- $B_n(t;f)$ is a polynomial in t, called a Bernstein Polynomial.

The Weierstrass Polynomial Approximation Theorem

- Let $t \in [0,1]$, and let $X_1 \dots X_n$ be i.i.d.r.v's satisfying $P(X_k = 1) = t, P(X_k = 0) = (1-t)$. Then define

$$B_n(t; f) = E\left(f\left(\frac{X_1 + \dots X_n}{n}\right)\right) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k} t^k (1-t)^{n-k}$$

- $B_n(t; f)$ is a polynomial in t, called a Bernstein Polynomial.
- Weierstrass Theorem: Let f be continuous on [0,1], and let $B_n(t; f)$ be as defined above. Then

$$\lim_{n\to\infty} \max_{0\le t\le 1} |f(t) - B_n(t; f)| = 0$$

The Weierstrass Polynomial Approximation Theorem

- Let $t \in [0,1]$, and let $X_1 \dots X_n$ be i.i.d.r.v's satisfying $P(X_k = 1) = t, P(X_k = 0) = (1-t)$. Then define

$$B_n(t; f) = E\left(f\left(\frac{X_1 + \dots X_n}{n}\right)\right) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k} t^k (1-t)^{n-k}$$

- $B_n(t; f)$ is a polynomial in t, called a Bernstein Polynomial.
- Weierstrass Theorem: Let f be continuous on [0,1], and let $B_n(t; f)$ be as defined above. Then

$$\lim_{n\to\infty} \max_{0 \le t \le 1} |f(t) - B_n(t; f)| = 0$$

- Proof: Given any small $\epsilon > 0$, uniform continuity allows us to choose δ so that $|f(x) - f(y)| < \epsilon$ whenever $|x - y| < \delta$. Note that for any value of t, the binomial theorem says that

$$f(t) = f(t) * (t + (1-t))^n = f(t) * \sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k}$$

The Weierstrass Polynomial Approximation Theorem

- Now write $|f(t) - B_n(t;f)| =$

$$\left|\sum_{k=0}^{n}(f(t) - f(\frac{k}{n}))\binom{n}{k}t^k(1-t)^{n-k}\right| \leq \sum_{|\frac{k}{n}-t|\leq\delta}|\ldots| + \sum_{|\frac{k}{n}-t|>\delta}|\ldots|$$

- The first sum is bounded by the arbitrarily small ϵ when k/n is close to t, using uniform continuity.

The Weierstrass Polynomial Approximation Theorem

- Now write $|f(t) - B_n(t; f)| =$

$$\left|\sum_{k=0}^{n}(f(t) - f(\frac{k}{n}))\binom{n}{k}t^k(1-t)^{n-k}\right| \leq \sum_{|\frac{k}{n}-t|\leq\delta} |\ldots| + \sum_{|\frac{k}{n}-t|>\delta} |\ldots|$$

- The first sum is bounded by the arbitrarily small ϵ when k/n is close to t, using uniform continuity.
- Writing M for max($|f|$) on [0,1], the second sum is bounded by

$$2M * \sum_{|\frac{k}{n}-t|>\delta} \binom{n}{k}t^k(1-t)^{n-k} = 2M * P\left\{\left|\frac{X_1 + \ldots X_n}{n} - t\right| > \delta\right\}$$

The Weierstrass Polynomial Approximation Theorem

- Now write $|f(t) - B_n(t; f)| =$

$$\left|\sum_{k=0}^{n}(f(t) - f(\frac{k}{n}))\binom{n}{k}t^k(1-t)^{n-k}\right| \leq \sum_{|\frac{k}{n}-t|\leq\delta} |\ldots| + \sum_{|\frac{k}{n}-t|>\delta} |\ldots|$$

- The first sum is bounded by the arbitrarily small ϵ when k/n is close to t, using uniform continuity.
- Writing M for max($|f|$) on [0,1], the second sum is bounded by

$$2M * \sum_{|\frac{k}{n}-t|>\delta} \binom{n}{k}t^k(1-t)^{n-k} = 2M * P\left\{\left|\frac{X_1 + \ldots X_n}{n} - t\right| > \delta\right\}$$

- The VWLLN says that the last probability can made small by making n large. Thus each of the 2 pieces dominating $|f(t) - B_n(t; f)|$ are small. QED

The Strong Law of Large Numbers

- <u>Strong Law of Large Numbers</u> Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ. Let$\{S_n\} = X_1 + X_2 + \cdots + X_n$ be the corresponding partial sums. Then

$$P\left(\frac{S_n}{n} \to \mu\right) = 1$$

The Strong Law of Large Numbers

- Strong Law of Large Numbers Let $\{X_k\}_{k=1}^{\infty}$ be iidrv's, each having expected value of μ. Let $\{S_n\} = X_1 + X_2 + \cdots + X_n$ be the corresponding partial sums. Then

$$P\left(\frac{S_n}{n} \to \mu\right) = 1$$

- Proof: This requires some measure theory. A subject for another talk.